

General Relativity, Black Holes, and Cosmology

Andrew J. S. Hamilton

24 February 2026

Contents

<i>Table of contents</i>	<i>page</i> iii
<i>List of illustrations</i>	xx
<i>List of tables</i>	xxvi
<i>List of exercises and concept questions</i>	xxvii
<i>Legal notice</i>	1
<i>Notation</i>	2
PART ONE FUNDAMENTALS	5
<i>Concept Questions</i>	7
<i>What's important?</i>	9
1 Special Relativity	10
1.1 Motivation	10
1.2 The postulates of special relativity	11
1.3 The paradox of the constancy of the speed of light	13
1.4 Simultaneity	17
1.5 Time dilation	18
1.6 Lorentz transformation	19
1.7 Paradoxes: Time dilation, Lorentz contraction, and the Twin paradox	22
1.8 The spacetime wheel	26
1.9 Scalar spacetime distance	31
1.10 4-vectors	33
1.11 Energy-momentum 4-vector	35
1.12 Photon energy-momentum	37
1.13 What things look like at relativistic speeds	39
1.14 Occupation number, phase-space volume, intensity, and flux	44
1.15 How to program Lorentz transformations on a computer	46

	<i>Concept Questions</i>	48
	<i>What's important?</i>	50
2	Fundamentals of General Relativity	51
2.1	Motivation	52
2.2	The postulates of General Relativity	53
2.3	Implications of Einstein's principle of equivalence	55
2.4	Metric	57
2.5	Timelike, spacelike, proper time, proper distance	58
2.6	Orthonormal tetrad basis γ_m	58
2.7	Basis of coordinate tangent vectors e_μ	59
2.8	4-vectors and tensors	60
2.9	Covariant derivatives	63
2.10	Torsion	66
2.11	Connection coefficients in terms of the metric	68
2.12	Torsion-free covariant derivative	68
2.13	Mathematical aside: What if there is no metric?	72
2.14	Coordinate 4-velocity	72
2.15	Geodesic equation	73
2.16	Coordinate 4-momentum	74
2.17	Affine parameter	74
2.18	Affine distance	75
2.19	Riemann tensor	77
2.20	Ricci tensor, Ricci scalar	80
2.21	Einstein tensor	80
2.22	Bianchi identities	80
2.23	Covariant conservation of the Einstein tensor	81
2.24	Einstein equations	81
2.25	Summary of the path from metric to the energy-momentum tensor	82
2.26	Energy-momentum tensor of a perfect fluid	83
2.27	Cosmological constant	83
2.28	Newtonian limit	84
3	More on the coordinate approach	91
3.1	Weyl tensor	91
3.2	Evolution equations for the Weyl tensor, and gravitational waves	92
3.3	Geodesic deviation	94
4	Action principle for point particles	96
4.1	Principle of least action for point particles	96
4.2	Generalized momentum	98
4.3	Lagrangian for a test particle	98

4.4	Massless test particle	100
4.5	Effective Lagrangian for a test particle	101
4.6	Nice Lagrangian for a test particle	102
4.7	Action for a charged test particle in an electromagnetic field	103
4.8	Symmetries and constants of motion	105
4.9	Conformal symmetries	106
4.10	(Super-)Hamiltonian	109
4.11	Conventional Hamiltonian	110
4.12	Conventional Hamiltonian for a test particle	110
4.13	Effective (super-)Hamiltonian for a test particle with electromagnetism	112
4.14	Nice (super-)Hamiltonian for a test particle with electromagnetism	113
4.15	Derivatives of the action	114
4.16	Hamilton-Jacobi equation	115
4.17	Canonical transformations	116
4.18	Symplectic structure	118
4.19	Symplectic scalar product and Poisson brackets	120
4.20	(Super-)Hamiltonian as a generator of evolution	121
4.21	Infinitesimal canonical transformations	122
4.22	Constancy of phase-space volume under canonical transformations	122
4.23	Poisson algebra of integrals of motion	123
	<i>Concept Questions</i>	125
	<i>What's important?</i>	127
5	Observational Evidence for Black Holes	128
6	Ideal Black Holes	131
6.1	Definition of a black hole	131
6.2	Ideal black hole	132
6.3	No-hair theorem	132
7	Schwarzschild Black Hole	134
7.1	Schwarzschild metric	134
7.2	Stationary, static	135
7.3	Spherically symmetric	136
7.4	Energy-momentum tensor	137
7.5	Birkhoff's theorem	137
7.6	Horizon	138
7.7	Proper time	139
7.8	Redshift	140
7.9	"Schwarzschild singularity"	140
7.10	Weyl tensor	141
7.11	Singularity	141

7.12	Gullstrand-Painlevé metric	143
7.13	Embedding diagram	152
7.14	Schwarzschild spacetime diagram	153
7.15	Gullstrand-Painlevé spacetime diagram	154
7.16	Eddington-Finkelstein spacetime diagram	155
7.17	Kruskal-Szekeres spacetime diagram	156
7.18	Antihorizon	158
7.19	Analytically extended Schwarzschild geometry	158
7.20	Penrose diagrams	161
7.21	Penrose diagrams as guides to spacetime	162
7.22	Future and past horizons	165
7.23	Oppenheimer-Snyder collapse to a black hole	165
7.24	Apparent horizon	166
7.25	True horizon	167
7.26	Penrose diagrams of Oppenheimer-Snyder collapse	168
7.27	Illusory horizon	169
7.28	Collapse of a shell of matter on to a black hole	171
7.29	The illusory horizon and black hole thermodynamics	173
7.30	Rindler space and Rindler horizons	174
7.31	Rindler observers who start at rest, then accelerate	176
7.32	Killing vectors	181
7.33	Killing tensors	185
7.34	Lie derivative	185
8	Reissner-Nordström Black Hole	192
8.1	Reissner-Nordström metric	192
8.2	Energy-momentum tensor	193
8.3	Weyl tensor	194
8.4	Horizons	194
8.5	Gullstrand-Painlevé metric	194
8.6	Radial null geodesics	196
8.7	Finkelstein coordinates	197
8.8	Kruskal-Szekeres coordinates	198
8.9	Analytically extended Reissner-Nordström geometry	201
8.10	Penrose diagram	201
8.11	Antiverse: Reissner-Nordström geometry with negative mass	203
8.12	Outgoing, ingoing	203
8.13	The inflationary instability	204
8.14	The X point	206
8.15	Extremal Reissner-Nordström geometry	206

8.16	Super-extremal Reissner-Nordström geometry	207
8.17	Reissner-Nordström geometry with imaginary charge	209
9	Kerr-Newman Black Hole	212
9.1	Boyer-Lindquist metric	212
9.2	Oblate spheroidal coordinates	213
9.3	Time and rotation symmetries	215
9.4	Ring singularity	215
9.5	Horizons	215
9.6	Angular velocity of the horizon	217
9.7	Ergospheres	217
9.8	Turnaround radius	217
9.9	Antiverse	218
9.10	Sisytube	218
9.11	Extremal Kerr-Newman geometry	218
9.12	Super-extremal Kerr-Newman geometry	220
9.13	Energy-momentum tensor	220
9.14	Weyl tensor	221
9.15	Electromagnetic field	221
9.16	Principal null congruences	221
9.17	Finkelstein coordinates	222
9.18	Doran coordinates	222
9.19	Penrose diagram	225
	<i>Concept Questions</i>	226
	<i>What's important?</i>	228
10	Homogeneous, Isotropic Cosmology	229
10.1	Observational basis	229
10.2	Cosmological Principle	236
10.3	Friedmann-Lemaître-Robertson-Walker metric	236
10.4	Spatial part of the FLRW metric: informal approach	236
10.5	Comoving coordinates	239
10.6	Spatial part of the FLRW metric: more formal approach	240
10.7	FLRW metric	241
10.8	Einstein equations for FLRW metric	241
10.9	Newtonian “derivation” of Friedmann equations	242
10.10	Hubble parameter	244
10.11	Critical density	245
10.12	Omega	245
10.13	Types of mass-energy	247
10.14	Redshifting	249

10.15	Evolution of the cosmic scale factor	251
10.16	Age of the Universe	251
10.17	Conformal time	253
10.18	Looking back along the lightcone	254
10.19	Hubble diagram	256
10.20	Recombination	257
10.21	Horizon	258
10.22	Inflation	259
10.23	Evolution of the size and density of the Universe	262
10.24	Evolution of the temperature of the Universe	264
10.25	Neutrino mass	268
10.26	Occupation number, number density, and energy-momentum	271
10.27	Occupation numbers in thermodynamic equilibrium	273
10.28	Maximally symmetric spaces	281
	 PART TWO TETRAD APPROACH TO GENERAL RELATIVITY	 293
	<i>Concept Questions</i>	295
	<i>What's important?</i>	297
11	The tetrad formalism	298
11.1	Tetrad	298
11.2	Vierbein	299
11.3	The line-element encodes the vierbein	299
11.4	Tetrad transformations	300
11.5	Tetrad vectors and tensors	301
11.6	Index and naming conventions for vectors and tensors	303
11.7	Gauge transformations	304
11.8	Directed derivatives	304
11.9	Tetrad covariant derivative	305
11.10	Relation between tetrad and coordinate connections	307
11.11	Antisymmetry of the tetrad connections	307
11.12	Torsion tensor	308
11.13	No-torsion condition	308
11.14	Tetrad connections in terms of the vierbein	308
11.15	Torsion-free covariant derivative	309
11.16	Riemann curvature tensor	310
11.17	Ricci, Einstein, Bianchi	313
11.18	Expressions with torsion	313
11.19	General relativity in two spacetime dimensions	314

12	Spin and Newman-Penrose tetrads	319
12.1	Spin tetrad formalism	319
12.2	Newman-Penrose tetrad formalism	322
12.3	Weyl tensor	324
12.4	Petrov classification of the Weyl tensor	327
13	The geometric algebra	330
13.1	Products of vectors	331
13.2	Geometric product	332
13.3	Reverse	334
13.4	The pseudoscalar and the Hodge dual	334
13.5	General products of multivectors	336
13.6	Reflection	338
13.7	Rotation	339
13.8	Rotor group	342
13.9	Active and passive rotations	346
13.10	A rotor is a spin- $\frac{1}{2}$ object	346
13.11	2D rotations and complex numbers	347
13.12	Quaternions	348
13.13	3D rotations and quaternions	350
13.14	Pauli matrices	351
13.15	Pauli spinors as quaternions, or scaled rotors	353
13.16	Spin axis	355
14	The spacetime algebra	357
14.1	Spacetime algebra	357
14.2	Complex quaternions	359
14.3	Lorentz transformations and complex quaternions	361
14.4	Spatial inversion (P) and Time reversal (T)	364
14.5	How to implement Lorentz transformations on a computer	366
14.6	Killing vector fields of Minkowski space	371
14.7	Dirac matrices	375
14.8	Dirac spinors	376
14.9	Dirac spinors as complex quaternions	377
14.10	Non-null Dirac spinor	382
14.11	Null Dirac Spinor	384
15	Geometric Differentiation and Integration	388
15.1	Covariant derivative of a multivector	389
15.2	Riemann tensor of bivectors	391
15.3	Torsion tensor of vectors	392
15.4	Covariant spacetime derivative	392

15.5	Torsion-full and torsion-free covariant spacetime derivative	394
15.6	Differential forms	395
15.7	Differential forms in an arbitrary frame	396
15.8	Wedge product of differential forms	397
15.9	Exterior derivative	397
15.10	Hodge dual form	399
15.11	Relation between coordinate- and tetrad-frame volume elements	401
15.12	Generalized Stokes' theorem	402
15.13	Exact and closed forms	404
15.14	Generalized Gauss' theorem	405
15.15	Dirac delta-function	406
15.16	Integration of multivector-valued forms	407
15.17	Covariant derivative of a spinor	413
15.18	Covariant derivative in a spinor basis	414
15.19	Covariant spacetime derivative of a spinor	415
15.20	Gauss' theorem for spinors	416
16	Action principle for electromagnetism and gravity	417
16.1	Euler-Lagrange equations for a generic field	419
16.2	Super-Hamiltonian formalism	420
16.3	Conventional Hamiltonian formalism	421
16.4	Symmetries and conservation laws	421
16.5	Electromagnetic action	422
16.6	Electromagnetic action in forms notation	429
16.7	Gravitational action	436
16.8	Variation of the gravitational action	439
16.9	Trading coordinates and momenta	441
16.10	Matter energy-momentum and the Einstein equations with matter	442
16.11	Spin angular-momentum	443
16.12	Lagrangian as opposed to Hamiltonian formulation	450
16.13	Gravitational action in multivector notation	452
16.14	Gravitational action in multivector forms notation	456
16.15	Space+time (3+1) split in multivector forms notation	473
16.16	ADM gauge condition	482
16.17	Bianchi identities in multivector forms notation	489
17	Conventional Hamiltonian (3+1) approach	495
17.1	ADM formalism	496
17.2	ADM gravitational equations of motion	506
17.3	Conformally scaled ADM	512
17.4	Bianchi spacetimes	514

17.5	Friedmann-Lemaître-Robertson-Walker spacetimes	526
17.6	BKL oscillatory collapse	527
17.7	Numerical considerations	536
17.8	BSSN formalism	537
17.9	Pretorius formalism	540
17.10	$M+N$ split	542
18	Singularity theorems	545
18.1	Congruences	546
18.2	Raychaudhuri equations	547
18.3	Raychaudhuri equations for a timelike geodesic congruence	548
18.4	Raychaudhuri equations for a null geodesic congruence	551
18.5	Sachs optical coefficients	553
18.6	Interior mass	554
18.7	Hypersurface-orthogonality for a timelike congruence	554
18.8	Hypersurface-orthogonality for a null congruence	557
18.9	Focusing theorems	560
18.10	Singularity theorems	562
	<i>Concept Questions</i>	565
	<i>What's important?</i>	566
19	Black hole waterfalls	567
19.1	Tetrads move through coordinates	567
19.2	Gullstrand-Painlevé waterfall	568
19.3	Boyer-Lindquist tetrad	574
19.4	Doran waterfall	576
20	General spherically symmetric spacetimes	582
20.1	Spherical spacetime	582
20.2	Spherical line-element	582
20.3	Rest diagonal line-element	585
20.4	Comoving diagonal line-element	585
20.5	Tetrad connections	586
20.6	Riemann, Einstein, and Weyl tensors	588
20.7	Einstein equations	589
20.8	Choose your frame	589
20.9	Interior mass	590
20.10	Energy-momentum conservation	591
20.11	Structure of the Einstein equations	593
20.12	Comparison to ADM (3+1) formulation	597
20.13	Spherical electromagnetic field	597
20.14	General relativistic stellar structure	598

20.15	Freely-falling dust without shell-crossing	600
20.16	Naked singularities in dust collapse	601
20.17	Thin spherical shells	605
20.18	Self-similar spherically symmetric spacetime	614
20.19	General cylindrically symmetric spacetimes	628
20.20	An infinite thin wall?	629
21	The interiors of accreting, spherical black holes	631
21.1	Boundary conditions, mass, charge, and equation of state	632
21.2	Black hole accreting a neutral relativistic plasma	635
21.3	Black hole accreting a charged relativistic plasma	636
21.4	Black hole accreting charged baryons and dark matter	637
21.5	The black hole collider	640
21.6	The mechanism of mass inflation	641
21.7	The far future?	644
21.8	Weak null singularity on the Cauchy horizon?	645
21.9	Black hole accreting a fluid with an ultrahard equation of state	649
21.10	Black hole accreting a conducting charged plasma	650
21.11	Holography	655
21.12	Weird stuff at the outer horizon?	655
22	Ideal rotating black holes	657
22.1	Separable geometries	657
22.2	Horizons	660
22.3	Conditions from Hamilton-Jacobi separability	660
22.4	Electrovac solutions from separation of Einstein's equations	663
22.5	Electrovac solutions of Maxwell's equations	667
22.6	Λ -Kerr-Newman boundary conditions	669
22.7	Taub-NUT geometry	672
23	Trajectories in ideal rotating black holes	677
23.1	Hamilton-Jacobi equation	677
23.2	Particle with magnetic charge	679
23.3	Killing vectors and Killing tensor	680
23.4	Turnaround	680
23.5	Constraints on the Hamilton-Jacobi parameters P_t and P_x	681
23.6	Principal null congruences	682
23.7	Carter integral $\mathcal{Q}$	683
23.8	Penrose process	692
23.9	Constant latitude trajectories in the Kerr-Newman geometry	692
23.10	Circular orbits in the Kerr-Newman geometry	693
23.11	General solution for circular orbits	695

23.12	Circular geodesics (orbits for particles with zero electric charge)	699
23.13	Null circular orbits	703
23.14	The silhouette of a black hole	705
23.15	Marginally stable circular orbits	705
23.16	Circular orbits at constant latitude in the Antiverse	706
23.17	Circular orbits at the horizon of an extremal black hole	707
23.18	Equatorial circular orbits in the Kerr geometry	709
23.19	Thin disk accretion	711
23.20	Circular orbits in the Reissner-Nordström geometry	714
23.21	Ray-tracing	715
23.22	Hypersurface-orthogonal congruences	725
23.23	The Doran congruence	732
23.24	Principal null congruences	733
23.25	Pretorius-Israel double-null congruence	735
24	The interiors of rotating black holes	738
24.1	Nonlinear evolution	738
24.2	Focussing along principal null directions	739
24.3	Conformally separable geometries	739
24.4	Conditions from conformal Hamilton-Jacobi separability	739
24.5	Tetrad-frame connections	739
24.6	Inevitability of mass inflation	740
24.7	The black hole collider	741
25	Black hole thermodynamics	746
	<i>Concept Questions</i>	749
	<i>What's important?</i>	751
26	Perturbations and gauge transformations	752
26.1	Notation for perturbations	752
26.2	Vierbein perturbation	752
26.3	Gauge transformations	753
26.4	Tetrad metric assumed constant	753
26.5	Perturbed coordinate metric	754
26.6	Tetrad gauge transformations	754
26.7	Coordinate gauge transformations	755
26.8	Scalar, vector, tensor decomposition of perturbations	757
27	Perturbations in a flat space background	761
27.1	Classification of vierbein perturbations	761
27.2	Metric, tetrad connections, and Einstein and Weyl tensors	764
27.3	Spin components of the Einstein tensor	766
27.4	Too many Einstein equations?	767

27.5	Action at a distance?	767
27.6	Comparison to electromagnetism	768
27.7	Harmonic gauge	772
27.8	Newtonian (Copernican) gauge	774
27.9	Synchronous gauge	775
27.10	Newtonian potential	777
27.11	Dragging of inertial frames	778
27.12	Quadrupole pressure	781
27.13	Gravitational waves	782
27.14	Energy-momentum carried by gravitational waves	785
	<i>Concept Questions</i>	790
28	An overview of cosmological perturbations	791
29	Cosmological perturbations in a flat FLRW background	796
29.1	Unperturbed line-element	796
29.2	Comoving Fourier modes	797
29.3	Classification of vierbein perturbations	797
29.4	Residual global gauge freedoms	799
29.5	Metric, tetrad connections, Einstein tensor, and energy-momentum tensor	801
29.6	Gauge choices	806
29.7	ADM gauge choices	807
29.8	Conformal Newtonian (Copernican) gauge	807
29.9	Energy-momentum carried by gravitational waves	812
29.10	Spatially flat gauge	815
29.11	Conformal synchronous gauge	815
30	Cosmological perturbations: a simplest set of assumptions	818
30.1	Perturbed FLRW line-element	819
30.2	Energy-momenta of perfect fluids	819
30.3	Entropy conservation at superhorizon scales	823
30.4	Unperturbed background	825
30.5	Generic behaviour of non-baryonic cold dark matter	826
30.6	Generic behaviour of radiation	827
30.7	Equations for the simplest set of assumptions	829
30.8	On the numerical computation of cosmological power spectra	832
30.9	Analytic solutions in various regimes	833
30.10	Superhorizon scales	834
30.11	Radiation-dominated, adiabatic initial conditions	837
30.12	Radiation-dominated, isocurvature initial conditions	841
30.13	Subhorizon scales	842
30.14	Fluctuations that enter the horizon during the matter-dominated epoch	845

30.15	Matter-dominated regime	847
30.16	Baryons post-recombination	849
30.17	Matter with dark energy	850
30.18	Matter with dark energy and curvature	851
30.19	Primordial power spectrum	854
30.20	Matter power spectrum	855
30.21	Nonlinear evolution of the matter power spectrum	859
30.22	Statistics of random fields	859
31	Non-equilibrium processes in the FLRW background	865
31.1	Conditions around the epoch of recombination	866
31.2	Overview of recombination	867
31.3	Energy levels and ionization state in thermodynamic equilibrium	868
31.4	Occupation numbers	872
31.5	Boltzmann equation	872
31.6	Collisions	874
31.7	Non-equilibrium recombination	876
31.8	Recombination: Peebles approximation	878
31.9	Recombination: Seager et al. approximation	884
31.10	Sobolev escape probability	886
32	Cosmological perturbations: the hydrodynamic approximation	888
32.1	Electron-photon (Thomson) scattering	890
32.2	Summary of equations in the hydrodynamic approximation	891
32.3	Standard cosmological parameters	896
32.4	The photon-baryon fluid in the tight-coupling approximation	898
32.5	WKB approximation	900
32.6	Including quadrupole pressure in the momentum conservation equation	901
32.7	Photon diffusion (Silk damping)	902
32.8	Viscous baryon drag damping	903
32.9	Photon-baryon wave equation with dissipation	904
32.10	Baryon loading	906
32.11	Neutrinos	907
33	Cosmological perturbations: Boltzmann treatment	908
33.1	Summary of equations in the Boltzmann treatment	910
33.2	Boltzmann equation in a perturbed FLRW geometry	913
33.3	Non-baryonic cold dark matter	916
33.4	Boltzmann equation for the temperature fluctuation	918
33.5	Spherical harmonics of the temperature fluctuation	920
33.6	The Boltzmann equation for massless particles	920
33.7	Energy-momentum tensor for massless particles	921

33.8	Nonrelativistic electron-photon (Thomson) scattering	921
33.9	The photon collision term for electron-photon scattering	922
33.10	Boltzmann equation for photons	925
33.11	Baryons	927
33.12	Boltzmann equation for relativistic neutrinos	928
33.13	Massive neutrinos	930
33.14	Appendix: Legendre polynomials	931
34	Fluctuations in the Cosmic Microwave Background	933
34.1	Radiative transfer of CMB photons	933
34.2	Harmonics of the CMB photon distribution	935
34.3	CMB in real space	945
34.4	Observing CMB power	950
34.5	Large-scale CMB fluctuations (Sachs-Wolfe effect)	950
34.6	Radiative transfer of neutrinos	952
34.7	Appendix: Integrals over spherical Bessel functions	955
35	Cosmological perturbations including polarization	956
35.1	Photon polarization	956
35.2	Photon density matrix	959
35.3	Temperature fluctuation for polarized photons	962
35.4	Summary of equations including polarization	963
35.5	Boltzmann equations for polarized photons	964
35.6	Spherical harmonics of the polarized photon distribution	964
35.7	Neutrino Boltzmann equations	968
35.8	Matter Boltzmann equations	969
35.9	Vector and tensor Einstein equations	969
35.10	Polarized Thomson scattering	970
35.11	Initial conditions for vector and tensor fluctuations	976
35.12	Appendix: Spin-weighted spherical harmonics	979
36	Polarization of the Cosmic Microwave Background	986
36.1	Radiative transfer of the polarized CMB	986
36.2	Harmonics of the polarized CMB photon distribution	987
36.3	Harmonics of the polarized CMB in real space	990
36.4	Polarized CMB power spectra	992
37	Gravitational lensing of the Cosmic Microwave Background	995
38	Inflation	996
38.1	Power spectrum of vacuum fluctuations	999
38.2	Generation of gravitational waves by inflation	1000
38.3	Slow-roll inflation	1003

38.4	Generation of scalar waves by inflation	1006

PART THREE SPINORS, GEOMETRIC ALGEBRAS, THE STANDARD MODEL, QUANTUM FIELD THEORY, AND BEYOND		1013
39	Representations of Lie groups	1015
39.1	Lie group and Lie algebra	1015
39.2	Classification of finitely generated simple Lie groups	1015
39.3	Self-dual compactification tori	1028
39.4	Simply-laced root systems	1033
40	The supergeometric algebra	1035
40.1	Spin basis vectors in 3D	1035
40.2	Spin weight	1036
40.3	Pauli representation of chiral basis vectors	1036
40.4	Basis spinors	1037
40.5	Pauli spinor	1038
40.6	Spinor metric	1038
40.7	Row basis spinors	1039
40.8	Inner products of basis spinors	1040
40.9	Lowering and raising spinor indices	1040
40.10	Outer products of basis spinors	1041
40.11	The 3D supergeometric algebra	1043
40.12	Conjugate Pauli spinor	1044
40.13	Scalar products of spinors and conjugate spinors	1045
40.14	Conjugate multivectors	1046
40.15	The supergeometric algebra in arbitrarily many spatial dimensions	1047
41	Super spacetime algebra	1067
41.1	Newman-Penrose formalism	1067
41.2	Chiral representation of γ -matrices	1069
41.3	Basis spinors	1070
41.4	Weyl spinors	1071
41.5	Dirac and Weyl spinors	1072
41.6	Spinor scalar product	1073
41.7	Super spacetime algebra	1076
41.8	Charge conjugation C	1080
41.9	Anticommutation of Dirac spinors	1085
41.10	Discrete transformations P, T	1086
41.11	The supergeometric algebra in arbitrarily many space and time dimensions	1088

42	Quantum field theory of a real scalar field	1095
42.1	Quanta	1097
42.2	Quantization of an elementary simple harmonic oscillator	1098
42.3	Real scalar field	1103
42.4	The Klein-Gordon propagator in Minkowski spacetime	1110
43	Quantization from the perspective of an accelerating observer	1116
43.1	Unruh radiation	1116
43.2	Plane wave seen by a uniformly accelerating observer	1119
44	Quantum field theory of Dirac spinors	1127
44.1	Dirac spinor field	1127
44.2	Dirac field with electromagnetism	1132
44.3	Particles and antiparticles	1134
44.4	Quantization of the Dirac spinor field	1136
44.5	The Dirac propagator	1142
45	Quantum field theory of the electromagnetic field	1146
45.1	Quantized electromagnetic field	1147
45.2	Electromagnetic propagator	1149
46	Quantum field theory of the gravitational field	1151
46.1	Quantized gravitational field	1153
46.2	Gravitational propagator	1155
47	Quantum electrodynamics	1156
47.1	Interactions	1156
47.2	Single-vertex interaction	1157
47.3	Two-vertex interaction	1159
47.4	Amplitudes and rates	1163
47.5	Strategy	1167
47.6	Spinor-photon scattering	1168
47.7	Cross-section	1173
47.8	Spinor-antispinor creation from and annihilation into photons	1175
47.9	Coulomb scattering	1179
47.10	Annihilation/creation of spinor-antispinor pairs	1187
47.11	Crossing symmetry and Mandelstam variables	1192
47.12	Ward-Takahashi identity	1196
47.13	Loops	1197
48	The Standard Model of Physics	1199
48.1	Patterns of the Standard Model	1200
48.2	Spin(10)	1202
48.3	Spin(10) gauge fields	1206

48.4	Masses of fundamental fermions	1207
48.5	Appendix: Gauge fields of subgroups of Spin(10)	1214
49	Unification of Cl(10) and the Dirac algebra in Cl(11,1)	1216
49.1	Striking features of the Spin(10) chart	1217
49.2	Spin(11,1)	1219
49.3	Spatial readjustment of geometry following grand symmetry breaking	1225
50	Symmetry breaking and the Higgs mechanism	1229
50.1	The minimal symmetry breaking chain	1230
50.2	Constraints on symmetry breaking	1231
50.3	Electroweak symmetry breaking	1231
50.4	Grand symmetry breaking	1236
50.5	SU(8) symmetry breaking	1240
50.6	Pati-Salam symmetry breaking	1241
50.7	Energy scales of unification	1245
50.8	Appendix: Embedding the Pati-Salam group in $SU(8)_R \times SU(8)_L$	1250
50.9	Appendix: Embedding Spin(10) in SU(16)	1252
	<i>Bibliography</i>	1253

Illustrations

1.1	Vermilion emits a flash of light	13
1.2	Spacetime diagram	14
1.3	Spacetime diagram of Vermilion emitting a flash of light	14
1.4	Cerulean's spacetime is skewed compared to Vermilion's	15
1.5	Distances perpendicular to the direction of motion are unchanged	16
1.6	Vermilion defines hypersurfaces of simultaneity	17
1.7	Cerulean defines hypersurfaces of simultaneity similarly	18
1.8	Light clocks	19
1.9	Spacetime diagram illustrating the construction of hypersurfaces of simultaneity	20
1.10	Time dilation and Lorentz contraction spacetime diagrams	23
1.11	A cube: are the lengths of its sides all equal?	24
1.12	Twin paradox spacetime diagram	25
1.13	Wheel	26
1.14	Spacetime wheel	27
1.15	The right quadrant of the spacetime wheel represents uniformly accelerating observers	29
1.16	Spacetime diagram illustrating timelike, lightlike, and spacelike intervals	32
1.17	The longest proper time between two events is a straight line	35
1.18	Superluminal motion of the M87 jet	40
1.19	The rules of 4-dimensional perspective	42
1.20	Tachyon spacetime diagram	46
2.1	The principle of equivalence implies that gravity curves spacetime	52
2.2	A 2-sphere must be covered with at least two charts	54
2.3	The principle of equivalence implies the gravitational redshift and the gravitational bending of light	56
2.4	Tetrad vectors γ_m and tangent vectors e_μ	58
2.5	Derivatives of tangent vectors e_μ defined by parallel transport	64
2.6	Shapiro time delay	88
2.7	Lensing diagram	89
2.8	The appearance of a source lensed by a point lens	90
4.1	Action principle	97

4.2	Rindler wedge	107
5.1	The M87 black hole imaged by the Event Horizon Telescope	129
6.1	Fishes in a black hole waterfall	131
7.1	The singularity is not a point	142
7.2	Waterfall model of a Schwarzschild black hole	143
7.3	A body cannot remain rigid as it approaches the Schwarzschild singularity	149
7.4	Embedding diagram of the Schwarzschild geometry	152
7.5	Schwarzschild spacetime diagram	154
7.6	Gullstrand-Painlevé spacetime diagram	155
7.7	Finkelstein spacetime diagram	156
7.8	Kruskal-Szekeres spacetime diagram	157
7.9	Morph Finkelstein to Kruskal-Szekeres spacetime diagram	158
7.10	Embedding diagram of the analytically extended Schwarzschild geometry	159
7.11	Analytically extended Kruskal-Szekeres spacetime diagram	159
7.12	Sequence of embedding diagrams of the analytically extended Schwarzschild geometry	160
7.13	Penrose spacetime diagram	162
7.14	Morph Kruskal-Szekeres to Penrose spacetime diagram	162
7.15	Penrose spacetime diagram of the analytically extended Schwarzschild geometry	163
7.16	Penrose diagram of the Schwarzschild geometry	163
7.17	Penrose diagram of the analytically extended Schwarzschild geometry	164
7.18	Oppenheimer-Snyder collapse of a pressureless star	166
7.19	Spacetime diagrams of Oppenheimer-Snyder collapse	167
7.20	Penrose diagrams of Oppenheimer-Snyder collapse	168
7.21	Penrose diagram of a collapsed spherical star	169
7.22	Visualization of falling into a Schwarzschild black hole	170
7.23	Collapse of a shell on to a pre-existing black hole	172
7.24	Finkelstein spacetime diagram of a shell collapsing on to a pre-existing black hole	173
7.25	Rindler diagram	174
7.26	Penrose Rindler diagram	176
7.27	Spacetime diagram of Minkowski space showing observers who start at rest and then accelerate	177
7.28	Formation of the Rindler illusory horizon	177
7.29	Penrose diagram of Rindler space	178
7.30	Killing vector field on a 2-sphere	182
8.1	Waterfall model of a Reissner-Nordström black hole	195
8.2	Spacetime diagram of the Reissner-Nordström geometry	197
8.3	Finkelstein spacetime diagram of the Reissner-Nordström geometry	198
8.4	Kruskal spacetime diagram of the Reissner-Nordström geometry	199
8.5	Kruskal spacetime diagram of the analytically extended Reissner-Nordström geometry	200
8.6	Penrose diagram of the Reissner-Nordström geometry	202
8.7	Penrose diagram illustrating why the Reissner-Nordström geometry is subject to the inflationary instability	204
8.8	Waterfall model of an extremal Reissner-Nordström black hole	207
8.9	Penrose diagram of the extremal Reissner-Nordström geometry	208

8.10	Waterfall model of an extremal Reissner-Nordström black hole	209
8.11	Penrose diagram of the Reissner-Nordström geometry with imaginary charge Q	210
9.1	Geometry of Kerr and Kerr-Newman black holes	214
9.2	Contours of constant ρ , and their normals, in Boyer-Lindquist coordinates	216
9.3	Geometry of extremal Kerr and Kerr-Newman black holes	219
9.4	Geometry of a super-extremal Kerr black hole	220
9.5	Waterfall model of a Kerr black hole	223
9.6	Penrose diagram of the Kerr-Newman geometry	224
10.1	Hubble diagram of Type Ia supernovae	230
10.2	Spectrum of the CMB	232
10.3	Power spectrum of fluctuations in the CMB	233
10.4	Compilation of CMB TT , EE , BB , TE , and lensing power spectra	234
10.5	Power spectrum of galaxies	235
10.6	Embedding diagram of the FLRW geometry	237
10.7	Poincaré disk	241
10.8	Newtonian picture of the Universe as a uniform density gravitating ball	242
10.9	Behaviour of the mass-energy density of various species as a function of cosmic time	248
10.10	Cosmic scale factor as a function of time in universes with various Ω_m and Ω_Λ	252
10.11	Distance versus redshift in the FLRW geometry	255
10.12	Spacetime diagram of an FRLW Universe	259
10.13	Evolution of redshifts of objects at fixed comoving distance	260
10.14	Cosmic scale factor and Hubble distance as a function of cosmic time	263
10.15	Mass-energy density of the Universe as a function of cosmic time	264
10.16	Temperature of the Universe as a function of cosmic time	265
10.17	Comoving number densities during electron-positron annihilation	266
10.18	A massive fermion flips between left- and right-handed as it propagates through spacetime	270
10.19	Embedding spacetime diagram of de Sitter space	283
10.20	Penrose diagram of de Sitter space	285
10.21	Embedding spacetime diagram of anti de Sitter space	287
10.22	Penrose diagram of anti de Sitter space	289
11.1	Tetrad vectors γ_m	298
11.2	Derivatives of tetrad vectors γ_m defined by parallel transport	306
13.1	Vectors, bivectors, and trivectors	331
13.2	Reflection of a vector through an axis	338
13.3	Rotation of a vector by a bivector	340
13.4	Right-handed rotation of a vector by angle θ	341
14.1	Lorentz boost of a vector by rapidity θ	363
14.2	Killing trajectories in Minkowski space	373
15.1	Partition of unity	403
17.1	Cosmic scale factors in BKL collapse	531
17.2	Cosmic scale factors in BKL collapse	535
18.1	Expansion, vorticity, and shear	553
18.2	Formation of caustics in a hypersurface-orthogonal timelike congruence	555

18.3	Caustics in the galaxy NGC 474	556
18.4	Formation of caustics in a hypersurface-orthogonal null congruence	559
18.5	Spacetime diagram of the dog-leg proposition	562
18.6	Null boundary of the future of a 2-dimensional spacelike surface	563
19.1	Waterfall model of Schwarzschild and Reissner-Nordström black holes	569
19.2	Waterfall model of a Kerr black hole	580
20.1	Spacetime diagram of a naked singularity in dust collapse	602
20.2	Effective potential of a shell between a bubble of vacuum and empty space	610
20.3	Effective potential of a magnetically charged shell enclosing a bubble of vacuum	613
21.1	Uncharged baryonic plasma falls into spherical black hole	636
21.2	Charged, nonconducting plasma falls into a spherical black hole	637
21.3	Charged baryonic matter and neutral dark matter fall into a spherical black hole	638
21.4	Smaller accretion rates provoke faster inflation	639
21.5	Collision rates in the black hole collider	641
21.6	Spacetime diagram illustrating qualitatively the three successive phases of mass inflation	642
21.7	Charged plasma with an ultrahard equation of state falls into a black hole	650
21.8	Charged plasma with near critical conductivity falls into a black hole, creating huge entropy inside the horizon	652
21.9	Accreting spherical charged black hole that creates even more entropy inside the horizon	653
21.10	Penrose diagram of entropy production inside a black hole	654
22.1	Geometry of a Kerr-NUT black hole with $c_\bullet = -1$	673
22.2	Geometry of a Kerr-NUT black hole with $c_\bullet = 0$	674
23.1	Silhouettes of Kerr black holes	677
23.2	Regions visible and invisible to an infaller at the inner horizon	684
23.3	Values of $1/P$ for circular orbits of a charged particle about a Kerr-Newman black hole	697
23.4	Location of stable and unstable circular orbits in the Kerr geometry	700
23.5	Location of stable and unstable circular orbits in the super-extremal Kerr geometry	701
23.6	Radii of null circular orbits for a Kerr black hole	704
23.7	Radii of marginally stable circular orbits for a Kerr black hole	706
23.8	Values of the Hamilton-Jacobi parameter P_t for circular orbits in the equatorial plane of a near-extremal Kerr black hole	709
23.9	Energy and angular momentum on the ISCO	710
23.10	Accretion efficiency of a Kerr black hole	711
23.11	Three-eyed ape	718
23.12	Outgoing and ingoing null coordinates	731
23.13	Pretorius-Israel double-null hypersurface-orthogonal congruence	736
24.1	Particles falling from infinite radius can be either outgoing or ingoing at the inner horizon only up to a maximum latitude	742
27.1	The two polarizations of gravitational waves	783
29.1	Evolution of the tensor potential h_{ab}	814
30.1	Evolution of dark matter and radiation in the simple model	830
30.2	Overdensities and velocities in the simple approximation	831
30.3	Regimes in the evolution of fluctuations	833

30.4	Superhorizon scales	834
30.5	Evolution of the scalar potential Φ at superhorizon scales	836
30.6	Radiation-dominated regime	837
30.7	Evolution of the potential Φ and the radiation monopole Θ_0	838
30.8	Evolution of the dark matter overdensity δ_c	839
30.9	Subhorizon scales	842
30.10	Growth of the dark matter overdensity δ_c through matter-radiation equality	844
30.11	Fluctuations that enter the horizon in the matter-dominated regime	846
30.12	Matter-dominated regime	847
30.13	Evolution of the potential Φ and the radiation monopole Θ_0 for long wavelength modes	848
30.14	Growth factor $g(a)$	852
30.15	Matter power spectrum for the simple model	858
31.1	Ion fractions in thermodynamic equilibrium	870
31.2	Recombination of Hydrogen	881
31.3	Departure coefficients in the recombination of Hydrogen	882
31.4	Recombination of Hydrogen and Helium	885
32.1	Overdensities and velocities in the hydrodynamic approximation	889
32.2	Photon and neutrino multipoles in the hydrodynamic approximation	890
32.3	Evolution of matter and photon monopoles in the hydrodynamic approximation	893
32.4	Matter power spectrum in the hydrodynamic approximation	894
33.1	Overdensities and velocities from a Boltzmann computation	908
33.2	Photon and neutrino multipoles	909
33.3	Evolution of matter and photon monopoles in a Boltzmann computation	911
33.4	Difference $\Psi - \Phi$ in scalar potentials	912
33.5	Matter power spectrum	913
34.1	Visibility function	935
34.2	Factors in the solution of the radiative transfer equation	937
34.3	ISW integrand	939
34.4	CMB transfer functions $T_\ell(\eta_0, k)$ for a selection of harmonics ℓ	940
34.4	(continued)	941
34.5	Thomson scattering source functions at recombination	943
34.6	CMB transfer functions in the rapid recombination approximation	944
34.7	CMB power spectrum in the rapid recombination approximation	948
34.8	Multipole contributions to the CMB power spectrum	949
35.1	Thomson scattering of polarized light	971
35.2	Angles between photon momentum, scattered photon momentum, and wavevector	973
38.1	CMB B -mode power spectrum	997
39.1	Predicted symmetry breaking scale of $\text{Spin}(5) \times \text{Spin}(6)$	1019
39.2	$\text{SU}(3)$ tile	1032
44.1	An antiparticle is a particle moving backwards in time	1129
47.1	Spinor-photon interaction	1158
47.2	Two-vertex spinor-photon interaction	1160
47.3	Two-vertex spinor-photon interaction with intermediate spinor going backwards in time	1161

47.4	Spinor-photon scattering in centre-of-mass and lab frames	1162
47.5	The amplitude is a sum over Feynman diagram	1165
47.6	Spinor-antispinor creation and annihilation	1176
47.7	Spinor-antispinor creation/annihilation in centre-of-mass and lab frames	1177
47.8	Coulomb scattering of spinors	1180
47.9	Coulomb scattering in centre-of-mass and lab frames	1181
47.10	Spinor-antispinor pair annihilation and creation	1187
47.11	Spinor-antispinor annihilation/creation in centre-of-mass and lab frames	1188
47.12	Loop in a Feynman diagram	1197
48.1	Chart of Spin(10) fermions	1204
48.2	Masses of fundamental fermions	1208
49.1	Chart of Spin(10) fermions	1220
50.1	Mexican hat potential of a Higgs field	1235
50.2	Running of coupling parameters with energy	1246

Tables

1.1	Trip across the Universe	29
10.1	Cosmic inventory	246
10.2	Properties of universes dominated by various species	247
10.3	Evolution of cosmic scale factor in universes dominated by various species	252
10.4	Effective entropy-weighted number of relativistic particle species	268
12.1	Petrov classification of the Weyl tensor	328
17.1	Classification of Bianchi spaces	519
17.2	Bianchi dreibein	520
17.3	Bianchi Killing vectors	521
17.4	Bianchi transformation matrix Ξ	522
23.1	Signs of P_t and P_x in various regions of the Kerr-Newman geometry	681
39.1	Spinor and multivector representations of Lie groups	1023
40.1	Symmetry of spinor metric	1050
40.2	Sign of $\gamma_a^\top \varepsilon = \pm \varepsilon \gamma_a$	1051
41.1	Symmetry of the conjugation operator	1091
48.1	Conserved charges in the Standard Model	1201
48.2	Masses of fundamental fermions	1208

Exercises and *Concept questions

1.1	*Does light move differently depending on who emits it?	13
1.2	Challenge problem: the paradox of the constancy of the speed of light	13
1.3	Pictorial derivation of the Lorentz transformation	20
1.4	3D model of the Lorentz transformation	20
1.5	Mathematical derivation of the Lorentz transformation	20
1.6	*Determinant of a Lorentz transformation	22
1.7	Time dilation	23
1.8	Lorentz contraction	23
1.9	*Is one side of a cube shorter than the other?	23
1.10	Twin paradox	24
1.11	*What breaks the symmetry between you and your twin?	26
1.12	Length of a particle accelerator that reaches the GUT or Planck scale	30
1.13	*Proper time, proper distance	33
1.14	Scalar product	34
1.15	The principle of longest proper time	34
1.16	Superluminal jets	40
1.17	The rules of 4-dimensional perspective	42
1.18	Circles on the sky	43
1.19	Lorentz transformation preserves angles on the sky	44
1.20	The aberration of starlight	44
1.21	*Apparent (affine) distance	44
1.22	Brightness of a star	45
1.23	Tachyons	46
2.1	The equivalence principle implies the gravitational redshift of light, Part 1	55
2.2	The equivalence principle implies the gravitational redshift of light, Part 2	56
2.3	*Does covariant differentiation commute with the metric?	66
2.4	*Parallel transport when torsion is present	67
2.5	*Does torsion vanish in reality?	70
2.6	*Can the metric be Minkowski in the presence of torsion?	70
2.7	Covariant curl and coordinate curl	70

2.8	Covariant divergence and coordinate divergence	71
2.9	*If torsion does not vanish, does torsion-free covariant differentiation commute with the metric?	72
2.10	Gravitational redshift in a stationary metric	75
2.11	Gravitational redshift in Rindler space	75
2.12	Gravitational redshift in a uniformly rotating space	76
2.13	*Can Minkowski space rotate?	76
2.14	Derivation of the Riemann tensor	77
2.15	Jacobi identity	81
2.16	Einstein tensor in 3 or more dimensions	82
2.17	Special and general relativistic corrections for clocks on satellites	84
2.18	Equations of motion in weak gravity	85
2.19	Deflection of light by the Sun	87
2.20	Shapiro time delay	87
2.21	Gravitational lensing	88
3.1	Weyl tensor in arbitrary dimensions	91
3.2	Number of components of the Riemann, Ricci, and Weyl tensors in arbitrary dimensions	92
3.3	Number of Bianchi identities	94
4.1	*Redundant time coordinates?	98
4.2	*Throw a clock up in the air	100
4.3	*Conventional Lagrangian	101
4.4	Geodesics in Rindler space	107
4.5	*Action vanishes along a null geodesic, but its gradient does not	115
4.6	Canonical transformations from a quadratic generating function	117
4.7	*How many integrals of motion can there be?	124
7.1	Schwarzschild metric in isotropic form	135
7.2	Derivation of the Schwarzschild metric	137
7.3	*Going forwards or backwards in time inside the horizon	139
7.4	*Is the singularity of a Schwarzschild black hole a point?	142
7.5	*Separation between infallers who fall in at different times	143
7.6	Geodesics in the Schwarzschild geometry	144
7.7	Null geodesics in the Schwarzschild geometry	146
7.8	Geodesics in the Schwarzschild geometry in arbitrary dimensions	147
7.9	General relativistic precession of Mercury	148
7.10	A body cannot remain rigid as it approaches the Schwarzschild singularity	149
7.11	Causal distance between infallers near the singularity	150
7.12	Maximum transverse velocity of a light signal inside the horizon	151
7.13	*Penrose diagram of Minkowski space	165
7.14	*Penrose diagram of a thin spherical shell collapsing on to a Schwarzschild black hole	172
7.15	*Spherical Rindler space	176
7.16	Rindler illusory horizon	179
7.17	Area of the Rindler horizon	180
7.18	Coordinates adapted to an observer accelerating in Minkowski space	180
7.19	*What use is a Lie derivative?	186

7.20	Equivalence of expressions for the Lie derivative	188
7.21	Commutator of Lie derivatives	188
7.22	Lie derivative of the metric	190
7.23	Lie derivative of the inverse metric	190
7.24	Lie derivative of the metric determinant	190
8.1	*Units of charge of a charged black hole	193
8.2	Blueshift of a photon crossing the inner horizon of a Reissner-Nordström black hole	206
10.1	Isotropic (Poincaré) form of the FLRW metric	240
10.2	FLRW equations in arbitrary dimensions	242
10.3	Omega in photons	247
10.4	Mass-energy in an FLRW Universe	249
10.5	*Mass of a ball of photons or of vacuum	249
10.6	Geodesics in the FLRW geometry	250
10.7	Age of an FLRW universe containing matter and vacuum	252
10.8	Age of an FLRW universe containing radiation and matter	253
10.9	Relation between conformal time and cosmic scale factor	254
10.10	Hubble diagram	257
10.11	Horizon size at recombination	261
10.12	The horizon problem	261
10.13	Relation between horizon and flatness problems	262
10.14	Distribution of non-interacting particles initially in thermodynamic equilibrium	274
10.15	The first law of thermodynamics with non-conserved particle number	275
10.16	Number, energy, pressure, and entropy of a relativistic ideal gas at zero chemical potential	275
10.17	A relation between thermodynamic integrals	276
10.18	Relativistic particles in the early Universe had approximately zero chemical potential	276
10.19	Entropy per particle	276
10.20	Photon temperature at high redshift versus today	277
10.21	Cosmic Neutrino Background	277
10.22	Abundance of electrons and positrons in thermodynamic equilibrium	278
10.23	Maximally symmetric spaces in arbitrary dimensions	282
10.24	Maximally symmetric spaces	290
10.25	*Milne Universe	291
10.26	*Stationary FLRW metrics with different curvature constants describe the same spacetime	291
11.1	*Schwarzschild vierbein	300
11.2	Generators of Lorentz transformations are antisymmetric	301
11.3	*Does a locally inertial frame remain locally inertial under parallel transport?	307
11.4	Riemann tensor	310
11.5	Antisymmetry of the Riemann tensor	311
11.6	Cyclic symmetry of the Riemann tensor	311
11.7	Symmetry of the Riemann tensor	311
11.8	Number of components of the Riemann tensor	311
11.9	*Must connections vanish if Riemann vanishes?	312
11.10	Black holes in 2 spacetime dimensions?	316

11.11	Tidal forces falling into a Schwarzschild black hole	316
11.12	Totally antisymmetric tensor	317
13.1	Schur's lemma	335
13.2	*How fast do bivectors rotate?	342
13.3	*What is the dimension of the rotor group in N dimensions?	343
13.4	*Is the rotor group the same as the group of even, unimodular elements of the geometric algebra?	343
13.5	The even geometric algebra in $N+1$ dimensions is isomorphic to the full geometric algebra in N dimensions	343
13.6	Lie groups generated by multivectors	343
13.7	Rotation of a vector	348
13.8	3D rotation matrices	351
13.9	*Properties of Pauli matrices	352
13.10	Translate a rotor into an element of $SU(2)$	353
13.11	Translate a Pauli spinor into a quaternion	354
13.12	Translate a quaternion into a Pauli spinor	355
13.13	Can a Pauli spinor be rotated into its complex conjugate?	355
13.14	Orthonormal eigenvectors of the spin operator	356
14.1	Null complex quaternions	360
14.2	Nilpotent complex quaternions	360
14.3	Lorentz boost	363
14.4	Factor a Lorentz rotor into a boost and a rotation	364
14.5	Topology of the group of Lorentz rotors	364
14.6	Interpolate a Lorentz transformation	369
14.7	Exponential and logarithm of a complex quaternion	369
14.8	Spline a Lorentz transformation	369
14.9	The wrong way to implement a Lorentz transformation	370
14.10	Transform a 4-vector into a desired frame	370
14.12	Translate a Dirac spinor into a complex quaternion	379
14.13	Translate a complex quaternion into a Dirac spinor	379
14.14	Pseudoscalar times a Dirac spinor	380
14.15	Relation between $\tilde{\psi}$ and $\psi^\dagger$	381
14.16	Translate a Dirac spinor into a pair of Pauli spinors	381
14.17	Is the group of Lorentz rotors isomorphic to $SU(4)$?	382
14.18	*Is $\tilde{\psi}\psi$ real or complex?	383
14.19	*Is $\psi\boldsymbol{\gamma}_m\psi$ a scalar or a 4-vector?	384
14.20	*The boost axis of a null spinor is Lorentz-invariant	385
14.21	*What makes Weyl spinors special?	387
15.1	*Commutator versus wedge product of multivectors	389
15.2	Leibniz rule for the covariant spacetime derivative	394
15.3	*Calculating with the totally antisymmetric tensor	401
15.4	Action principle for strings and branes in arbitrary dimensions	407
16.1	*What are the units of the electromagnetic action?	425
16.2	*What are the units of the gravitational action?	438

16.3	*Can the coordinate metric be Minkowski in the presence of torsion?	447
16.4	*What kinds of metric or vierbein admit torsion?	448
16.5	*Why the names matter energy-momentum and spin angular-momentum?	448
16.6	Energy-momentum and spin angular-momentum of the electromagnetic field	448
16.7	Energy-momentum and spin angular-momentum of a Dirac field	449
16.8	Electromagnetic field in the presence of torsion	449
16.9	Dirac spinor field in the presence of torsion	449
16.10	Commutation of multivector forms	458
16.11	*Scalar product of the interval form ϵ	459
16.12	Triple products involving products of the interval form ϵ	461
16.13	Lie derivative of a form	473
16.14	Gravitational equations in arbitrary spacetime dimensions	485
16.15	Volume of a ball and area of a sphere	488
16.16	Bianchi identities	493
17.1	*Does Nature pick out a preferred foliation of time?	498
17.2	Energy and momentum constraints	511
17.3	Geodesics in Bianchi spacetimes	521
17.4	Bianchi Type IX space with all equal axes is a 3-sphere	521
17.5	Kasner spacetime	530
17.6	Schwarzschild interior	530
17.7	Kasner spacetime for a perfect fluid	531
17.8	Oscillatory Belinskii-Khalatnikov-Lifshitz (BKL) instability	533
18.1	Raychaudhuri equations for a non-geodesic timelike congruence	550
18.2	*Can vorticity be non-zero while shear vanishes?	554
18.3	*How do singularity theorems apply to the Kerr geometry?	564
18.4	*How do singularity theorems apply to the Reissner-Nordström geometry?	564
19.1	Tetrad frame of a rotating wheel	567
19.2	Coordinate transformation from Schwarzschild to Gullstrand-Painlevé	570
19.3	Velocity of a person who free-falls radially from rest	570
19.4	Dragging of inertial frames around a Kerr-Newman black hole	576
19.5	River model of the Friedmann-Lemaître-Robertson-Walker metric	581
20.1	Apparent horizon	584
20.2	Birkhoff's theorem	594
20.3	Spherically symmetric spacetimes in arbitrary dimensions	594
20.4	*Naked singularities in spherical spacetimes	595
20.5	Constant density star	599
20.6	Oppenheimer-Snyder collapse	601
20.7	Free fall of a thin, pressureless, spherical shell in vacuo	608
20.8	Self-similar line-element	617
21.1	A collisionless two-stream model of inflation	646
22.1	Explore other separable solutions	663
22.2	Explore separable solutions in an arbitrary number N of spacetime dimensions	663
23.1	Boundary of the region between the horizons visible to an infaller at the inner horizon	683

23.2	Near the Kerr-Newman singularity	687
23.3	When must t and ϕ progress forwards on a geodesic?	688
23.4	Inside the sisytube	689
23.5	Gödel's Universe	690
23.6	Negative energy trajectories outside the horizon	692
23.7	Circular geodesics at constant latitude?	694
23.8	*Are principal null geodesics circular orbits?	708
23.9	Icarus	712
23.10	Interstellar	713
23.11	What would you like to ray-trace?	725
23.12	Expansion, vorticity, and shear along the principal null congruences of the Λ -Kerr-Newman geometry	734
24.1	*Which Einstein equations are redundant?	741
24.2	Can accretion fuel outgoing and ingoing streams at the inner horizon?	741
24.3	Inflationary Kasner solution	743
25.1	Entropy in Hawking radiation	747
25.2	Area of the horizon	747
26.1	*What are the units of the vierbein perturbation φ_{mn} ?	753
26.2	*Non-infinitesimal tetrad transformations in perturbation theory?	755
26.3	*Should not the Lie derivative of a tetrad tensor be a tetrad tensor?	756
26.4	*Variation of unperturbed quantities under coordinate gauge transformations?	757
27.1	Classification of perturbations in arbitrary dimensions	763
27.2	*Are gauge-invariant potentials Lorentz-invariant?	769
27.3	*What parts of Maxwell's equations can be discarded?	771
27.4	Einstein tensor in harmonic gauge	774
27.5	*Independent evolution of scalar, vector, and tensor modes	775
27.6	Gravity Probe B and the geodetic and frame-dragging precession of gyroscopes	779
27.7	Scalar potentials outside a spherical body	782
27.8	*Units of the gravitational quadrupole radiation formula	785
27.9	Hulse-Taylor binary	786
29.1	*Global curvature as a perturbation?	801
29.2	*How can the primordial curvature fluctuation be constant outside the horizon, rather than redshifting to nothing?	806
29.3	Scalar, vector, tensor components of energy-momentum conservation	809
29.4	*Can the Universe at large rotate?	811
29.5	*Energy-momentum in gravitational waves	813
29.6	Evolution of tensor perturbations (gravitational waves) in FLRW spacetimes	813
29.7	*What frame does the CMB define?	816
29.8	*Are congruences of comoving observers in cosmology hypersurface-orthogonal?	816
30.1	Entropy perturbation	821
30.2	*Entropy perturbation when number is conserved	822
30.3	Vector fluctuation	822
30.4	Relation between entropy and the primordial curvature fluctuation ζ	823

30.5	*If the Friedmann equations enforce conservation of entropy, where does the entropy of the Universe come from?	824
30.6	*What is meant by the horizon in cosmology?	826
30.7	Redshift of matter-radiation equality	826
30.8	Generic behaviour of dark matter	827
30.9	Generic behaviour of radiation	828
30.10	*Can neutrinos be treated as a fluid?	828
30.11	Program the equations for the simplest set of cosmological assumptions	829
30.12	Radiation-dominated fluctuations	840
30.13	*Does the radiation monopole oscillate after recombination?	847
30.14	Growth of baryon fluctuations after recombination	849
30.15	*Curvature scale	852
30.16	Power spectrum of matter fluctuations: simple approximation	856
31.1	Proton and neutron fractions	866
31.2	*Level populations of hydrogen near recombination	871
31.3	*Ionization state of hydrogen near recombination	871
31.4	*Atomic structure notation	872
31.5	Detailed balance	876
31.6	Recombination	883
32.1	Thomson scattering rate	891
32.2	Program the equations in the hydrodynamic approximation	893
32.3	Power spectrum of matter fluctuations: hydrodynamic approximation	894
32.4	Effect of massive neutrinos on the matter power spectrum	895
32.5	Behaviour of radiation in the presence of damping	906
32.6	Diffusion scale	906
32.7	Generic behaviour of neutrinos	907
33.1	Program the Boltzmann equations	912
33.2	Power spectrum of matter fluctuations: Boltzmann treatment	912
33.3	Boltzmann equation factors in a general gauge	916
33.4	Moments of the non-baryonic cold dark matter Boltzmann equation	917
33.5	Initial conditions in the presence of neutrinos	929
34.1	CMB power spectrum in the instantaneous and rapid recombination approximations	947
34.2	CMB power spectra from CAMB	949
34.3	Cosmic Neutrino Background	954
35.1	*Relation of the polarization vector to the electromagnetic potential	958
35.2	*Elliptically polarized light	961
35.3	Boltzmann code including polarization	964
35.4	* E and B modes versus Stokes parameters	967
35.5	*Fluctuations with $ m \geq 3$?	968
35.6	*Are neutrinos polarized?	969
35.7	Photon diffusion including polarization	974
35.8	Generic behaviour of scalar, vector, and tensor fluctuations of neutrinos	977
35.9	Initial evolution of vector fluctuations of neutrinos	977

35.10	Initial evolution of tensor fluctuations of neutrinos	978
36.1	Neutrino harmonics including vectors and tensors	989
36.2	*Scalar, vector, tensor power spectra?	993
36.3	CMB polarized power spectrum	994
38.1	*Largest and smallest scales of inflation	998
38.2	*Did scales of galaxy clustering exit the horizon early or late during inflation?	998
38.3	Inflation driven by a scalar field with a Higgs-like potential	1005
39.1	Representations of $\text{Spin}(3)$	1024
39.2	$\text{Spin}(4)$ is isomorphic to $\text{SU}(2) \times \text{SU}(2)$	1026
39.3	$\text{SU}(4)$ is isomorphic to $\text{Spin}(6)$	1026
39.4	Representations of products of simple Lie groups	1026
40.1	Consistency of spinor and multivector scalar products	1042
40.2	*Imaginary spinor metric?	1046
40.3	Generalize the supergeometric algebra to an arbitrary number of dimensions	1048
41.1	*Lorentz transformation of the phase of a spinor	1072
41.2	Consistency of spinor and multivector scalar products	1078
41.3	*Chiral scalar	1078
41.4	Complex conjugate of a product of spinors and multivectors	1086
41.5	Generalize the super spacetime algebra to an arbitrary number of space and time dimensions	1088
42.1	*Is vacuum nothing, or is it something?	1101
42.2	*What are the units of the scalar field?	1104
43.1	*Does an observer in the uniform gravitational field of the Earth perceive Unruh radiation?	1121
43.2	*Does an observer inside a uniformly accelerating box perceive Unruh radiation?	1121
43.3	*Does an observer uniformly accelerating in a circle perceive Unruh radiation?	1121
43.4	Pair production seen by an observer moving along a Killing trajectory	1121
43.5	Plane wave seen by an observer moving in a circle	1123
43.6	Pair production seen by an adiabatically accelerating observer	1124
43.7	Pair production seen by an observer in de Sitter space	1126
46.1	Hilbert gravitational Lagrangian in perturbation theory	1152
47.1	Relation between numbers of vertices, lines, and loops in a Feynman diagram	1167
47.2	Numbers of loops in a Feynman diagram	1197

Legal notice

If you have obtained a copy of this proto-book from anywhere other than my website
<https://jila.colorado.edu/~ajsh/>

then that copy is illegal. This version of the proto-book is linked at
https://jila.colorado.edu/~ajsh/courses/ast5770_21/text.html

For the time being, this proto-book is free. Do not fall for third party scams that attempt to profit from this proto-book.

© A. J. S. Hamilton 2014-21

Notation

Except where actual units are needed, units are such that the speed of light is one, $c = 1$, and Newton's gravitational constant is one, $G = 1$.

The metric signature is $-+++$.

Greek (**brown**) letters $\kappa, \lambda, \dots$, denote spacetime (4D, usually) coordinate indices. Latin (black) letters $k, l, \dots$, denote spacetime (4D, usually) tetrad indices. Early-alphabet greek letters $\alpha, \beta, \dots$ denote spatial (3D, usually) coordinate indices. Early-alphabet latin letters $a, b, \dots$ denote spatial (3D, usually) tetrad indices. To avoid distraction, colouring is applied only to coordinate indices, not to the coordinates themselves. Early-alphabet latin letters $a, b, \dots$ are also used to denote spinor indices.

Sequences of indices, as encountered in multivectors (Chapter 13) and differential forms (Chapter 15), are denoted by capital letters. Greek (**brown**) capital letters $\Lambda, \Pi, \dots$ denote sequences of spacetime (4D, usually) coordinate indices. Latin (black) capital letters $K, L, \dots$ denote sequences of spacetime (4D, usually) tetrad indices. Early-alphabet capital letters denote sequences of spatial (3D, usually) indices, coloured brown $A, B, \dots$ for coordinate indices, and black $A, B, \dots$ for tetrad indices.

Specific (non-dummy) components of a vector are labelled by the corresponding coordinate (**brown**) or tetrad (black) direction, for example $A^\mu = \{A^t, A^x, A^y, A^z\}$ or $A^m = \{A^t, A^x, A^y, A^z\}$. Sometimes it is convenient to use numerical indices, as in $A^\mu = \{A^0, A^1, A^2, A^3\}$ or $A^m = \{A^0, A^1, A^2, A^3\}$. Allowing the same label to denote either a coordinate or a tetrad index risks ambiguity, but it should be apparent from the context (or colour) what is meant. Some texts distinguish coordinate and tetrad indices for example by a caret on the latter (there is no widespread convention), but this produces notational overload.

Boldface denotes abstract vectors, in either 3D or 4D. In 4D, $\mathbf{A} = A^\mu \mathbf{e}_\mu = A^m \boldsymbol{\gamma}_m$, where $\mathbf{e}_\mu$ denote coordinate tangent axes, and $\boldsymbol{\gamma}_m$ denote tetrad axes.

Repeated paired dummy indices are summed over, the implicit summation convention. In special and general relativity, one index of a pair must be up (contravariant), while the other must be down (covariant). If the space being considered is Euclidean, then both indices may be down.

$\partial/\partial x^\mu$ denotes coordinate partial derivatives, which commute. ∂_m denotes tetrad directed derivatives, which do not commute. D_μ and D_m denote respectively coordinate-frame and tetrad-frame covariant derivatives.

Choice of metric signature

There is a tendency, by no means unanimous, for general relativists to prefer the $-+++$ metric signature, while particle physicists prefer $+---$.

For someone like me who does general relativistic visualization, there is no contest: the choice has to be $-+++$, so that signs remain consistent between 3D spatial vectors and 4D spacetime vectors. For example, the 3D industry knows well that quaternions provide the most efficient and powerful way to implement spatial rotations. As shown in Chapter 13, complex quaternions provide the best way to implement Lorentz transformations, with the subgroup of real quaternions continuing to provide spatial rotations. Compatibility requires $-+++$. Actually, OpenGL and other graphics languages put spatial coordinates in the first three indices, leaving time to occupy the fourth index; but in these notes I stick to the physics convention of putting time in the zeroth index.

In practical calculations it is convenient to be able to switch transparently between boldface and index notation in both 3D and 4D contexts. This is where the $+---$ signature poses greater potential for misinterpretation in 3D. For example, with this signature, what is the sign of the 3D scalar product

$$\mathbf{a} \cdot \mathbf{b} ?$$

Is it $\mathbf{a} \cdot \mathbf{b} = \sum_{a=1}^3 a_a b^a$ or $\mathbf{a} \cdot \mathbf{b} = \sum_{a=1}^3 a^a b^a$? To be consistent with common 3D usage, it must be the latter. With the $+---$ signature, it must be that $\mathbf{a} \cdot \mathbf{b} = -a_a b^a$, where the repeated indices signify implicit summation over spatial indices. So you have to remember to introduce a minus sign in switching between boldface and index notation.

As another example, what is the sign of the 3D vector product

$$\mathbf{a} \times \mathbf{b} ?$$

Is it $\mathbf{a} \times \mathbf{b} = \sum_{b,c=1}^3 \varepsilon_{abc} a^b b^c$ or $\mathbf{a} \times \mathbf{b} = \sum_{b,c=1}^3 \varepsilon^a_{bc} a^b b^c$ or $\mathbf{a} \times \mathbf{b} = \sum_{b,c=1}^3 \varepsilon^{abc} a^b b^c$? Well, if you want to switch transparently between boldface and index notation, and you decide that you want boldface consistently to signify a vector with a raised index, then maybe you'd choose the middle option. To be consistent with standard 3D convention for the sign of the vector product, maybe you'd choose ε^a_{bc} to have positive sign for abc an even permutation of xyz .

Finally, what is the sign of the 3D spatial gradient operator

$$\nabla \equiv \frac{\partial}{\partial \mathbf{x}} ?$$

Is it $\nabla = \partial/\partial x^a$ or $\nabla = \partial/\partial x_a$? Convention dictates the former, in which case it must be that some boldface 3D vectors must signify a vector with a raised index, and others a vector with a lowered index. Oh dear.

PART ONE

FUNDAMENTALS

Concept Questions

1. What does c = universal constant mean? What is speed? What is distance? What is time?
2. $c + c = c$. How can that be possible?
3. The first postulate of special relativity asserts that spacetime forms a 4-dimensional continuum. The fourth postulate of special relativity asserts that spacetime has no absolute existence. Isn't that a contradiction?
4. The principle of special relativity says that there is no absolute spacetime, no absolute frame of reference with respect to which position and velocity are defined. Yet does not the cosmic microwave background define such a frame of reference?
5. How can two people moving relative to each other at near c both think each other's clock runs slow?
6. How can two people moving relative to each other at near c both think the other is Lorentz-contracted?
7. All paradoxes in special relativity have the same solution. In one word, what is that solution?
8. All conceptual paradoxes in special relativity can be understood by drawing what kind of diagram?
9. Your twin takes a trip to α Cen at near c , then returns to Earth at near c . Meeting your twin, you see that the twin has aged less than you. But from your twin's perspective, it was you that receded at near c , then returned at near c , so your twin thinks you aged less. Is it true?
10. Blobs in the jet of the galaxy M87 have been tracked by the Hubble Space Telescope to be moving at about $6c$. Does this violate special relativity?
11. If you watch an object move at near c , does it actually appear Lorentz-contracted? Explain.
12. You speed towards the centre of our Galaxy, the Milky Way, at near c . Does the centre appear to you closer or farther away?
13. You go on a trip to the centre of the Milky Way, 30,000 lightyears distant, at near c . How long does the trip take you?
14. You surf a light ray from a distant quasar to Earth. How much time does the trip take, from your perspective?
15. If light is a wave, what is waving?
16. As you surf the light ray, how fast does it appear to vibrate?
17. How does the phase of a light ray vary along the light ray? Draw surfaces of constant phase on a spacetime diagram.

18. You see a distant galaxy at a redshift of $z = 1$. If you could see a clock on the galaxy, how fast would the clock appear to tick? Could this be tested observationally?
19. You take a trip to α Cen at near c , then instantaneously accelerate to return at near c . If you are looking through a telescope at a clock on the Earth while you instantaneously accelerate, what do you see happen to the clock?
20. In what sense is time an imaginary spatial dimension?
21. In what sense is a Lorentz boost a rotation by an imaginary angle?
22. You know what it means for an object to be rotating at constant angular velocity. What does it mean for an object to be boosting at a constant rate?
23. A wheel is spinning so that its rim is moving at near c . The rim is Lorentz-contracted, but the spokes are not. How can that be?
24. You watch a wheel rotate at near the speed of light. The spokes appear bent. How can that be?
25. Does a sunbeam appear straight or bent when you pass by it at near the speed of light?
26. Energy and momentum are unified in special relativity. Explain.
27. In what sense is mass equivalent to energy in special relativity? In what sense is mass different from energy?
28. Why is the Minkowski metric unchanged by a Lorentz transformation?
29. What is the best way to program Lorentz transformations on a computer?

What's important?

1. The postulates of special relativity.
2. Understanding conceptually the unification of space and time implied by special relativity.
 - a. Spacetime diagrams.
 - b. Simultaneity.
 - c. Understanding the paradoxes of relativity — time dilation, Lorentz contraction, the twin paradox.
3. The mathematics of spacetime transformations.
 - a. Lorentz transformations.
 - b. Invariant spacetime distance.
 - c. Minkowski metric.
 - d. 4-vectors.
 - e. Energy-momentum 4-vector. $E = mc^2$.
 - f. The energy-momentum 4-vector of massless particles, such as photons.
4. What things look like at relativistic speeds.

Special Relativity

Special relativity is a fundamental building block of general relativity. General relativity postulates that the local structure of spacetime is that of special relativity.

The primary goal of this Chapter is to convey a clear conceptual understanding of special relativity. Everyday experience gives the impression that time is absolute, and that space is entirely distinct from time, as Galileo and Newton postulated. Special relativity demands, in apparent contradiction to experience, the revolutionary notion that space and time are united into a single 4-dimensional entity, called **spacetime**. The revolution forces conclusions that appear paradoxical: how can two people moving relative to each other both measure the speed of light to be the same, both think each other's clock runs slow, and both think the other is Lorentz-contracted?

In fact special relativity does not contradict everyday experience. It is just that we humans move through our world at speeds that are so much smaller than the speed of light that we are not aware of relativistic effects. The correctness of special relativity is confirmed every day in particle accelerators that smash particles together at highly relativistic speeds.

See <https://jila.colorado.edu/~ajsh/sr/> for animated versions of several of the diagrams in this Chapter.

1.1 Motivation

The history of the development of special relativity is rich and human, and it is beyond the intended scope of this book to give any reasonable account of it. If you are interested in the history, I recommend starting with the popular account by Thorne (1994).

As first proposed by James Clerk Maxwell in 1864, light is an electromagnetic wave. Maxwell believed (Goldman, 1984) that electromagnetic waves must be carried by some medium, the luminiferous aether, just as sound waves are carried by air. However, Maxwell knew that his equations of electromagnetism had empirical validity without any need for the hypothesis of an aether.

For Albert Einstein, the theory of special relativity was motivated by the curious circumstance that Maxwell's equations of electromagnetism seemed to imply that the speed of light was independent of the motion of an observer. Others before Einstein had noticed this curious feature of Maxwell's equations. Joseph

Larmor, Hendrick Lorentz, and Henri Poincaré all noticed that the form of Maxwell's equations could be preserved if lengths and times measured by an observer were somehow altered by motion through the aether. The transformations of special relativity were discovered before Einstein by Lorentz (1904), the name “Lorentz transformations” being conferred by Poincaré (1905).

Einstein's great contribution was to propose (Einstein, 1905) that there was no aether, no absolute space-time. From this simple and profound idea stemmed his theory of special relativity.

1.2 The postulates of special relativity

The theory of special relativity can be derived formally from a small number of postulates:

1. Space and time form a 4-dimensional continuum;
2. The existence of globally inertial frames;
3. The speed of light is constant;
4. The principle of special relativity.

The first two postulates are assertions about the structure of spacetime, while the last two postulates form the heart of special relativity. Most books mention just the last two postulates, but I think it is important to know that special (and general) relativity simply postulate the 4-dimensional character of spacetime, and that special relativity postulates moreover that spacetime is flat.

1. Space and time form a 4-dimensional continuum. The correct mathematical word for continuum is **manifold**. A 4-dimensional manifold is defined mathematically to be a topological space that is locally homeomorphic to Euclidean 4-space $\mathbb{R}^4$.

The postulate that spacetime forms a 4-dimensional continuum is a generalization of the classical Galilean concept that space and time form separate 3 and 1 dimensional continua. The postulate of a 4-dimensional spacetime continuum is retained in general relativity.

Physicists widely believe that this postulate must ultimately break down, that space and time are quantized over extremely small intervals of space and time, the Planck length $\sqrt{G\hbar/c^3} \approx 10^{-35}$ m, and the Planck time $\sqrt{G\hbar/c^5} \approx 10^{-43}$ s, where G is Newton's gravitational constant, $\hbar \equiv h/(2\pi)$ is Planck's constant divided by 2π , and c is the speed of light.

2. The existence of globally inertial frames. Statement: “There exist global spacetime frames with respect to which unaccelerated objects move in straight lines at constant velocity.”

A spacetime **frame** is a system of coordinates for labelling space and time. Four coordinates are needed, because spacetime is 4-dimensional. A frame in which unaccelerated objects move in straight lines at constant velocity is called an **inertial** frame. One can easily think of non-inertial frames: a rotating frame, an accelerating frame, or simply a frame with some bizarre Dahlian labelling of coordinates.

A **globally inertial** frame is an inertial frame that covers all of space and time. The postulate that globally inertial frames exist is carried over from classical mechanics (Newton's first law of motion).

Notice the subtle shift from the Newtonian perspective. The postulate is not that particles move in straight lines, but rather that there exist spacetime frames with respect to which particles move in straight lines.

Implicit in the assumption of the existence of globally inertial frames is the assumption that the geometry of spacetime is flat, the geometry of Euclid, where parallel lines remain parallel to infinity. In general relativity, this postulate is replaced by the weaker postulate that local (not global) inertial frames exist. A **locally inertial** frame is one which is inertial in a “small neighbourhood” of a spacetime point. In general relativity, spacetime can be curved.

3. The speed of light is constant. Statement: “The speed of light c is a universal constant, the same in any inertial frame.”

This postulate is the nub of special relativity. The immediate challenge of this Chapter, §1.3, is to confront its paradoxical implications, and to resolve them.

Measuring speed requires being able to measure intervals of both space and time: speed is distance travelled divided by time elapsed. Inertial frames constitute a special class of spacetime coordinate systems; it is with respect to distance and time intervals in these special frames that the speed of light is asserted to be constant.

In general relativity, arbitrarily weird coordinate systems are allowed, and light need move neither in straight lines nor at constant velocity with respect to bizarre coordinates (why should it, if the labelling of space and time is totally arbitrary?). However, general relativity asserts the existence of locally inertial frames, and the speed of light is a universal constant in those frames.

In 1983, the General Conference on Weights and Measures officially defined the speed of light to be

$$c \equiv 299,792,458 \text{ m s}^{-1}, \quad (1.1)$$

and the metre, instead of being a primary measure, became a secondary quantity, defined in terms of the second and the speed of light.

4. The principle of special relativity. Statement: “The laws of physics are the same in any inertial frame, regardless of position or velocity.”

Physically, this means that there is no absolute spacetime, no absolute frame of reference with respect to which position and velocity are defined. Only relative positions and velocities between objects are meaningful.

Mathematically, the principle of special relativity requires that the equations of special relativity be **Lorentz covariant**.

It is to be noted that the principle of special relativity does *not* imply the constancy of the speed of light, although the postulates are consistent with each other. Moreover the constancy of the speed of light does *not* imply the Principle of Special Relativity, although for Einstein the former appears to have been the inspiration for the latter.

An example of the application of the principle of special relativity is the construction of the energy-momentum 4-vector of a particle, which should have the same form in any inertial frame (§1.11).

1.3 The paradox of the constancy of the speed of light

The postulate that the speed of light is the same in any inertial frame leads immediately to a paradox. Resolution of this paradox compels a revolution in which space and time are united from separate 3 and 1-dimensional continua into a single 4-dimensional continuum.

Figure 1.1 shows Vermilion emitting a flash of light, which expands away from her in all directions. Vermilion thinks that the light moves outward at the same speed in all directions. So Vermilion thinks that she is at the centre of the expanding sphere of light.

Figure 1.1 shows also Cerulean, who is moving away from Vermilion at about half the speed of light. But, says special relativity, Cerulean also thinks that the light moves outward at the same speed in all directions from him. So Cerulean should be at the centre of the expanding light sphere too. But he's not, is he. Paradox!

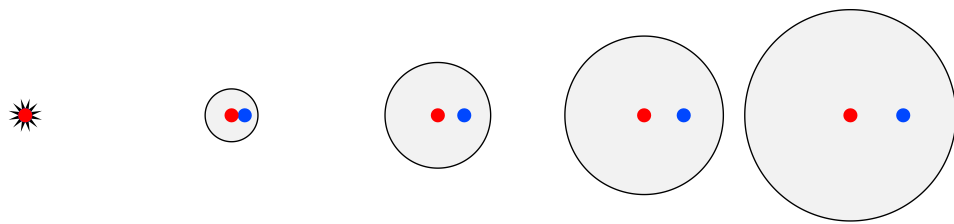


Figure 1.1 Vermilion emits a flash of light, which (from left to right) expands away from her in all directions. Since the speed of light is constant in all directions, she finds herself at the centre of the expanding sphere of light. Cerulean is moving to the right at half of the speed of light relative to Vermilion. Special relativity declares that Cerulean too thinks that the speed of light is constant in all directions. So should not Cerulean think that he too is at the centre of the expanding sphere of light? Paradox!

Concept question 1.1 Does light move differently depending on who emits it? Would the light have expanded differently if Cerulean had emitted the light?

Exercise 1.2 Challenge problem: the paradox of the constancy of the speed of light. Can you figure out a solution to the paradox? Somehow you have to arrange that both Vermilion and Cerulean regard themselves as being in the centre of the expanding sphere of light.

1.3.1 Spacetime diagram

A spacetime diagram suggests a way of thinking, first advocated by Minkowski (1909), that leads to the solution of the paradox of the constancy of the speed of light. Indeed, spacetime diagrams provide the way to resolve all conceptual paradoxes in special relativity, so it is thoroughly worthwhile to understand them.

A **spacetime diagram**, Figure 1.2, is a diagram in which the vertical axis represents time, while the horizontal axis represents space. Really there are three dimensions of space, which can be thought of as

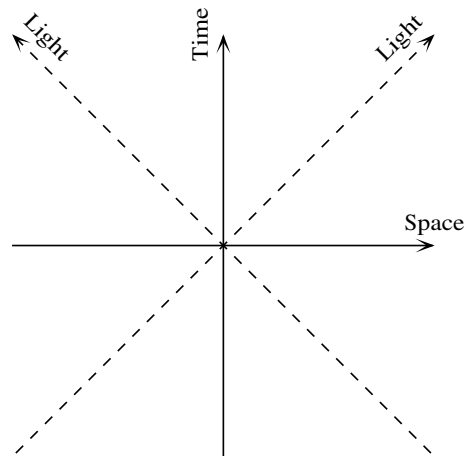


Figure 1.2 A spacetime diagram shows events in space and time. In a spacetime diagram, time goes upward, while space dimensions are horizontal. Really there should be 3 space dimensions, but usually it suffices to show 1 spatial dimension, as here. In a spacetime diagram, the units of space and time are chosen so that light goes one unit of distance in one unit of time, i.e. the units are such that the speed of light is one, $c = 1$. Thus light moves upward and outward at 45° from vertical in a spacetime diagram.

filling additional horizontal dimensions. But for simplicity a spacetime diagram usually shows just one spatial dimension.

In a spacetime diagram, the units of space and time are chosen so that light goes one unit of distance in one unit of time, i.e. the units are such that the speed of light is one, $c = 1$. Thus light always moves upward at 45° from vertical in a spacetime diagram. Each point in 4-dimensional spacetime is called an **event**. Light

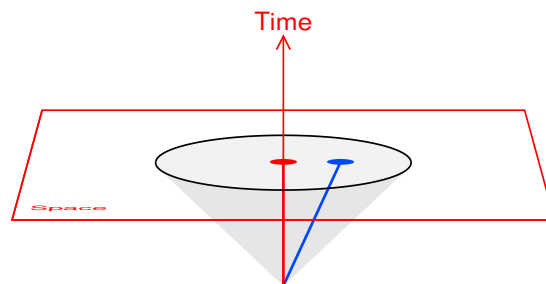


Figure 1.3 Spacetime diagram of Vermilion emitting a flash of light. This is a spacetime diagram version of the situation illustrated in Figure 1.1. The lines along which Vermilion and Cerulean move through spacetime are called their worldlines. Each point in 4-dimensional spacetime is called an event. Light signals converging to or expanding from an event follow a 3-dimensional hypersurface called the lightcone. In the diagram, the sphere of light expanding from the emission event is following the future lightcone. There is also a past lightcone, not shown here.

signals converging to or expanding from an event follow a 3-dimensional hypersurface called the **lightcone**. Light converging on to an event is on the **past lightcone**, while light emerging from an event is on the **future lightcone**.

Figure 1.3 shows a spacetime diagram of Vermilion emitting a flash of light, and Cerulean moving relative to Vermilion at about $\frac{1}{2}$ the speed of light. This is a spacetime diagram version of the situation illustrated in Figure 1.1. The lines along which Vermilion and Cerulean move through spacetime are called their **world-lines**.

Consider again the challenge problem. The problem is to arrange that both Vermilion and Cerulean are at the centre of the lightcone, from their own points of view.

Here's a clue. Cerulean's concept of space and time may not be the same as Vermilion's.

1.3.2 Centre of the lightcone

The solution to the paradox is that Cerulean's spacetime is skewed compared to Vermilion's, as illustrated by Figure 1.4. The thing to notice in the diagram is that Cerulean is in the centre of the lightcone, according to the way Cerulean perceives space and time. Vermilion remains at the centre of the lightcone according to the way Vermilion perceives space and time. In the diagram Vermilion and her space are drawn at one "tick" of her clock past the point of emission, and likewise Cerulean and his space are drawn at one "tick" of his identical clock past the point of emission. Of course, from Cerulean's point of view his spacetime is quite normal, and it is Vermilion's spacetime that is skewed.

In special relativity, the transformation between the spacetime frames of two inertial observers is called a

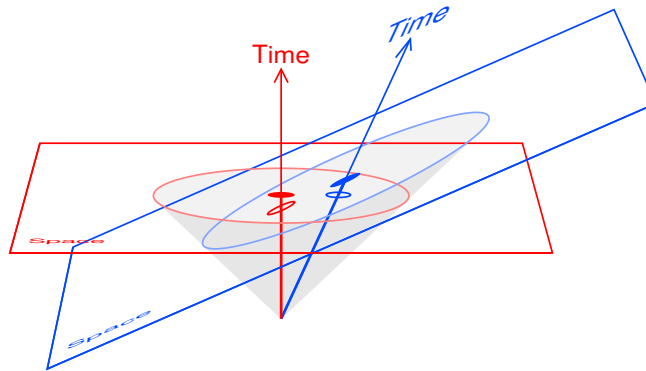


Figure 1.4 The solution to how both Vermilion and Cerulean can consider themselves to be at the centre of the lightcone. Cerulean's spacetime is skewed compared to Vermilion's. Cerulean is in the centre of the lightcone, according to the way Cerulean perceives space and time, while Vermilion remains at the centre of the lightcone according to the way Vermilion perceives space and time. In the diagram Vermilion (red) and her space are drawn at one "tick" of her clock past the point of emission, and likewise Cerulean (blue) and his space are drawn at one "tick" of his identical clock past the point of emission.

Lorentz transformation. In general, a Lorentz transformation consists of a spatial rotation about some spatial axis, combined with a **Lorentz boost** by some velocity in some direction.

Only space along the direction of motion gets skewed with time. Distances perpendicular to the direction of motion remain unchanged. Why must this be so? Consider two hoops which have the same size when at rest relative to each other. Now set the hoops moving towards each other. Which hoop passes inside the other? Neither! For suppose Vermilion thinks Cerulean's hoop passed inside hers; by symmetry, Cerulean must think Vermilion's hoop passed inside his; but both cannot be true; the only possibility is that the hoops remain the same size in directions perpendicular to the direction of motion.

If you have understood all this, then you have understood the crux of special relativity, and you can now go away and figure out all the mathematics of Lorentz transformations. The mathematical problem is: what is the relation between the spacetime coordinates $\{t, x, y, z\}$ and $\{t', x', y', z'\}$ of a spacetime interval, a 4-vector, in Vermilion's versus Cerulean's frames, if Cerulean is moving relative to Vermilion at velocity v in, say, the x direction? The solution follows from requiring

1. that both observers consider themselves to be at the centre of the lightcone, as illustrated by Figure 1.4, and
2. that distances perpendicular to the direction of motion remain unchanged, as illustrated by Figure 1.5.

An alternative version of the second condition is that a Lorentz transformation at velocity v followed by a Lorentz transformation at velocity $-v$ should yield the unit transformation.

Note that the postulate of the existence of globally inertial frames implies that Lorentz transformations are linear, that straight lines (4-vectors) in one inertial spacetime frame transform into straight lines in other inertial frames.

You will solve this problem in the next section but two, §1.6. As a prelude, the next two sections, §1.4 and §1.5 discuss simultaneity and time dilation.

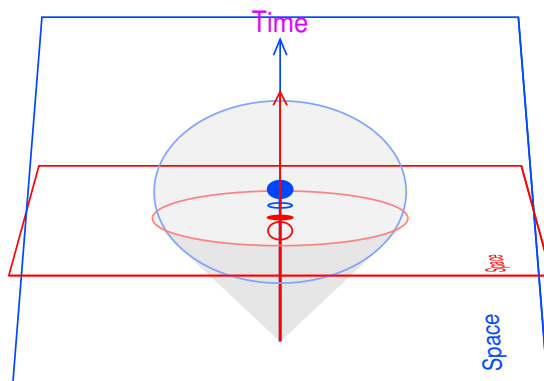


Figure 1.5 Same as Figure 1.4, but with Cerulean moving into the page instead of to the right. This is just Figure 1.4 spatially rotated by 90° in the horizontal plane. Distances perpendicular to the direction of motion are unchanged.

1.4 Simultaneity

Most (all?) of the apparent paradoxes of special relativity arise because observers moving at different velocities relative to each other have different notions of simultaneity.

1.4.1 Operational definition of simultaneity

How can simultaneity, the notion of events occurring at the same time at different places, be defined operationally?

One way is illustrated in the sequences of spacetime diagrams in Figure 1.6. Vermilion surrounds herself with a set of mirrors, equidistant from Vermilion. She sends out a flash of light, which reflects off the mirrors back to Vermilion. How does Vermilion know that the mirrors are all the same distance from her? Because the reflected flash from the mirrors arrives back to Vermilion all at the same instant. Vermilion asserts that the light flash must have hit all the mirrors simultaneously. Vermilion also asserts that the instant when the light hit the mirrors must have been the instant, as registered by her wristwatch, precisely half way between the moment she emitted the flash and the moment she received it back again. If it takes, say, 2 seconds between flash and receipt, then Vermilion concludes that the mirrors are 1 lightsecond away from her. The spatial hyperplane passing through these events is a hypersurface of simultaneity. More generally, from Vermilion's perspective, each horizontal hyperplane in the spacetime diagram is a hypersurface of simultaneity.

Cerulean defines surfaces of simultaneity using the same operational setup: he encompasses himself with mirrors, arranging them so that a flash of light returns from them to him all at the same instant. But whereas Cerulean concludes that his mirrors are all equidistant from him and that the light bounces off them all at the same instant, Vermilion thinks otherwise. From Vermilion's point of view, the light bounces off Cerulean's mirrors at different times and moreover at different distances from Cerulean, as illustrated in Figure 1.7. Only so can the speed of light be constant, as Vermilion sees it, and yet the light return to Cerulean all at the same instant.

Of course from Cerulean's point of view all is fine: he thinks his mirrors are equidistant from him, and

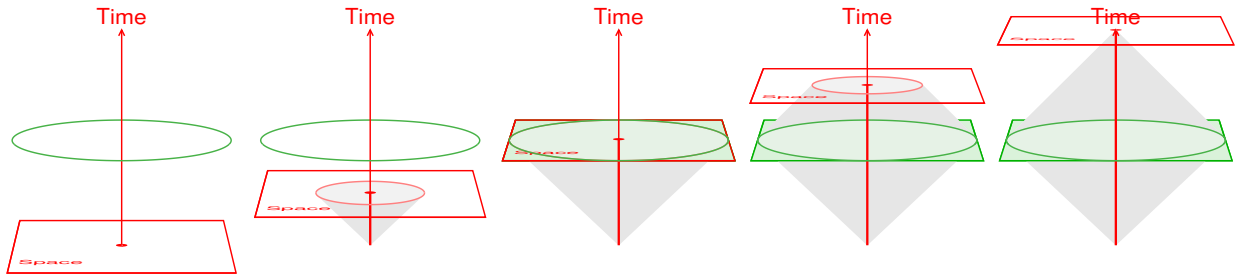


Figure 1.6 How Vermilion defines hypersurfaces of simultaneity. She surrounds herself with (green) mirrors all at the same distance. She sends out a light beam, which reflects off the mirrors, and returns to her all at the same moment. She knows that the mirrors are all at the same distance precisely because the light returns to her all at the same moment. The events where the light bounced off the mirrors defines a hypersurface of simultaneity for Vermilion.

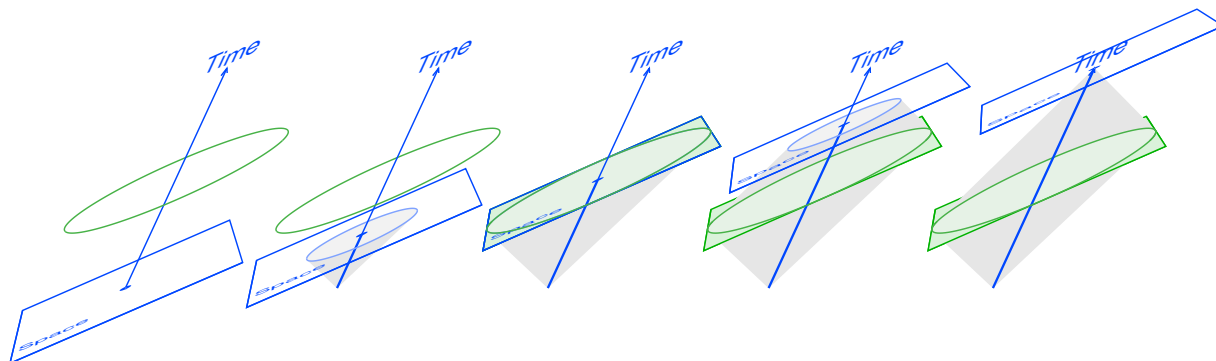


Figure 1.7 Cerulean defines hypersurfaces of simultaneity using the same operational setup as Vermilion: he bounces light off (green) mirrors all at the same distance from him, arranging them so that the light returns to him all at the same time. But from Vermilion's frame, Cerulean's experiment looks skewed, as shown here.

that the light bounces off them all at the same instant. The inevitable conclusion is that Cerulean must measure space and time along axes that are skewed relative to Vermilion's. Events that happen at the same time according to Cerulean happen at different times according to Vermilion; and vice versa. Cerulean's hypersurfaces of simultaneity are not the same as Vermilion's.

From Cerulean's point of view, Cerulean remains always at the centre of the lightcone. Thus for Cerulean, as for Vermilion, the speed of light is constant, the same in all directions.

1.5 Time dilation

Vermilion and Cerulean construct identical clocks, Figure 1.8, consisting of a light beam which bounces off a mirror. Tick, the light beam hits the mirror, tock, the beam returns to its owner. As long as Vermilion and Cerulean remain at rest relative to each other, both agree that each other's clock tick-tocks at the same rate as their own.

But now suppose Cerulean goes off at velocity v relative to Vermilion, in a direction perpendicular to the direction of the mirror. As far as Cerulean is concerned, his clock tick-tocks at the same rate as before, a tick at the mirror, a tock on return. But from Vermilion's point of view, although the distance between Cerulean and his mirror at any instant remains the same as before, the light has farther to go. And since the speed of light is constant, Vermilion thinks it takes longer for Cerulean's clock to tick-tock than her own. Thus Vermilion thinks Cerulean's clock runs slow relative to her own.

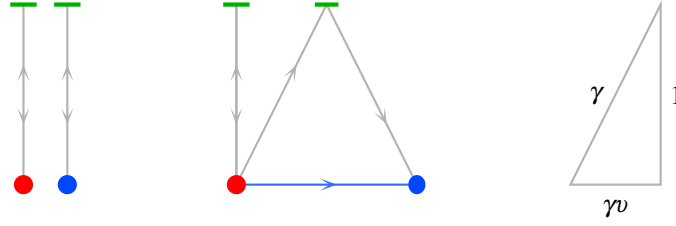


Figure 1.8 Vermilion and Cerulean construct identical clocks, consisting of a light beam that bounces off a (green) mirror and returns to them. In the left panel, Cerulean is at rest relative to Vermilion. They both agree that their clocks are identical. In the middle panel, Cerulean is moving to the right at speed v relative to Vermilion. The vertical distance to the mirror is unchanged by Cerulean's motion in a direction orthogonal to the direction to the mirror. Whereas Cerulean thinks his clock ticks at the usual rate, Vermilion sees the path of the light taken by Cerulean's clock is longer, by a factor γ , than the path of light taken by her own clock. Since the speed of light is constant, Vermilion thinks Cerulean's clock takes longer to tick, by a factor γ , than her own. The sides of the triangle formed by the distance 1 to the mirror, the length γ of the lightpath to Cerulean's clock, and the distance γv travelled by Cerulean, form a right-angled triangle, illustrated in the right panel.

1.5.1 Lorentz gamma factor

How much slower does Cerulean's clock run, from Vermilion's point of view? In special relativity the factor is called the **Lorentz gamma factor** γ , introduced by the Dutch physicist Hendrik A. Lorentz in 1904, one year before Einstein proposed his theory of special relativity.

In units where the speed of light is one, $c = 1$, Vermilion's mirror in Figure 1.8 is one tick away from her, and from her point of view the vertical distance between Cerulean and his mirror is the same, one tick. But Vermilion thinks that the distance travelled by the light beam between Cerulean and his mirror is γ ticks. Cerulean is moving at speed v , so Vermilion thinks he moves a distance of γv ticks during the γ ticks of time taken by the light to travel from Cerulean to his mirror. Thus, from Vermilion's point of view, the vertical line from Cerulean to his mirror, Cerulean's light beam, and Cerulean's path form a triangle with sides 1, γ , and γv , as illustrated in Figure 1.8. Pythagoras' theorem implies that

$$1^2 + (\gamma v)^2 = \gamma^2. \quad (1.2)$$

From this it follows that the Lorentz gamma factor γ is related to Cerulean's velocity v by

$$\gamma = \frac{1}{\sqrt{1 - v^2}}, \quad (1.3)$$

which is Lorentz's famous formula.

1.6 Lorentz transformation

A Lorentz transformation is a rotation of space and time. Lorentz transformations form a 6-dimensional group, with 3 dimensions from spatial rotations, and 3 dimensions from Lorentz boosts.

If you wish to understand special relativity mathematically, then it is essential for you to go through the exercise of deriving the form of Lorentz transformations for yourself. Indeed, this problem is the challenge problem posed in §1.3, recast as a mathematical exercise. For simplicity, it is enough to consider the case of a Lorentz boost by velocity v along the x -axis.

You can derive the form of a Lorentz transformation either pictorially (geometrically), or algebraically. Ideally you should do both.

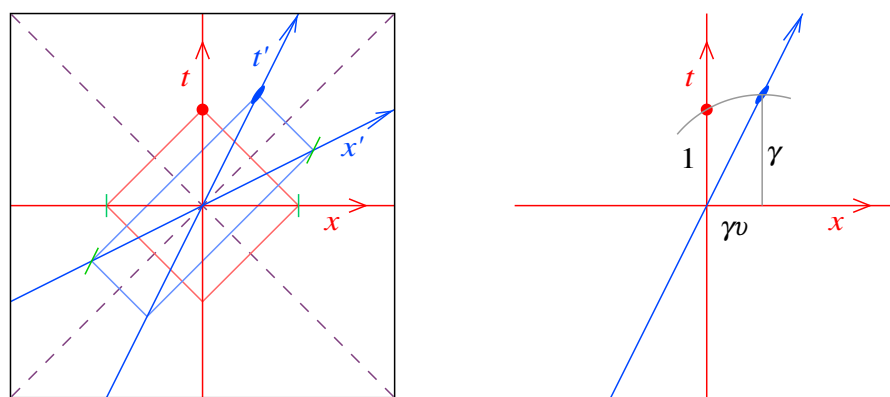


Figure 1.9 Spacetime diagram representing the experiments shown in Figures 1.6 and 1.7. The right panel shows a detail of how the spacetime diagram can be drawn using only a straight edge and a compass. If Cerulean's position is drawn first, then Vermilion's position follows from drawing the arc as shown.

Exercise 1.3 Pictorial derivation of the Lorentz transformation. Construct, with ruler and compass, a spacetime diagram that looks like the one in Figure 1.9. You should recognize that the square represents the paths of lighttrays that Vermilion uses to define a hypersurface of simultaneity, while the rectangle represents the same thing for Cerulean. Notice that Cerulean's worldline and line of simultaneity are diagonals along his light rectangle, so the angles between those lines and the lightcone are equal. Notice also that the areas of the square and the rectangle are the same, which expresses the fact that the area is multiplied by the determinant of the Lorentz transformation matrix, which must be one (why?). Use your geometric construction to derive the mathematical form of the Lorentz transformation.

Exercise 1.4 3D model of the Lorentz transformation. Make a 3D spacetime diagram of the Lorentz transformation, something like that in Figure 1.4, with not only an x -dimension, as in Exercise 1.3, but also a y -dimension. You can use a 3D computer modelling program, or you can make a real 3D model. Make the lightcone from flexible paperboard, the spatial hypersurface of simultaneity from stiff paperboard, and the worldline from wooden dowel.

Exercise 1.5 Mathematical derivation of the Lorentz transformation. Relative to person A (Vermilion, unprimed frame), person B (Cerulean, primed frame) moves at velocity v along the x -axis. Derive

the form of the Lorentz transformation between the coordinates (t, x, y, z) of a 4-vector in A's frame and the corresponding coordinates (t', x', y', z') in B's frame from the assumptions:

1. that the transformation is linear;
2. that the spatial coordinates in the directions orthogonal to the direction of motion are unchanged;
3. that the speed of light c is the same for both A and B, so that $x = t$ in A's frame transforms to $x' = t'$ in B's frame, and likewise $x = -t$ in A's frame transforms to $x' = -t'$ in B's frame;
4. the definition of speed; if B is moving at speed v relative to A, then $x = vt$ in A's frame transforms to $x' = 0$ in B's frame;
5. spatial isotropy; specifically, show that if A thinks B is moving at velocity v , then B must think that A is moving at velocity $-v$, and symmetry (spatial isotropy) between these two situations then fixes the Lorentz γ factor.

Your logic should be precise, and explained in clear, concise English.

You should find that the Lorentz transformation for a Lorentz boost by velocity v along the x -axis is

$$\begin{aligned} t' &= \gamma t - \gamma v x & t &= \gamma t' + \gamma v x' \\ x' &= -\gamma v t + \gamma x & x &= \gamma v t' + \gamma x' \\ y' &= y & y &= y' \\ z' &= z & z &= z' \end{aligned} \quad , \quad (1.4)$$

The transformation can be written more elegantly in matrix notation:

$$\begin{pmatrix} t' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix} , \quad (1.5)$$

with inverse

$$\begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \gamma & \gamma v & 0 & 0 \\ \gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} t' \\ x' \\ y' \\ z' \end{pmatrix} . \quad (1.6)$$

A Lorentz transformation at velocity v followed by a Lorentz transformation at velocity v in the opposite direction, i.e. at velocity $-v$, yields the unit transformation, as it should:

$$\begin{pmatrix} \gamma & \gamma v & 0 & 0 \\ \gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad (1.7)$$

The determinant of the Lorentz transformation is one, as it should be:

$$\begin{vmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} = \gamma^2(1 - v^2) = 1 . \quad (1.8)$$

Indeed, requiring that the determinant be one provides another derivation of the formula (1.3) for the Lorentz gamma factor.

Concept question 1.6 Determinant of a Lorentz transformation. Why must the determinant of a Lorentz transformation be one?

1.7 Paradoxes: Time dilation, Lorentz contraction, and the Twin paradox

There are several classic paradoxes in special relativity. One of them has already been met above, the paradox of the constancy of the speed of light in §1.3. This section collects three famous paradoxes: time dilation, Lorentz contraction, and the Twin paradox.

If you wish to understand special relativity conceptually, then you should work through all these paradoxes yourself. As remarked in §1.4, most (all?) paradoxes in special relativity arise because different observers have different notions of simultaneity, and most (all?) paradoxes can be solved using spacetime diagrams.

The Twin paradox is particularly helpful because it illustrates several different facets of special relativity, not only time dilation, but also how light travel time modifies what an observer actually sees.

1.7.1 Time dilation

If a timelike interval $\{t, r\}$ corresponds to motion at velocity v , then $r = vt$. The proper time along the interval is

$$\tau = \sqrt{t^2 - r^2} = t\sqrt{1 - v^2} = \frac{t}{\gamma} . \quad (1.9)$$

This is Lorentz time dilation: the proper time interval τ experienced by a moving person is a factor γ less than the time interval t according to an onlooker.

1.7.2 Fitzgerald-Lorentz contraction

Consider a rocket of proper length l , so that in the rocket's own rest frame (primed) the back and front ends of the rocket move through time t' with coordinates

$$\{t', x'\} = \{t', 0\} \quad \text{and} \quad \{t', l\} . \quad (1.10)$$

From the perspective of an observer who sees the rocket move at velocity v in the x -direction, the worldlines of the back and front ends of the rocket are at

$$\{t, x\} = \{\gamma t', \gamma v t'\} \quad \text{and} \quad \{\gamma t' + \gamma v l, \gamma v t' + \gamma l\} . \quad (1.11)$$

However, the observer measures the length of the rocket simultaneously in their own frame, not the rocket frame. Solving for $\gamma t' = t$ at the back and $\gamma t' + \gamma v l = t$ at the front gives

$$\{t, x\} = \{t, vt\} \quad \text{and} \quad \left\{t, vt + \frac{l}{\gamma}\right\} \quad (1.12)$$

which says that the observer measures the front end of the rocket to be a distance l/γ ahead of the back end. This is Lorentz contraction: an object of proper length l is measured by a moving person to be shorter by a factor γ .

Exercise 1.7 Time dilation. On a spacetime diagram such as that in the left panel of Figure 1.10, show how two observers moving relative to each other can both consider the other's clock to run slow compared to their own.

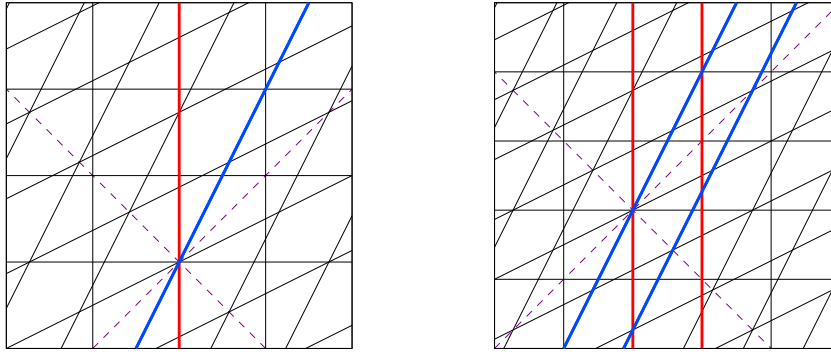


Figure 1.10 (Left) Time dilation, and (right) Lorentz contraction spacetime diagrams.

Exercise 1.8 Lorentz contraction. On a spacetime diagram such as that in the right panel Figure 1.10, show how two observers moving relative to each other can both consider the other to be contracted along the direction of motion.

Concept question 1.9 Is one side of a cube shorter than the other? Figure 1.11 shows a picture of a 3-dimensional cube. Is one edge shorter than the other? Projected on to the page, it appears so, but in reality all the edges have equal length. In what ways is this situation similar or dissimilar to time dilation and Lorentz contraction in 4-dimensional relativity?

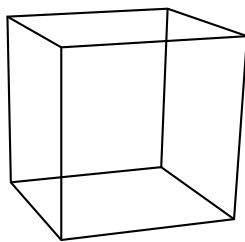


Figure 1.11 A cube. Are the lengths of its sides all equal?

Exercise 1.10 Twin paradox. Your twin leaves you on Earth and travels to the spacestation Alpha, $\ell = 3$ lyr away, at a good fraction of the speed of light, then immediately returns to Earth at the same speed. Figure 1.12 shows on a spacetime diagram the corresponding worldlines of both you and your twin. Aside from part 1 and the first part of 2, you should derive your answers mathematically, using logic and Lorentz transformations. However, the diagram is accurately drawn, and you should be able to check your answers by measuring.

1. On a spacetime diagram such that in Figure 1.12, label the worldlines of you and your twin. Draw the worldline of a light signal which travels from you on Earth, hits Alpha just when your twin arrives, and immediately returns to Earth. Draw the twin's "now" (line of simultaneity) when just arriving at Alpha, and the twin's "now" (line of simultaneity) just departing from Alpha (in the first case the twin is moving toward Alpha, while in the second case the twin is moving back toward Earth).
2. From the diagram, measure the twin's speed v relative to you, in units where the speed of light is unity, $c = 1$. Deduce the Lorentz gamma factor γ , and the redshift factor $1 + z = [(1 + v)/(1 - v)]^{1/2}$, in the cases (i) where the twin is receding, and (ii) where the twin is approaching.
3. Choose the spacetime origin to be the event where the twin leaves Earth. Argue that the position 4-vector of the twin on arrival at Alpha is

$$\{t, x, y, z\} = \{\ell/v, \ell, 0, 0\} . \quad (1.13)$$

Lorentz transform this 4-vector to determine the position 4-vector of the twin on arrival at Alpha, in the twin's frame. Express your answer first in terms of ℓ , v , and γ , and then in (light)years. State in words what this position 4-vector means.

4. How much do you and your twin age respectively during the round trip to Alpha and back? What is the ratio of these ages? Express your answers first in terms of ℓ , v , and γ , and then in years.
5. What is the distance between the Earth and Alpha from the twin's point of view? What is the ratio of this distance to the distance between Earth and Alpha from your point of view? Explain how you arrived at your result. Express your answer first in terms of ℓ , v , and γ , and then in lightyears.
6. You watch your twin through a telescope. How much time do you see (through the telescope) elapse on your twin's wristwatch between launch and arrival on Alpha? How much time passes on your own

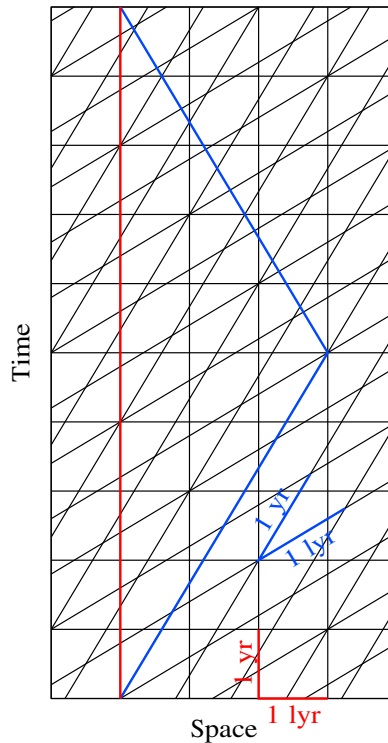


Figure 1.12 Twin paradox spacetime diagram.

- wristwatch during this time? What is the ratio of these two times? Express your answers first in terms of ℓ , v , and γ , and then in years.
7. On arrival at Alpha, your twin looks back through a telescope at your wristwatch. How much time does your twin see (through the telescope) has elapsed since launch on your watch? How much time has elapsed on the twin's own wristwatch during this time? What is the ratio of these two times? Express your answers first in terms of ℓ , v , and γ , and then in years.
 8. You continue to watch your twin through a telescope. How much time elapses on your twin's wristwatch, as seen by you through the telescope, during the twin's journey back from Alpha to Earth? How much time passes on your own watch as you watch (through the telescope) the twin journey back from Alpha to Earth? What is the ratio of these two times? Express your answers first in terms of ℓ , v , and γ , and then in years.
 9. During the journey back from Alpha to Earth, your twin likewise continues to look through a telescope at the time registered on your watch. How much time passes on your wristwatch, as seen by your twin through the telescope, during the journey back? How much time passes on the twin's wristwatch from

the twin's point of view during the journey back? What is the ratio of these two times? Express your answers first in terms of ℓ , v , and γ , and then in years.

Concept question 1.11 What breaks the symmetry between you and your twin? From your point of view, you saw the twin recede from you at velocity v on the outbound journey, then approach you at velocity v on the inbound journey. But the twin saw the essentially same thing: from the twin's point of view, the twin saw you recede at velocity v on the outbound journey, then approach the twin at velocity v on the inbound journey. Isn't the situation symmetrical, so shouldn't you and the twin age identically? What breaks the symmetry, allowing your twin to age less?

1.8 The spacetime wheel

1.8.1 Wheel

Figure 1.13 shows an ordinary 3-dimensional wheel. As the wheel rotates, a point on the wheel describes an invariant circle. The coordinates $\{x, y\}$ of a point on the wheel relative to its centre change, but the distance r between the point and the centre remains constant

$$r^2 = x^2 + y^2 = \text{constant} . \quad (1.14)$$

More generally, the coordinates $\{x, y, z\}$ of the interval between any two points in 3-dimensional space (a vector) change when the coordinate system is rotated in 3 dimensions, but the separation r of the two points remains constant

$$r^2 = x^2 + y^2 + z^2 = \text{constant} . \quad (1.15)$$

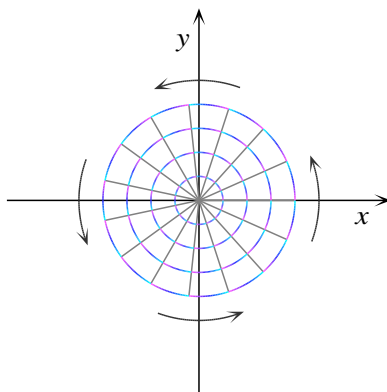


Figure 1.13 A wheel.

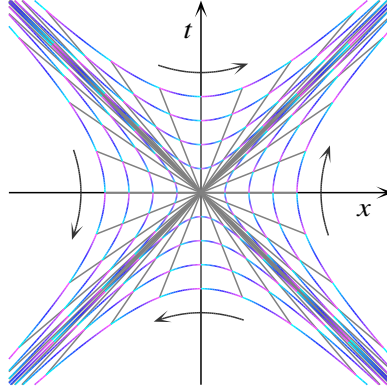


Figure 1.14 A spacetime wheel.

1.8.2 Spacetime wheel

Figure 1.14 shows a spacetime wheel. The diagram here is a spacetime diagram, with time t vertical and space x horizontal. A rotation between time t and space x is a Lorentz boost in the x -direction. As the spacetime wheel boosts, a point on the wheel describes an invariant hyperbola. The spacetime coordinates $\{t, x\}$ of a point on the wheel relative to its centre change, but the spacetime separation s between the point and the centre remains constant

$$s^2 = -t^2 + x^2 = \text{constant} . \quad (1.16)$$

More generally, the coordinates $\{t, x, y, z\}$ of the interval between any two events in 4-dimensional spacetime (a 4-vector) change when the coordinate system is boosted or rotated, but the spacetime separation s of the two events remains constant

$$s^2 = -t^2 + x^2 + y^2 + z^2 = \text{constant} . \quad (1.17)$$

1.8.3 Lorentz boost as a rotation by an imaginary angle

The $-$ sign instead of a $+$ sign in front of the t^2 in the spacetime separation formula (1.17) means that time t can often be treated mathematically as if it were an imaginary spatial dimension. That is, $t = iw$ where $i \equiv \sqrt{-1}$ and w is a “fourth spatial coordinate.”

A Lorentz boost by a velocity v can likewise be treated as a rotation by an imaginary angle. Consider a normal spatial rotation in which a primed frame is rotated in the wx -plane clockwise by an angle a about the origin, relative to the unprimed frame. The relation between the coordinates $\{w', x'\}$ and $\{w, x\}$ of a point in the two frames is

$$\begin{pmatrix} w' \\ x' \end{pmatrix} = \begin{pmatrix} \cos a & -\sin a \\ \sin a & \cos a \end{pmatrix} \begin{pmatrix} w \\ x \end{pmatrix} . \quad (1.18)$$

Now set $t = iw$ and $\alpha = ia$ with t and α both real. In other words, take the spatial coordinate w to be imaginary, and the rotation angle a likewise to be imaginary. Then the rotation formula above becomes

$$\begin{pmatrix} t' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh \alpha & -\sinh \alpha \\ -\sinh \alpha & \cosh \alpha \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} \quad (1.19)$$

This agrees with the usual Lorentz transformation formula (1.5) if the boost velocity v and boost angle α are related by

$$v = \tanh \alpha , \quad (1.20)$$

so that

$$\gamma = \cosh \alpha , \quad \gamma v = \sinh \alpha . \quad (1.21)$$

The boost angle α is commonly called the **rapidity**. This provides a convenient way to add velocities in special relativity: the rapidities simply add (for boosts along the same direction), just as spatial rotation angles add (for rotations about the same axis). Thus a boost by velocity $v_1 = \tanh \alpha_1$ followed by a boost by velocity $v_2 = \tanh \alpha_2$ in the same direction gives a net velocity boost of $v = \tanh \alpha$ where

$$\alpha = \alpha_1 + \alpha_2 . \quad (1.22)$$

The equivalent formula for the velocities themselves is

$$v = \frac{v_1 + v_2}{1 + v_1 v_2} , \quad (1.23)$$

the special relativistic velocity addition formula.

1.8.4 Trip across the Universe at constant acceleration

Suppose that you took a trip across the Universe in a spaceship, accelerating all the time at one Earth gravity g . How far would you travel in how much time?

The spacetime wheel offers a cute way to solve this problem, since the rotating spacetime wheel can be regarded as representing spacetime frames undergoing constant acceleration. Points on the right quadrant of the rotating spacetime wheel, Figure 1.15, represent worldlines of persons who accelerate with constant acceleration in their own frame. The spokes of the spacetime wheel are lines of simultaneity for the accelerating persons.

If the units of space and time are chosen so that the speed of light and the gravitational acceleration are both one, $c = g = 1$, then the proper time experienced by the accelerating person is the rapidity α , and the time and space coordinates of the accelerating person, relative to a person who remains at rest, are those of a point on the spacetime wheel, namely

$$\{t, x\} = \{\sinh \alpha, \cosh \alpha\} . \quad (1.24)$$

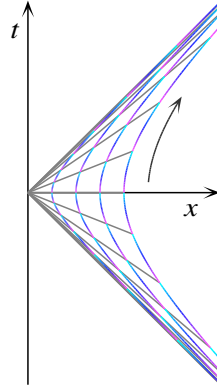


Figure 1.15 The right quadrant of the spacetime wheel represents the worldlines and lines of simultaneity of persons who accelerate in the x direction with uniform acceleration in their own frames.

In the case where the acceleration is one Earth gravity, $g = 9.80665 \text{ m s}^{-2}$, the unit of time is

$$\frac{c}{g} = \frac{299,792,458 \text{ m s}^{-1}}{9.80665 \text{ m s}^{-2}} = 0.97 \text{ yr} , \quad (1.25)$$

Table 1.1: Trip across the Universe.

Time elapsed on spaceship in years	Time elapsed on Earth in years	Distance travelled in lightyears	To
α	$\sinh \alpha$	$\cosh \alpha - 1$	
0	0	0	Earth (starting point)
1	1.175	.5431	
2	3.627	2.762	
2.34	5.12	4.22	Proxima Cen
3.962	26.3	25.3	Vega
6.60	368	367	Pleiades
10.9	2.7×10^4	2.7×10^4	Centre of Milky Way
15.4	2.44×10^6	2.44×10^6	Andromeda galaxy
18.4	4.9×10^7	4.9×10^7	Virgo cluster
19.2	1.1×10^8	1.1×10^8	Coma cluster
25.3	5×10^{10}	5×10^{10}	Edge of observable Universe

just short of one year. For simplicity, Table 1.1, which tabulates some milestones along the way, takes the unit of time to be exactly one year, which would be the case if you were accelerating at $0.97g = 9.5 \text{ m s}^{-2}$.

After a slow start, you cover ground at an ever increasing rate, crossing 50 billion lightyears, the distance to the edge of the currently observable Universe, in just over 25 years of your own time.

Does this mean you go faster than the speed of light? No. From the point of view of a person at rest on Earth, you never go faster than the speed of light. From your own point of view, distances along your direction of motion are Lorentz-contracted, so distances that are vast from Earth's point of view appear much shorter to you. Fast as the Universe rushes by, it never goes faster than the speed of light.

This rosy picture of being able to flit around the Universe has drawbacks. Firstly, it would take a huge amount of energy to keep you accelerating at g . Secondly, you would use up a huge amount of Earth time travelling around at relativistic speeds. If you took a trip to the edge of the Universe, then by the time you got back not only would all your friends and relations be dead, but the Earth would probably be gone, swallowed by the Sun in its red giant phase, the Sun would have exhausted its fuel and shrivelled into a cold white dwarf star, and the Solar System, having orbited the Galaxy a thousand times, would be lost somewhere in its milky ways.

Technical point. The Universe is expanding, so the distance to the edge of the currently observable Universe is increasing. Thus it would actually take longer than indicated in the table to reach the edge of the currently observable Universe. Moreover if the Universe is accelerating, as evidence from the Hubble diagram of Type Ia Supernovae indicates, then you will never be able to reach the edge of the currently observable Universe, however fast you go.

Exercise 1.12 Length of a particle accelerator that reaches the GUT or Planck scale. Consider a linear particle accelerator able to accelerate particles at constant acceleration g in the particles' own frame.

1. How long must the particle accelerator be to reach a Lorentz gamma factor of γ ?
2. Estimate the acceleration g for a contemporary accelerator such as the Large Hadron Collider.
3. Estimate the length of a particle accelerator needed to accelerate a proton, rest mass 1 GeV, to a GUT energy of 10^{16} GeV, or alternatively to a Planck energy of 10^{19} GeV.
4. Show that a GUT density of 1 GUT mass per (GUT length)³ is about 10^{81} times the density of water. Approximately what is the Planck density relative to the density of water?
5. To what Lorentz γ factor would you have to accelerate two rocks so that they achieve a GUT or Planck density when slammed together? How long would the particle accelerator be to achieve this γ factor?

Solution.

1. The rapidity α achieved by a particle that accelerates at constant acceleration g in its own frame for a proper time τ is

$$\alpha = \frac{g\tau}{c} . \quad (1.26)$$

The Lorentz gamma factor γ is related to the rapidity by $\gamma = \cosh \alpha$, equation (1.21). The distance x

the particle moves in the background frame is

$$x = \frac{c^2}{g}(\cosh \alpha - 1) = \frac{c^2}{g}(\gamma - 1) . \quad (1.27)$$

In the highly relativistic regime, $\gamma \gg 1$, the distance travelled is

$$x \approx \frac{c^2 \gamma}{g} . \quad (1.28)$$

The distance x increases linearly with γ .

2. The Large Hadron Collider (LHC) accelerates protons and heavier nuclei to energies of order 1 TeV, whereat a proton has a gamma factor of $\gamma \approx 10^3$. The acceleration occurs over scales of kilometres, or 10^3 m. So the acceleration is about one per metre,

$$\frac{g}{c^2} \approx 1 \text{ m}^{-1} . \quad (1.29)$$

3. A GUT energy of 10^{16} GeV requires a gamma factor of 10^{16} , hence a particle accelerator of length

$$x \approx 10^{16} \text{ m} \approx 1 \text{ lyr} . \quad (1.30)$$

A Planck energy of 10^{19} GeV requires a particle accelerator of length

$$x \approx 10^{19} \text{ m} \approx 1000 \text{ lyr} . \quad (1.31)$$

4. The Planck energy 10^{19} GeV is 10^3 higher than the GUT density 10^{16} GeV. The Planck density is then $(10^3)^4 = 10^{12}$ times higher than the GUT density of $10^{81} \text{ gm cm}^{-3}$. The Planck density is $10^{93} \text{ gm cm}^{-3}$.
 5. When two objects are slammed together at Lorentz factor γ , the energy of each object is enhanced by a factor γ , and the length of each object is contracted along the direction of motion by another factor of γ , so overall the density is increased by a factor of γ^2 . To reach a GUT density of $10^{81} \text{ gm cm}^{-3}$ by slamming together two rocks of initial density say 10 gm cm^{-3} would require a gamma factor of $\sqrt{10^{80}} = 10^{40}$. Which would require a particle accelerator of length 10^{40} m, or 10^{24} lyr, or about 10^{14} times the size of the observable Universe.
-

1.9 Scalar spacetime distance

The fact that Lorentz transformations leave unchanged a certain distance, the spacetime distance, between any two events in spacetime is one the most fundamental features of Lorentz transformations. The scalar spacetime distance Δs between two events separated by $\{\Delta t, \Delta x, \Delta y, \Delta z\}$ is given by

$$\begin{aligned} \Delta s^2 &= -\Delta t^2 + \Delta r^2 \\ &= -\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2 . \end{aligned} \quad (1.32)$$

A quantity such as Δs^2 that remains unchanged under any Lorentz transformation is called a **scalar**. You should check yourself that Δs^2 is unchanged under Lorentz transformations (see Exercise 1.14). Lorentz transformations can be defined as linear spacetime transformations that leave Δs^2 invariant.

The single scalar spacetime squared interval Δs^2 replaces the two scalar quantities

$$\begin{array}{ll} \text{time interval} & \Delta t \\ \text{spatial interval} & \Delta r = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \end{array} \quad (1.33)$$

of classical Galilean spacetime.

1.9.1 Timelike, lightlike, spacelike

The scalar spacetime distance squared Δs^2 , equation (1.32), between two events can be negative, zero, or positive. A spacetime interval $\{\Delta t, \Delta x, \Delta y, \Delta z\} \equiv \{\Delta t, \Delta \mathbf{r}\}$ is called

$$\begin{array}{lll} \text{timelike} & \text{if } \Delta t > \Delta r & \text{or equivalently if } \Delta s^2 < 0, \\ \text{null or lightlike} & \text{if } \Delta t = \Delta r & \text{or equivalently if } \Delta s^2 = 0, \\ \text{spacelike} & \text{if } \Delta t < \Delta r & \text{or equivalently if } \Delta s^2 > 0, \end{array} \quad (1.34)$$

as illustrated in Figure 1.16.

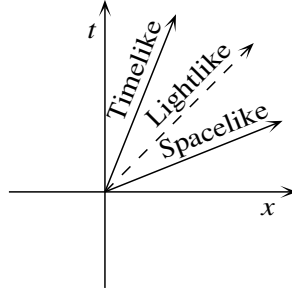


Figure 1.16 Spacetime diagram illustrating timelike, lightlike, and spacelike intervals.

1.9.2 Proper time, proper distance

The scalar spacetime distance squared Δs^2 has a physical meaning.

If an interval $\{\Delta t, \Delta \mathbf{r}\}$ is timelike, $\Delta t > \Delta r$, then the square root of minus the spacetime interval squared is the **proper time** $\Delta \tau$ along it

$$\Delta \tau = \sqrt{-\Delta s^2} = \sqrt{\Delta t^2 - \Delta r^2}. \quad (1.35)$$

This is the time experienced by an observer moving along that interval.

If an interval $\{\Delta t, \Delta \mathbf{r}\}$ is spacelike, $\Delta t < \Delta r$, then the spacetime interval equals the **proper distance** Δl along it

$$\Delta l = \sqrt{\Delta s^2} = \sqrt{\Delta r^2 - \Delta t^2} . \quad (1.36)$$

This is the distance between two events measured by an observer for whom those events are simultaneous.

Concept question 1.13 Proper time, proper distance. Justify the assertions (1.35) and (1.36).

1.9.3 Minkowski metric

It is convenient to denote an interval using an index notation,

$$\Delta x^m \equiv \{\Delta t, \Delta \mathbf{r}\} \equiv \{\Delta t, \Delta x, \Delta y, \Delta z\} . \quad (1.37)$$

The indices run over $m = t, x, y, z$, or sometimes $m = 0, 1, 2, 3$. The scalar spacetime length squared Δs^2 of an interval Δx^m can be abbreviated

$$\Delta s^2 = \eta_{mn} \Delta x^m \Delta x^n , \quad (1.38)$$

where η_{mn} is the **Minkowski metric**

$$\eta_{mn} \equiv \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad (1.39)$$

Equation (1.38) uses the **implicit summation convention**, according to which paired indices, one lowered and one raised, are implicitly summed over.

1.10 4-vectors

1.10.1 Contravariant 4-vector

Under a Lorentz transformation, a coordinate interval Δx^m transforms as

$$\Delta x^m \rightarrow \Delta x'^m = L^m_n \Delta x^n , \quad (1.40)$$

where L^m_n denotes a Lorentz transformation. The paired indices n on the right hand side of equation (1.40), one lowered and one raised, are implicitly summed over. In matrix notation, L^m_n is a 4×4 matrix. For example, for a Lorentz boost by velocity v along the x -axis, L^m_n is the matrix on the right hand side of equation (1.5).

In special relativity a **contravariant 4-vector** is defined to be a quantity

$$a^m \equiv \{a^t, a^x, a^y, a^z\} , \quad (1.41)$$

that transforms under Lorentz transformations like an interval Δx^m of spacetime,

$$a^m \rightarrow a'^m = L^m_n a^n . \quad (1.42)$$

The indices run over $m = t, x, y, z$, or sometimes $m = 0, 1, 2, 3$.

1.10.2 Covariant 4-vector

In special and general relativity, besides the contravariant 4-vector a^m , with raised indices, it is convenient to introduce a **covariant 4-vector** a_m , with lowered indices, obtained by multiplying the contravariant 4-vector by the metric,

$$a_m \equiv \eta_{mn} a^n . \quad (1.43)$$

With the Minkowski metric (1.39), the covariant components a_m are

$$a_m = \{-a^t, a^x, a^y, a^z\} , \quad (1.44)$$

which differ from the contravariant components a^m only in the sign of the time component.

The reason for introducing the two species of vector is that their implicitly summed product

$$\begin{aligned} a^m a_m &\equiv \eta_{mn} a^m a^n \\ &= a_t a^t + a_x a^x + a_y a^y + a_z a^z \\ &= -(a^t)^2 + (a^x)^2 + (a^y)^2 + (a^z)^2 \end{aligned} \quad (1.45)$$

is a Lorentz scalar, a fact you will prove in Exercise 1.14.

The notation may seem overly elaborate, but it proves extremely useful in general relativity, where the metric is more complicated than Minkowski. Further discussion of the formalism of 4-vectors is deferred to Chapter 2.

Exercise 1.14 Scalar product. Suppose that a^m and b^m are two 4-vectors. Show that $a_m b^m$ is a scalar, that is, it is unchanged by any Lorentz transformation. [Hint: For the Minkowski metric of special relativity, $a_m b^m = -a^t b^t + a^x b^x + a^y b^y + a^z b^z$. Show that $a'_m b'^m = a_m b^m$. You may assume without proof the familiar result that the 3D scalar product $\mathbf{a} \cdot \mathbf{b} = a^x b^x + a^y b^y + a^z b^z$ of two 3-vectors is unchanged by any spatial rotation, so it suffices to consider a Lorentz boost, say in the x direction.]

Exercise 1.15 The principle of longest proper time. Consider a person whose worldline goes from spacetime event P_0 to spacetime event P_1 at velocity v_1 relative to some inertial frame, and then from P_1 to spacetime event P_2 at velocity v_2 , as illustrated in Figure 1.17. Assume for simplicity that the velocities are both in the (positive or negative) x -direction. Show that the proper time along a straight line from P_0 to P_2 is always greater than or equal to the sum of the proper times along the two straight lines from P_0 to P_1 followed by P_1 to P_2 . Hence conclude that the longest proper time between two events is a straight line. What does this imply about the twin paradox? [Hint: It is simplest to use rapidities α rather than

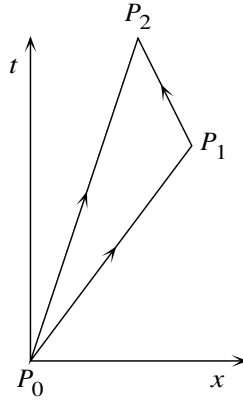


Figure 1.17 The longest proper time between P_0 and P_2 is a straight line.

velocities. Let the segment from P_0 to P_1 be $\{t_1, x_1\} = \tau_1\{\cosh \alpha_1, \sinh \alpha_1\}$, and the segment from P_1 to P_2 be $\{t_2, x_2\} = \tau_2\{\cosh \alpha_2, \sinh \alpha_2\}$. The segment from P_0 to P_2 is the sum of these, $\{t, x\} = \{t_1 + t_2, x_1 + x_2\}$. Show that

$$\tau^2 - (\tau_1 + \tau_2)^2 = 4\tau_1\tau_2 \sinh^2\left(\frac{\alpha_2 - \alpha_1}{2}\right), \quad (1.46)$$

which is a minimum for $\alpha_2 = \alpha_1$.]

1.11 Energy-momentum 4-vector

The foremost example of a 4-vector other than the interval Δx^m is the energy-momentum 4-vector.

One of the great insights of modern physics is that conservation laws are associated with symmetries. The Principle of Special Relativity asserts that the laws of physics should take the same form at any point. There is no preferred origin in spacetime in special relativity. In special relativity, spacetime has translation symmetry with respect to both time and space. Associated with those symmetries are laws of conservation of energy and momentum:

Symmetry	Conservation law
Time translation	Energy
Space translation	Momentum

Since one-dimensional time and three-dimensional space are united in special relativity, this suggests that the single component of energy and the three components of momentum should be combined into a 4-vector:

$$\left. \begin{array}{l} \text{energy} = \text{time component} \\ \text{momentum} = \text{space component} \end{array} \right\} \text{ of a 4-vector.} \quad (1.47)$$

The Principle of Special Relativity requires that the equation of energy-momentum conservation

$$\frac{\text{energy}}{\text{momentum}} = \text{constant} \quad (1.48)$$

should take the same form in any inertial frame. The equation should be **Lorentz covariant**, that is, the equation should transform like a Lorentz 4-vector.

1.11.1 Construction of the energy-momentum 4-vector

The energy-momentum 4-vector of a particle of mass m at position $\{t, \mathbf{r}\}$ moving at velocity $\mathbf{v} = d\mathbf{r}/dt$ can be derived by requiring

1. that is a 4-vector, and
2. that it goes over to the Newtonian limit as $v \rightarrow 0$.

In the Newtonian limit, the 3-momentum $\mathbf{p}$ equals mass m times velocity $\mathbf{v}$,

$$\mathbf{p} = m\mathbf{v} = m \frac{d\mathbf{r}}{dt} . \quad (1.49)$$

To obtain a 4-vector, two things must be done to the Newtonian momentum:

1. replace $\mathbf{r}$ by a 4-vector $x^n = \{t, \mathbf{r}\}$, and
2. replace dt by a scalar; the only available scalar measure of time is the proper time interval $d\tau$ along the worldline of the particle.

The result is the energy-momentum 4-vector p^n :

$$\begin{aligned} p^n &= m \frac{dx^n}{d\tau} \\ &= m \left\{ \frac{dt}{d\tau}, \frac{d\mathbf{r}}{d\tau} \right\} \\ &= m \{ \gamma, \gamma \mathbf{v} \} . \end{aligned} \quad (1.50)$$

The components of the energy-momentum 4-vector are the special relativistic versions of energy E and momentum $\mathbf{p}$,

$$p^n = \{E, \mathbf{p}\} = \{m\gamma, m\gamma \mathbf{v}\} . \quad (1.51)$$

1.11.2 Special relativistic energy

From equation (1.51), the special relativistic energy E is the product of the rest mass and the Lorentz γ -factor,

$$E = m\gamma \quad (\text{units } c = 1) , \quad (1.52)$$

or, restoring standard units,

$$E = mc^2 \gamma . \quad (1.53)$$

For small velocities v , the Taylor expansion of the Lorentz factor γ is

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = 1 + \frac{1}{2} \frac{v^2}{c^2} + \dots \quad (1.54)$$

Thus for small velocities, the special relativistic energy E Taylor expands as

$$\begin{aligned} E &= mc^2 \left(1 + \frac{1}{2} \frac{v^2}{c^2} + \dots \right) \\ &= mc^2 + \frac{1}{2} mv^2 + \dots \end{aligned} \quad (1.55)$$

The first term, mc^2 , is the rest-mass energy. The second term, $\frac{1}{2}mv^2$, is the nonrelativistic kinetic energy. Higher-order terms give relativistic corrections to the kinetic energy.

Einstein did not discard the constant term, but rather interpreted it seriously as indicating that mass contains energy, the rest-mass energy

$$E = mc^2, \quad (1.56)$$

perhaps the most famous equation in all of physics.

1.11.3 Rest mass is a scalar

The scalar quantity constructed from the energy-momentum 4-vector $p^n = \{E, \mathbf{p}\}$ is

$$\begin{aligned} p_n p^n &= -E^2 + p^2 \\ &= -m^2(\gamma^2 - \gamma^2 v^2) \\ &= -m^2, \end{aligned} \quad (1.57)$$

minus the square of the rest mass. The minus sign is associated with the choice $-+++$ of metric signature in this book.

Elementary texts sometimes state that special relativity implies that the mass of a particle increases as its velocity increases, but this is a confusing way of thinking. Mass is rest mass m , a scalar, not to be confused with energy. That being said, Einstein's famous equation (1.56) does suggest that rest mass is a form of energy, and indeed that proves to be the case. Rest mass is routinely converted into energy in chemical or nuclear reactions that liberate heat.

1.12 Photon energy-momentum

The energy-momentum 4-vectors of photons are of special interest because when you move through a scene at near the speed of light, the scene appears distorted by the Lorentz transformation of the photon 4-vectors that you see.

A photon has zero rest mass

$$m = 0 . \quad (1.58)$$

Its scalar energy-momentum squared is thus zero,

$$p_n p^n = -E^2 + p^2 = -m^2 = 0 . \quad (1.59)$$

Consequently the 3-momentum of a photon equals its energy (in units $c = 1$),

$$p \equiv |\mathbf{p}| = E . \quad (1.60)$$

The energy-momentum 4-vector of a photon therefore takes the form

$$\begin{aligned} p^n &= \{E, \mathbf{p}\} \\ &= E\{1, \mathbf{n}\} \\ &= h\nu\{1, \mathbf{n}\} \end{aligned} \quad (1.61)$$

where ν is the photon frequency. The photon velocity is $\mathbf{n}$, a unit vector. The photon speed is one, the speed of light.

1.12.1 Lorentz transformation of the photon energy-momentum 4-vector

The energy-momentum 4-vector p^m of a photon follows the usual rules for 4-vectors under Lorentz transformations. In the case that the emitter (primed frame) is moving at velocity v along the x -axis relative to the observer (unprimed frame), the transformation is

$$\begin{pmatrix} p'^t \\ p'^x \\ p'^y \\ p'^z \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p^t \\ p^x \\ p^y \\ p^z \end{pmatrix} = \begin{pmatrix} \gamma(p^t - vp^x) \\ \gamma(p^x - vp^t) \\ p^y \\ p^z \end{pmatrix} . \quad (1.62)$$

Equivalently

$$h\nu' \begin{pmatrix} 1 \\ n'^x \\ n'^y \\ n'^z \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} h\nu \begin{pmatrix} 1 \\ n^x \\ n^y \\ n^z \end{pmatrix} = h\nu \begin{pmatrix} \gamma(1 - n^x v) \\ \gamma(n^x - v) \\ n^y \\ n^z \end{pmatrix} . \quad (1.63)$$

These mathematical relations imply the rules of 4-dimensional perspective, §1.13.2.

1.12.2 Redshift

The wavelength λ of a photon is related to its frequency ν by

$$\lambda = c/\nu . \quad (1.64)$$

Astronomers define the **redshift** z of a photon by the shift of the observed wavelength λ_{obs} compared to its emitted wavelength λ_{em} ,

$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} . \quad (1.65)$$

In relativity, it is often more convenient to use the **redshift factor** $1 + z$,

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} . \quad (1.66)$$

Sometimes it is useful to use a **blueshift factor** which is just the reciprocal of the redshift factor,

$$\frac{1}{1 + z} \equiv \frac{\lambda_{\text{em}}}{\lambda_{\text{obs}}} = \frac{\nu_{\text{obs}}}{\nu_{\text{em}}} . \quad (1.67)$$

1.12.3 Special relativistic Doppler shift

If the emitter frame (primed) is moving with velocity v in the x -direction relative to the observer frame (unprimed) then the emitted and observed frequencies are related by, equation (1.63),

$$h\nu_{\text{em}} = h\nu_{\text{obs}}\gamma(1 - n^x v) . \quad (1.68)$$

The redshift factor is therefore

$$\begin{aligned} 1 + z &= \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} \\ &= \gamma(1 - n^x v) \\ &= \gamma(1 - \mathbf{n} \cdot \mathbf{v}) . \end{aligned} \quad (1.69)$$

Equation (1.69) is the general formula for the special relativistic Doppler shift. In special cases,

$$1 + z = \begin{cases} \sqrt{\frac{1-v}{1+v}} & \text{velocity directly towards observer } (\mathbf{v} \text{ aligned with } \mathbf{n}) , \\ \gamma & \text{velocity in the transverse direction } (\mathbf{v} \cdot \mathbf{n} = 0) , \\ \sqrt{\frac{1+v}{1-v}} & \text{velocity directly away from observer } (\mathbf{v} \text{ anti-aligned with } \mathbf{n}) . \end{cases} \quad (1.70)$$

1.13 What things look like at relativistic speeds

1.13.1 Light travel time effects

When you move through a scene at near the speed of light, the scene appears distorted not only by time dilation and Lorentz contraction, but also by differences in the light travel time from different parts of the scene. The effect of differential light travel times is comparable to the effects of time dilation and Lorentz contraction, and cannot be ignored.

An excellent way to see the importance of light travel time is to work through the twin paradox, Exercise 1.10. Nature provides a striking example of the importance of light travel time in the form of superluminal (faster-than-light) jets in galaxies, the subject of Exercise 1.16.

Exercise 1.16 Superluminal jets.

Radio observations of galaxies show in many cases twin jets emerging from the nucleus of the galaxy. The jets are typically narrow and long, often penetrating beyond the optical extent of the galaxy. The jets are frequently one-sided, and in some cases that are favourable to observation the jets are found to be superluminal. A celebrated example is the giant elliptical galaxy M87 at the centre of the Local Supercluster, whose jet is observed over a broad range of wavelengths, including optical wavelengths. Hubble Space Telescope observations, Figure 1.18, show blobs in the M87 jet moving across the sky at approximately $6c$.

1. Draw a spacetime diagram of the situation, in Earth's frame of reference. Assume that the velocity of the galaxy M87 relative to Earth is negligible. Let the x -axis be the direction to M87. Choose the y -axis so the jet lies in the xy -plane. Let the jet be moving at velocity v at angle θ away from the direction towards us on Earth, so that its spatial velocity relative to Earth is $\mathbf{v} \equiv \{v_x, v_y\} = \{-v \cos \theta, v \sin \theta\}$.

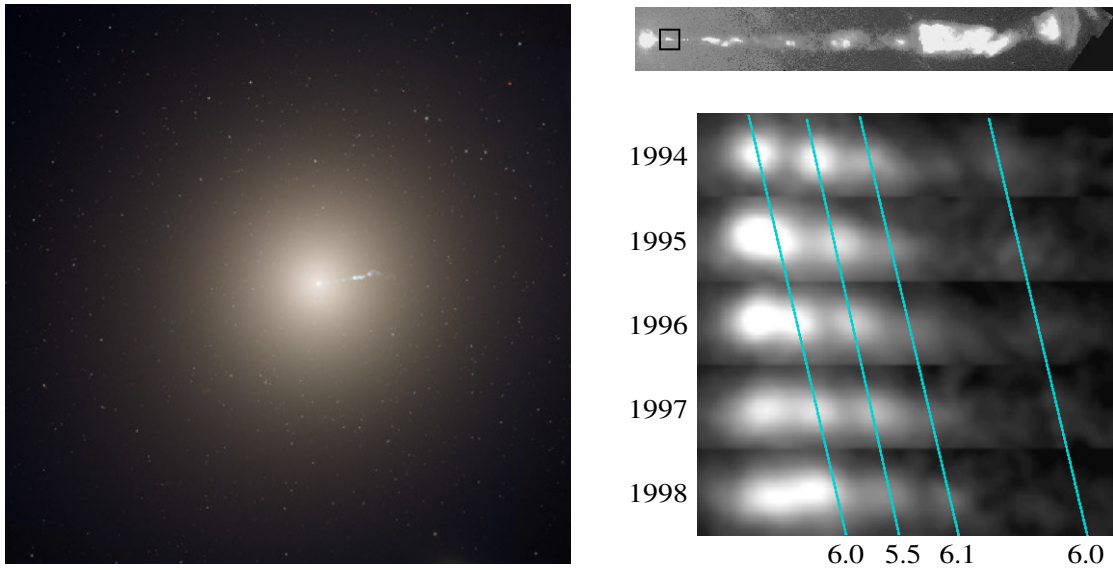


Figure 1.18 The left panel shows an image of the galaxy M87 taken with the Advanced Camera for Surveys on the Hubble Space Telescope. A jet, bluish compared to the starry background of the galaxy, emerges from the galaxy's central nucleus. Radio observations, not shown here, reveal that there is a second jet in the opposite direction. Credit: STScI/AURA. The right panel is a sequence of Hubble images showing blobs in the jet moving superluminally, at approximately $6c$ (Biretta, Sparks, and Macchetto, 1999; Meyer et al., 2013). The slanting lines track the moving features, with speeds given in units of c . The upper strip shows where in the jet the blobs were located. Credit: John Biretta, STScI.

2. In Earth coordinates $\{t, x, y\}$, the jet moves in time t a distance $\mathbf{l} = \{l_x, l_y\} = \mathbf{v}t$. Argue that during an Earth time t , the jet has moved a distance l_x nearer to the Earth (the distances l_x and l_y are both tiny compared to the distance to M87), so the apparent time as seen through a telescope is not t , but rather t diminished by the light travel time l_x (units $c = 1$). Hence conclude that the apparent transverse velocity on the sky is

$$v_{\text{app}} = \frac{v \sin \theta}{1 - v \cos \theta} . \quad (1.71)$$

3. Sketch the apparent velocity v_{app} as a function of θ for some given velocity v . In terms of v and the Lorentz factor γ , what are the values of θ and of v_{app} at the point where v_{app} reaches its maximum? What can you conclude about the jet in M87?
4. What is the expected redshift $1 + z$, or equivalently blueshift $1/(1 + z)$, of the jet as a function of v and θ ? By expressing v in terms of v_{app} and θ using equation (1.71), show that the blueshift factor is

$$\frac{1}{1 + z} = \sqrt{1 + 2v_{\text{app}} \cot \theta - v_{\text{app}}^2} . \quad (1.72)$$

[Hint: Remember to use the correct redshift formula, equation (1.69).]

5. In terms of v_{app} , at what value of θ is the blueshift (i) infinite, or (ii) zero? What are these angles in the case of M87? If the redshift of the jet were measurable, could you deduce the velocity v and opening angle θ ? Unfortunately the redshift of a superluminal jet is not usually observable, because the emission is a continuum of synchrotron emission over a broad range of wavelengths, with no sharp atomic or ionic lines to provide a redshift.
6. Why is the opposing jet not visible?

1.13.2 The rules of 4-dimensional perspective

The distortion of a scene when you move through it at near the speed of light can be calculated most directly from the Lorentz transformation of the energy-momentum 4-vectors of the photons that you see. The result is what I call the “Rules of 4-dimensional perspective.”

Figure 1.19 illustrates the rules of 4-dimensional perspective, also called “special relativistic beaming,” which describe how a scene appears when you move through it at near light speed.

On the left, you are at rest relative to the scene. Imagine painting the scene on a celestial sphere around you. The arrows represent the directions of light rays (photons) from the scene on the celestial sphere to you at the center.

On the right, you are moving to the right through the scene, at 0.8 times the speed of light. The celestial sphere is stretched along the direction of your motion by the Lorentz gamma-factor $\gamma = 1/\sqrt{1 - 0.8^2} = 5/3$ into a celestial ellipsoid. You, the observer, are not at the centre of the ellipsoid, but rather at one of its foci (the left one, if you are moving to the right). The focus of the celestial ellipsoid, where you the observer are, is displaced from centre by $\gamma v = 4/3$. The scene appears relativistically aberrated, which is to say concentrated ahead of you, and expanded behind you.

The lengths of the arrows are proportional to the energies, or frequencies, of the photons that you see.

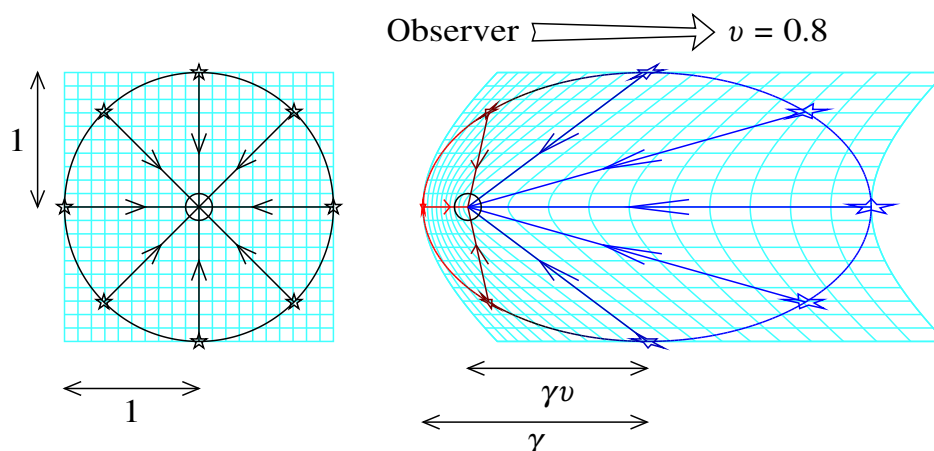


Figure 1.19 The rules of 4-dimensional perspective. In special relativity, the scene seen by an observer moving through the scene (right) is relativistically beamed compared to the scene seen by an observer at rest relative to the scene (left). On the left, the observer at the center of the circle is at rest relative to the surrounding scene. On the right, the observer is moving to the right through the same scene at $v = 0.8$ times the speed of light. The arrowed lines represent energy-momenta of photons. The length of an arrowed line is proportional to the perceived energy of the photon. The scene ahead of the moving observer appears concentrated, blueshifted, and farther away, while the scene behind appears expanded, redshifted, and closer.

When you are moving through the scene at near light speed, the arrows ahead of you, in your direction of motion, are longer than at rest, so you see the photons blue-shifted, increased in energy, increased in frequency. Conversely, the arrows behind you are shorter than at rest, so you see the photons red-shifted, decreased in energy, decreased in frequency. Since photons are good clocks, the change in photon frequency also tells you how fast or slow clocks attached to the scene appear to you to run.

This table summarizes the four effects of relativistic beaming on the appearance of a scene ahead of you and behind you as you move through it at near the speed of light:

Effect	Ahead	Behind
Aberration	Concentrated	Expanded
Colour	Blueshifted	Redshifted
Brightness	Brighter	Dimmer
Time	Speeded up	Slowed down

Mathematical details of the rules of 4-dimensional perspective are explored in the next several Exercises.

Exercise 1.17 The rules of 4-dimensional perspective.

1. In terms of the photon energy-momentum 4-vector p^k in an unprimed frame, what is the photon energy momentum 4-vector p'^k in a primed frame of reference moving at speed v in the x direction relative to the unprimed frame? Argue that the photon 4-vectors in the unprimed and primed frames are related

geometrically by the “celestial ellipsoid” transformation illustrated in Figure 1.19. Bear in mind that the photon vector is pointed *towards* the observer.

2. **Aberration.** The photon 4-vector seen by an observer is the null vector $p^k = E(1, -\mathbf{n})$, where E is the photon energy, and $\mathbf{n}$ is a unit 3-vector in the direction away from the observer, the minus sign taking into account the fact that the photon vector is pointed towards the observer. An object appears in the unprimed frame at angle θ to the x -direction and in the primed frame at angle θ' to the x -direction. Show that $\mu' \equiv \cos \theta'$ and $\mu \equiv \cos \theta$ are related by

$$\mu' = \frac{\mu + v}{1 + v\mu} . \quad (1.73)$$

3. **Redshift.** By what factor $a = E'/E$ is the observed photon frequency from the object changed? Express your answer as a function of γ , v , and μ .
4. **Brightness.** Photons at frequency E in the unprimed frame appear at frequency E' in the primed frame. Argue that the brightness $F(E)$, the number of photons per unit time per unit solid angle per log interval of frequency (about E in the unprimed frame, and E' in the primed frame),

$$F(E) \equiv \frac{dN(E)}{dt \, d\Omega \, d \ln E} , \quad (1.74)$$

goes as

$$\frac{F'(E')}{F(E)} = \frac{E'}{E} \frac{d\mu}{d\mu'} = a^3 . \quad (1.75)$$

[Hint: Photons number conservation implies that $dN'(E') = dN(E)$.]

5. **Time.** By what factor does the rate at which a clock ticks appear to change?

Exercise 1.18 Circles on the sky. Show that a circle on the sky Lorentz transforms to a circle on the sky. Let the primed frame be moving at velocity v in the x -direction, let θ be the angle between the x -direction and the direction $\mathbf{m}$ to the center of the circle, and let α be the angle between the circle axis $\mathbf{m}$ and the photon direction $\mathbf{n}$. Show that the angle θ' in the primed frame is given by

$$\tan \theta' = \frac{\sin \theta}{\gamma(v \cos \alpha + \cos \theta)} , \quad (1.76)$$

and that the angular radius α' in the primed frame is given by

$$\tan \alpha' = \frac{\sin \alpha}{\gamma(\cos \alpha + v \cos \theta)} . \quad (1.77)$$

[This result was first obtained by Penrose (1959) and Terrell (1959), prior to which it had been widely thought that circles would appear Lorentz-contracted and therefore squashed. The following simple proof was told to me by Engelbert Schucking (NYU). The set of null 4-vectors $p^k = E\{1, -\mathbf{n}\}$ on the circle satisfies the Lorentz-invariant equation $x_k p^k = 0$, where $x^k = |x|\{-\cos \alpha, \mathbf{m}\}$ is a spacelike 4-vector whose spatial components $|x|\mathbf{m}$ point to the center of the circle. Note that $|x|$ is a magnitude of a 3-vector, not a Lorentz-invariant scalar.]

Exercise 1.19 Lorentz transformation preserves angles on the sky. From equation (1.73), show that the angular metric $do^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$ on the sky Lorentz transforms as

$$do'^2 = \frac{1 - v^2}{(1 + v \cos \theta)^2} do^2 . \quad (1.78)$$

This kind of transformation, which multiplies the metric by an overall factor, called a conformal factor, is called a **conformal transformation**. The conformal transformation (1.78) of the angular metric shrinks and expands patches on the sky while preserving their shapes, that is, while preserving angles between lines.

Exercise 1.20 The aberration of starlight. The aberration of starlight was discovered by James Bradley (1728) through precision measurements of the position of γ Draconis observed from London with a specially commissioned “zenith sector.” Stellar aberration results from the annual motion of the Earth about the Sun. Calculate the size of the effect, in arcseconds. Are special relativistic effects important? How does the observational signature of stellar aberration differ from that of stellar parallax?

Concept question 1.21 Apparent (affine) distance. The rules of 4-dimensional perspective illustrated in Figure 1.19 suggest that when you move through a scene at near lightspeed, the scene ahead looks farther away (and not Lorentz-contracted at all). Is the scene really farther away, or is it just an illusion? **Answer.** What is reality? In a deep sense, reality is what can be observed (by something, not necessarily a person). So yes, the scene ahead really is farther away. Let the observer take a tape measure that is at rest relative to the observer, and lay it out to the emitter. The laying has to be done in advance, because the emitter is moving. Observers who move at different velocities lay out tapes that move at different velocities. The observer moving faster toward the emitter indeed sees the emitter farther away, according to their tape measure. The distance measured in this fashion is called the **affine distance**, §2.18, a measure of distance along the past lightcone of the observer.

1.14 Occupation number, phase-space volume, intensity, and flux

Exercise 1.17 asked you to discover how the appearance of an emitter changes when the observer boosts into a different frame. The change (1.75) in brightness can be derived at a more fundamental level from the concepts of **occupation number** and **phase-space volume**.

The intensity of light can be described by the number dN of photons in a 3-volume element d^3r of space (as measured by an observer in their own rest frame) with momenta in a 3-volume element d^3p of momentum (again as measured by an observer). The 6-dimensional product $d^3r d^3p$ of spatial and momentum 3-volumes, called the phase-space volume, is Lorentz-invariant, unchanged by a boost or rotation of the observer’s frame (see §10.26.1 for a proof). Indeed, as shown in §4.22, the phase-space volume element $d^3r d^3p$ is invariant under a wide range of transformations (called canonical transformations, §4.17).

In quantum mechanics, the phase volume divided by $(2\pi\hbar)^3$ (which is the same as h^3 ; but in quantum mechanics $\hbar$ is a more natural unit; for example, angular momentum is quantized in units of $\hbar$, and spin in units of $\frac{1}{2}\hbar$) counts the number of free states of particles, here photons. Particles typically have spin, and an

associated discrete number of distinct spin states. Photons have spin 1, and two spin states. The occupation number $f(t, \mathbf{r}, \mathbf{p})$ is defined to be the number of photons per state at time t and spatial position $\mathbf{r}$ with momenta $\mathbf{p}$. The number dN of photons is the product of the occupation number f , the number g of spin states, and the number $d^3r d^3p/(2\pi\hbar)^3$ of free quantum states,

$$dN(t, \mathbf{r}, \mathbf{p}) = f(t, \mathbf{r}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3} . \quad (1.79)$$

The number dN of photons, the occupation number f , the number g of spin states, and the phase volume $d^3r d^3p/(2\pi\hbar)^3$ are all Lorentz invariant.

Astronomers conventionally define the **intensity** I_ν of light observed from an object to be the energy received per unit time t per unit area A (of the telescope mirror or lens) per unit solid angle o per unit frequency ν . Often intensity is quoted per unit wavelength λ or per unit energy E instead of per unit frequency ν , and the intensity is subscripted accordingly, I_λ or I_E . The intensity measures are related by $I_\nu d\nu = I_\lambda d\lambda = I_E dE$ with $\lambda = c/\nu$ and $E = 2\pi\hbar\nu$. The intensity I_E per unit energy is related to the occupation number f by

$$I_E \equiv \frac{E dN}{dt dA do dE} = cf \frac{g p^3}{(2\pi\hbar)^3} , \quad (1.80)$$

the spatial and momentum 3-volumes being $d^3r = c dt dA$ and $d^3p = p^2 dp do$. The p^3 factor in equation (1.80) reproduces the brightness factor $a^3 \equiv (E'/E)^3$ in equation (1.75).

Stars typically appear to astronomers as point sources. Astronomers define the **flux** F_ν from a source to be the intensity I_ν integrated over the solid angle of the source. Again, flux is often quoted per unit wavelength λ or per unit energy E , and subscripted accordingly, F_λ or F_E .

Concept question 1.22 Brightness of a star. How does the flux from a star change when an observer boosts into another frame? The flux that an observer, or a telescope, actually sees depends on the spectrum of the light incident on the observer (the flux as a function of photon energy) and on the sensitivity of the detector as a function of photon energy. But imagine a perfect detector that sees all photons incident on it, of any photon energy.

Solution. The flux F_E in an interval dE of energy is

$$F_E \equiv \frac{E dN}{dt dA dE} = c \frac{g p^3}{(2\pi\hbar)^3} \int f do . \quad (1.81)$$

Since the solid angle varies as $do \propto p^{-2}$, while the occupation number f is Lorentz invariant, and the photon energy and momentum are related by $E = pc$, the flux F_E varies as

$$F_E \propto E , \quad (1.82)$$

that is, the flux is proportional to the blueshift factor. Physically, the observed number of photons per unit time increases in proportion to the photon frequency. The flux integrated over $d \ln E$ counts the total number

of photons observed per unit time, which again increases in proportion to the blueshift factor,

$$\int F_E d \ln E \propto E . \quad (1.83)$$

The flux integrated over dE counts the total energy observed per unit time, which increases as the square of the blueshift factor,

$$\int F_E dE \propto E^2 . \quad (1.84)$$

1.15 How to program Lorentz transformations on a computer

3D gaming programmers are familiar with the fact that the best way to program spatial rotations on a computer is with quaternions. Compared to standard rotation matrices, quaternions offer increased speed and require less storage, and their algebraic properties simplify interpolation and splining.

Section 1.8 showed that a Lorentz boost is mathematically equivalent to a rotation by an imaginary angle. This suggests that Lorentz transformations might be treated as complexified spatial rotations, which proves to be true. Indeed, the best way to program Lorentz transformation on a computer is with complex quaternions, §14.5.

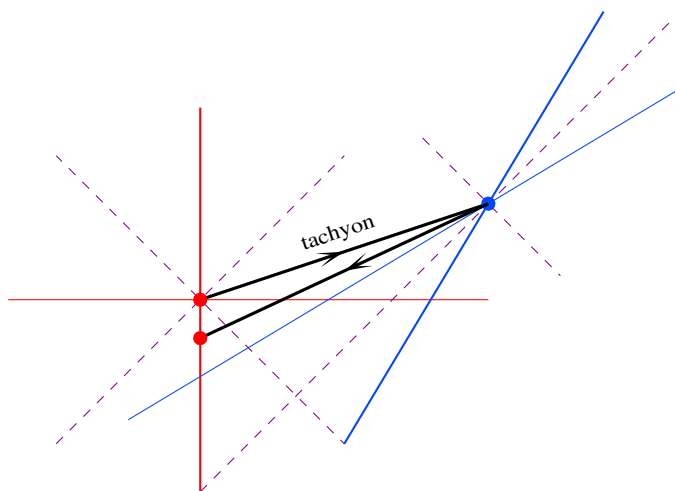


Figure 1.20 Tachyon spacetime diagram.

Exercise 1.23 Tachyons. A tachyon is a hypothetical particle that moves faster than the speed of light.

The purpose of this problem is to discover that the existence of tachyons would imply a violation of causality.

1. On a spacetime diagram such as that in Figure 1.20, show how a tachyon emitted by Vermilion at speed $v > 1$ can appear to go backwards in time, with $v < -1$, in another frame, that of Cerulean.
 2. What is the smallest velocity that Cerulean must be moving relative to Vermilion in order that the tachyon appears to go backwards in Cerulean's time?
 3. Suppose that Cerulean returns the tachyonic signal at the same speed $v > 1$ relative to his own frame. Show on the spacetime diagram how Cerulean's tachyonic signal can reach Vermilion before she sent out the original tachyon.
 4. What is the smallest velocity that Cerulean must be moving relative to Vermilion in order that his tachyon reach Vermilion before she sent out her tachyon?
 5. Why is the situation problematic?
 6. If it is possible for Vermilion to send out a particle with $v > 1$, do you think it should also be possible for her to send out a particle backward in time, with $v < -1$, from her point of view? Explain how she might do this, or not, as the case may be.
-

Concept Questions

1. What assumption of general relativity makes it possible to introduce a coordinate system?
2. Is the speed of light a universal constant in general relativity? If so, in what sense?
3. What does “locally inertial” mean? How local is local?
4. Why is spacetime locally inertial?
5. What assumption of general relativity makes it possible to introduce clocks and rulers?
6. Consider two observers at the same point and with the same instantaneous velocity, but one is accelerating and the other is in free-fall. What is the relation between the proper time or proper distance along an infinitesimal interval measured by the two observers? What assumption of general relativity implies this?
7. Does Einstein’s principle of equivalence imply that two unequal masses will fall at the same rate in a gravitational field? Explain.
8. In what respects is Einstein’s principle of equivalence (gravity is equivalent to acceleration) stronger than the weak principle of equivalence (gravitating mass equals inertial mass)?
9. Standing on the surface of the Earth, you hold an object of negative mass in your hand, and drop it. According to the principle of equivalence, does the negative mass fall up or down?
10. Same as the previous question, but what does Newtonian gravity predict?
11. You have a box of negative mass particles, and you remove energy from it. Do the particles move faster or slower? Does the entropy of the box increase or decrease? Does the pressure exerted by the particles on the walls of the box increase or decrease?
12. You shine two light beams along identical directions in a gravitational field. The two light beams are identical in every way except that they have two different frequencies. Does the equivalence principle imply that the interference pattern produced by each of the beams individually is the same?
13. What is a “straight line,” according to the principle of equivalence?
14. If all objects move on straight lines, how is it that when, standing on the surface of the Earth, you throw two objects in the same direction but with different velocities, they follow two different trajectories?
15. In relativity, what is the generalization of the “shortest distance between two points”?
16. What kinds of general coordinate transformations are allowed in general relativity?

17. In general relativity, what is a scalar? A 4-vector? A tensor? Which of the following is a scalar/vector/tensor/none-of-the-above? (a) a set of coordinates x^μ ; (b) a coordinate interval dx^μ ; (c) proper time τ ?
18. What does general covariance mean?
19. What does parallel transport mean?
20. Why is it important to define covariant derivatives that behave like tensors?
21. Is covariant differentiation a derivation? That is, is covariant differentiation a linear operation, and does it obey the Leibniz rule for the derivative of a product?
22. What is the covariant derivative of the metric tensor? Explain.
23. What does a connection coefficient $\Gamma_{\mu\nu}^\kappa$ mean physically? Is it a tensor? Why, or why not?
24. An astronaut is in free-fall in orbit around the Earth. Can the astronaut detect that there is a gravitational field?
25. Can a gravitational field exist in flat space?
26. How can you tell whether a given metric is equivalent to the Minkowski metric of flat space?
27. How many degrees of freedom does the metric have? How many of these degrees of freedom can be removed by arbitrary transformations of the spacetime coordinates, and therefore how many physical degrees of freedom are there in spacetime?
28. If you insist that the spacetime is spherical, how many physical degrees of freedom are there in the spacetime?
29. If you insist that the spacetime is spatially homogeneous and isotropic (the cosmological principle), how many physical degrees of freedom are there in the spacetime?
30. In general relativity, you are free to prescribe any spacetime (any metric) you like, including metrics with wormholes and metrics that connect the future to the past so as to violate causality. True or false?
31. If it is true that in general relativity you can prescribe any metric you like, then why aren't you bumping into wormholes and causality violations all the time?
32. How much mass does it take to curve space significantly (significantly meaning by of order unity)?
33. What is the relation between the energy-momentum 4-vector of a particle and the energy-momentum tensor?
34. It is straightforward to go from a prescribed metric to the energy-momentum tensor. True or false?
35. It is straightforward to go from a prescribed energy-momentum tensor to the metric. True or false?
36. Does the principle of equivalence imply Einstein's equations?
37. What do Einstein's equations mean physically?
38. What does the Riemann curvature tensor $R_{\kappa\lambda\mu\nu}$ mean physically? Is it a tensor?
39. The Riemann tensor splits into compressive (Ricci) and tidal (Weyl) parts. What do these parts mean, physically?
40. Einstein's equations imply conservation of energy-momentum, but what does that mean?
41. Do Einstein's equations describe gravitational waves?
42. Do photons (massless particles) gravitate?
43. How do different forms of mass-energy gravitate?
44. How does negative mass gravitate?

What's important?

1. The postulates of general relativity. How do the various postulates imply the mathematical structure of general relativity?
2. The road from spacetime curvature to energy-momentum:
metric $g_{\mu\nu}$
→ connection coefficients $\Gamma_{\mu\nu}^{\kappa}$
→ Riemann curvature tensor $R_{\kappa\lambda\mu\nu}$
→ Ricci tensor $R_{\kappa\mu}$ and scalar R
→ Einstein tensor $G_{\kappa\mu} = R_{\kappa\mu} - \frac{1}{2}g_{\kappa\mu}R$
→ energy-momentum tensor $T_{\kappa\mu}$
3. 4-velocity and 4-momentum. Geodesic equation.
4. Bianchi identities guarantee conservation of energy-momentum.

Fundamentals of General Relativity

As of writing (2013), general relativity continues to beat all-comers in the Darwinian struggle to be top theory of gravity and spacetime (Will, 2005). Despite its success, most physicists accept that general relativity cannot ultimately be correct, because of the difficulty in reconciling it with that other pillar of physics, quantum mechanics. The other three known forces of Nature, the electromagnetic, weak, and colour (strong) forces, are described by renormalizable quantum field theories, the so-called Standard Model of Physics, that agree extraordinarily well with experiment, and whose predictions have continued to be confirmed by ever more precise measurements. Attempts to quantize general relativity in a similar fashion fail. The attempt to unite general relativity and quantum mechanics continues to exercise some of the brightest minds in physics.

One place where general relativity predicts its own demise is at singularities inside black holes. What physics replaces general relativity at singularities? This is a deep question, providing one of the motivations for this book's emphasis on black hole interiors.

The aim of this Chapter is to give a condensed introduction to the fundamentals of general relativity, using the traditional coordinate-based approach to general relativity. The approach is neither the most insightful nor the most powerful, but it is the fastest route to connecting the metric to the energy-momentum content of spacetime. The Chapter does not attempt to convey a deep conceptual understanding, which I think is difficult to gain from the mathematics by itself. Later Chapters, starting with Chapter 7 on the Schwarzschild geometry, present visualizations intended to aid conceptual understanding.

One of the drawbacks of the coordinate approach is that it works with frames that are aligned at each point with the tangent vectors e_μ to the coordinates at that point. General relativity postulates the existence of locally inertial frames, so the coordinates at any point can always be arranged such that the tangent vectors at that one point are orthonormal, and the spacetime is locally flat (Minkowski) about that point. But in a curved spacetime it is impossible to arrange the coordinate tangent vectors e_μ to be orthonormal everywhere. Thus the coordinate approach inevitably presents quantities in a frame that is skewed compared to the natural, orthonormal frame. It is like looking at a scene with your eyes crossed. The problem is not so bad if the spacetime is empty of energy-momentum, as in the Schwarzschild and Kerr geometries for ideal spherical and rotating black holes, but it becomes a significant handicap in realistic spacetimes that contain energy-momentum.

The coordinate approach is adequate to deal with ideal black holes, Chapter 6 to 9, and with the Friedmann-

Lemaître-Robertson-Walker spacetime of a homogeneous, isotropic cosmology, Chapter 10. After that, the book restarts essentially from scratch. Chapter 11 introduces the tetrad formalism, the springboard for further explorations of gravity, black holes, and cosmology.

The convention in this book is that greek (brown) dummy indices label curved spacetime coordinates, while latin (black) dummy indices label locally inertial (more generally, tetrad) coordinates.

2.1 Motivation

Special relativity was unsatisfactory almost from the outset. Einstein had conceived special relativity by abolishing the aether. Yet for something that had no absolute substance, the spacetime of special relativity had strikingly absolute properties: in special relativity, two particles on parallel trajectories would remain parallel for ever, just as in Euclidean geometry.

Moreover whereas special relativity neatly accommodated the electromagnetic force, which propagated at the speed of light, it did not accommodate the other force known at the beginning of the 20th century, gravity. Plainly Newton's theory of gravity could not be correct, since it posited instantaneous transmission of the gravitational force, whereas special relativity seemed to preclude anything from moving faster than light, Exercise 1.23. You might think that gravity, an inverse square law like electromagnetism, might satisfy a similar set of equations, but this is not so. Whereas an electromagnetic wave carries no electric charge, and therefore does not interact with itself, any wave of gravity must carry energy, and therefore must interact with itself. This proves to be a considerable complication.

A partial solution, the principle of equivalence of gravity and acceleration, occurred to Einstein while working on an invited review on special relativity (Einstein, 1907). Einstein realised that “if a person falls freely, he will not feel his own weight,” an idea that Einstein would later refer to as “the happiest thought of my life.” The principle of equivalence meant that gravity could be reinterpreted as a curvature of spacetime. In this picture, the trajectories of two freely-falling particles that pass either side of a massive body are caused

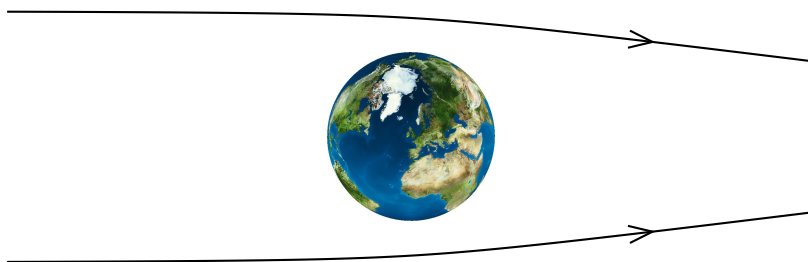


Figure 2.1 Particles initially on parallel trajectories passing either side of the Earth are caused to converge by the Earth's gravity. According to Einstein's principle of equivalence, the situation is equivalent to one where the particles are moving in straight lines in local free-fall frames. This allows the gravitational force to be reinterpreted as being produced by a curvature of spacetime induced by the presence of the Earth.

to converge not because of a gravitational force, but rather because the massive body curves spacetime, and the particles follow straight lines in the curved spacetime, Figure 2.1.

Einstein's principle of equivalence is only half the story. The principle of equivalence determines how particles must move in a spacetime of given curvature, but it does not determine how spacetime is itself curved by mass. That was a much more difficult problem, which Einstein took several more years to crack. The eventual solution was Einstein's equations, the final version of which he set out in a presentation to the Prussian Academy at the end of November 1915 (Einstein, 1915).

Contemporaneously with Einstein's discovery, David Hilbert derived Einstein's equations independently and elegantly from an action principle (Hilbert, 1915). In the present Chapter, Einstein's equations are simply postulated, since their real justification is that they reproduce experiment and observation. A derivation of Einstein's equations from the Hilbert action is deferred to Chapter 16.

2.2 The postulates of General Relativity

General relativity follows from three postulates:

1. Spacetime is a 4-dimensional differentiable manifold;
2. Einstein's principle of equivalence;
3. Einstein's equations.

2.2.1 Spacetime is a 4-dimensional differentiable manifold

A 4-dimensional **manifold** is defined mathematically to be a topological space that is locally homeomorphic to Euclidean 4-space $\mathbb{R}^4$. A homeomorphism is a continuous map that has a continuous inverse.

The postulate that spacetime is a 4-dimensional manifold means that it is possible to set up a coordinate system, possibly in patches, called **charts**,

$$x^\mu \equiv \{x^0, x^1, x^2, x^3\} \quad (2.1)$$

such that each point of a chart of the spacetime has a unique coordinate.

It is not always possible to cover a manifold with a single chart, that is, with a coordinate system such that every point of spacetime has a unique coordinate. A simple example of a 2-dimensional manifold that cannot be covered with a single chart is the 2-sphere S^2 , the 2-dimensional surface of a 3-dimensional sphere, as illustrated in Figure 2.2. Inevitably, lines of constant coordinate must cross somewhere on the 2-sphere. At least two charts are required to cover a 2-sphere.

When more than one chart is necessary, neighbouring charts are required to overlap, in order that the structure of the manifold be consistent across the overlap. General relativity postulates that the mapping between the coordinates of overlapping charts be at least doubly differentiable. A manifold subject to this property is called differentiable.

In practice one often uses coordinate systems that misbehave at some points, but in an innocuous fashion. The 2-sphere again provides a classic example, where the standard choice of polar coordinates $x^\mu = \{\theta, \phi\}$

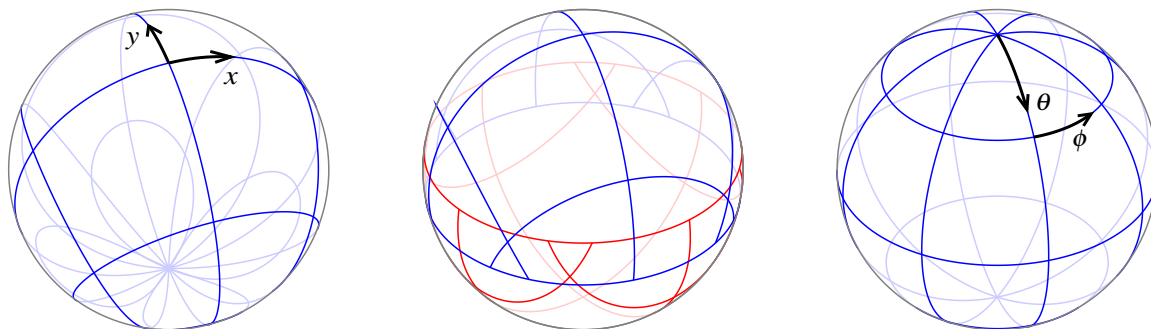


Figure 2.2 The 2-sphere is a 2-manifold, a topological space that is locally homeomorphic to Euclidean 2-space $\mathbb{R}^2$. Any attempt to cover the surface of a 2-sphere with a single chart, that is, with coordinates x and y such that each point on the sphere is specified by a unique coordinate $\{x, y\}$, fails at at least one point. In the left panel, a coordinate grid draped over the sphere fails at one point, the south pole, where coordinate lines cross. At least two charts are required to cover the surface of a 2-sphere, as illustrated in the middle panel, where one chart covers the north pole, the other the south pole. Where the two charts overlap, the two sets of coordinates are related differentially. The right panel shows standard polar coordinates θ, ϕ on the 2-sphere. The polar coordinatization fails at the north and south poles, where lines of longitude cross, the azimuthal angle ϕ is not unique, and a person passing smoothly through the pole would see the azimuthal angle jump by π . Such misbehaving points, called coordinate singularities, are however innocuous: they can be removed by cutting out a patch around the coordinate singularity, and pasting on a separate chart.

misbehaves at the north and south poles, Figure 2.2. A person passing smoothly through a pole sees the azimuthal coordinate jump discontinuously by π . This is called a coordinate singularity. It is innocuous because it can be removed by excising a patch around the pole, and pasting on a separate chart.

2.2.2 Principle of equivalence

The weak principle of equivalence states that: “Gravitating mass equals inertial mass.” General relativity satisfies the weak principle of equivalence, but then so also does Newtonian gravity.

Einstein’s principle of equivalence is actually two separate statements: “The laws of physics in a gravitating frame are equivalent to those in an accelerating frame,” and “The laws of physics in a non-accelerating, or free-fall, frame are locally those of special relativity.”

Einstein’s principle of equivalence implies that it is possible to remove the effects of gravity locally by going into a non-accelerating, or free-fall, frame. The structure of spacetime in a non-accelerating, or free-fall, frame is locally inertial, with the local structure of Minkowski space. By locally inertial is meant that at each point of spacetime it is possible to choose coordinates such that (a) the metric at that point is Minkowski, and (b) the first derivatives of the metric are all zero¹. In other words, Einstein’s principle of equivalence asserts the **existence of locally inertial frames**.

¹ Actually, general relativity goes a step further. The metric is the scalar product of coordinate tangent axes, equation (2.26). General relativity postulates, §2.10.1, that in a locally inertial frame the first derivatives not only of the metric, but also of the tangent axes themselves, vanish. See also Concept question 2.6.

Since special relativity is a metric theory, and the principle of equivalence asserts that general relativity looks locally like special relativity, general relativity inherits from special relativity the property of being a **metric theory**. A notable consequence is that the proper times and distances measured by an accelerating observer are the same as those measured by a freely-falling observer at the same point and with the same instantaneous velocity.

2.2.3 Einstein's equations

Einstein's equations comprise a 4×4 symmetric matrix of equations

$$\boxed{G_{\mu\nu} = 8\pi G T_{\mu\nu}} . \quad (2.2)$$

Here G is the Newtonian gravitational constant, $G_{\mu\nu}$ is the **Einstein tensor**, and $T_{\mu\nu}$ is the **energy-momentum tensor**.

Physically, Einstein's equations signify

$$(\text{compressive part of}) \text{ curvature} = \text{energy-momentum content} . \quad (2.3)$$

Einstein's equations generalize Poisson's equation

$$\nabla^2 \Phi = 4\pi G \rho \quad (2.4)$$

where Φ is the Newtonian gravitational potential, and ρ the mass-energy density. Poisson's equation is the time-time component of Einstein's equations in the limit of a weak gravitational field and slowly moving matter, §2.28.

2.3 Implications of Einstein's principle of equivalence

2.3.1 The gravitational redshift of light

Einstein's principle of equivalence implies that light will redshift in a gravitational field. In a weak gravitational field, the gravitational redshift of light can be deduced quantitatively from the equivalence principle without any further assumption (such as Einstein's equations), Exercises 2.1 and 2.2. A fully general relativistic treatment for the redshift between observers at rest in a stationary gravitational field is given in Exercise 2.10.

Exercise 2.1 The equivalence principle implies the gravitational redshift of light, Part 1. A rigorous general relativistic version of this exercise is Exercise 2.11. A person standing at rest on the surface of the Earth is to a good approximation in a uniform gravitational field, with gravitational acceleration g . The principle of equivalence asserts that the situation is equivalent to that of a frame uniformly accelerating at g . Assume that the non-accelerating, free-fall frame is Minkowski to a good approximation. Define the

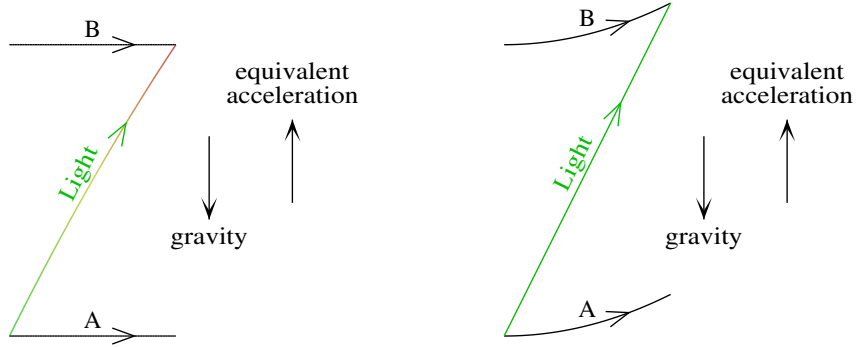


Figure 2.3 Einstein's principle of equivalence implies the gravitational redshift of light, and the gravitational bending of light. In the left panel, persons A and B are at rest relative to each other in a uniform gravitational field. They are shown moving to the right to bring out the evolution of the system in time. A sends a beam of light upward to B. The principle of equivalence asserts that the uniform gravitational field is equivalent to a uniformly accelerating frame. The right panel shows the equivalent uniformly accelerating situation as perceived by a person in free-fall. In the free-fall frame, the light moves on a straight line, and has constant frequency. Back in the gravitating/accelerating frame in the left panel, the light appears to bend, and to redshift as it climbs from A to B.

potential Φ by the usual Newtonian formula $g = -\nabla\Phi$. Show that for small differences in their gravitational potentials, B perceives the light emitted by A to be redshifted by (with units restored)

$$z = \frac{\Phi_{\text{obs}} - \Phi_{\text{em}}}{c^2} . \quad (2.5)$$

Exercise 2.2 The equivalence principle implies the gravitational redshift of light, Part 2.

A rigorous general relativistic version of this exercise is Exercise 2.12. Consider a person who, at rest in Minkowski space, whirls a clock around them on the end of string, so fast that the clock is moving at near the speed of light. The person sees the clock redshifted by the Lorentz γ -factor (the string is of fixed length, so the light travel time from clock to observer is always the same, and does not affect the redshift). Tugged on by the string, the clock experiences a centripetal acceleration towards the whirling person. According to the principle of equivalence, the centripetal acceleration is equivalent to a centrifugal gravitational force. In a Newtonian approximation, if the clock is whirling around at angular velocity ω , then the effective centrifugal potential at radius r from the observer is

$$\Phi = -\frac{1}{2}\omega^2 r^2 . \quad (2.6)$$

Show that, for nonrelativistic velocities $\omega r \ll c$, the observer perceives the light emitted from the clock to be redshifted by (with units restored)

$$z = -\frac{\Phi}{c^2} . \quad (2.7)$$

2.3.2 The gravitational bending of light

The principle of equivalence also implies that light will appear to bend in a gravitational field, as illustrated by Figure 2.3. However, a quantitative prediction for the bending of light requires full general relativity. The bending of light in a weak gravitational field is the subject of Exercise 2.18.

2.4 Metric

Postulate (1), §2.2.1, of general relativity means that it is possible to choose coordinates

$$x^\mu \equiv \{x^0, x^1, x^2, x^3\} \quad (2.8)$$

covering a patch of spacetime.

Postulate (2), §2.2.2, of general relativity implies that at each point of spacetime it is possible to choose locally inertial coordinates

$$\xi^m \equiv \{\xi^0, \xi^1, \xi^2, \xi^3\} \quad (2.9)$$

such that the metric is Minkowski,

$$ds^2 = \eta_{mn} d\xi^m d\xi^n, \quad (2.10)$$

in an infinitesimal neighbourhood of the point. Infinitesimal neighbourhood means that the metric is the Minkowski metric η_{mn} at the point, and that the first derivatives of the metric vanish at the point. The spacetime distance squared ds^2 is a **scalar**, a quantity that is unchanged by the choice of coordinates. Whereas in special relativity the Minkowski formula (1.32) for the spacetime distance Δs^2 held for finite intervals Δx^m , in general relativity the metric formula (2.10) holds only for infinitesimal intervals $d\xi^m$.

General relativity requires, postulate (1), that two sets of coordinates are differentially related, so locally inertial intervals $d\xi^m$ and coordinate intervals dx^μ are related by the Leibniz rule,

$$d\xi^m = \frac{\partial \xi^m}{\partial x^\mu} dx^\mu. \quad (2.11)$$

It follows that the scalar spacetime distance squared is

$$ds^2 = \eta_{mn} \frac{\partial \xi^m}{\partial x^\mu} \frac{\partial \xi^n}{\partial x^\nu} dx^\mu dx^\nu, \quad (2.12)$$

which can be written in terms of coordinate intervals dx^μ as

$$\boxed{ds^2 = g_{\mu\nu} dx^\mu dx^\nu}, \quad (2.13)$$

where $g_{\mu\nu}$ is the **metric**, a 4×4 symmetric matrix

$$g_{\mu\nu} = \eta_{mn} \frac{\partial \xi^m}{\partial x^\mu} \frac{\partial \xi^n}{\partial x^\nu}. \quad (2.14)$$

The metric is the essential mathematical object that converts an infinitesimal interval dx^μ to a proper measurement of an interval of time or space.

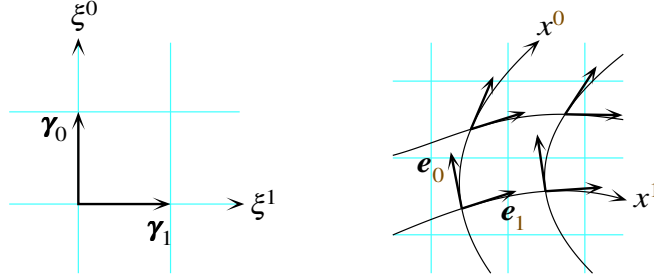


Figure 2.4 (Left) The tetrad vectors γ_m form an orthonormal basis of vectors tangent to a set of locally inertial coordinates ξ^m at a point. (Right) The coordinate tangent vectors e_μ are the basis of vectors tangent to the coordinates at each point. The background square grid represents a locally inertial frame, the existence of which is asserted by general relativity.

2.5 Timelike, spacelike, proper time, proper distance

General relativity inherits from special relativity the physical meaning of the scalar spacetime distance squared ds^2 along an interval dx^μ . The scalar spacetime distance squared can be negative, zero, or positive, and accordingly timelike, lightlike, or spacelike:

$$\begin{aligned} \text{timelike:} \quad & ds^2 < 0, \quad d\tau = \sqrt{-ds^2} = \text{interval of proper time}, \\ \text{lightlike:} \quad & ds^2 = 0, \\ \text{spacelike:} \quad & ds^2 > 0, \quad dl = \sqrt{ds^2} = \text{interval of proper distance}. \end{aligned} \quad (2.15)$$

2.6 Orthonormal tetrad basis γ_m

You are familiar with the idea that in ordinary 3-dimensional Euclidean geometry it is often convenient to treat vectors in an abstract coordinate-independent formalism. Thus for example a 3-vector is commonly written as an abstract quantity $\mathbf{r}$. The coordinates of the vector $\mathbf{r}$ may be $\{x, y, z\}$ in some particular coordinate system, but one recognizes that the vector $\mathbf{r}$ has a meaning, a magnitude and a direction, that is independent of the coordinate system adopted. In an arbitrary Cartesian coordinate system, the Euclidean 3-vector $\mathbf{r}$ can be expressed

$$\mathbf{r} = \sum_a \hat{\mathbf{x}}_a x_a = \hat{\mathbf{x}} x + \hat{\mathbf{y}} y + \hat{\mathbf{z}} z \quad (2.16)$$

where $\hat{\mathbf{x}}_a \equiv \{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}\}$ are unit vectors along each of the coordinate axes. The unit vectors satisfy a Euclidean metric

$$\hat{\mathbf{x}}_a \cdot \hat{\mathbf{x}}_b = \delta_{ab}. \quad (2.17)$$

The same kind of abstract notation is useful in general relativity. Because the spacetime of general relativity is only locally inertial, not globally inertial, vectors must be thought of as living not in the spacetime manifold

itself, but rather in the **tangent space** of the manifold. The existence and structure of such a tangent space follows from the postulate of the existence of locally inertial frames. Let ξ^m be a set of locally inertial coordinates at a point of spacetime. Define the vectors γ_m , called a **tetrad**, to be tangent to the locally inertial coordinates at the point in question,

$$\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\} , \quad (2.18)$$

as illustrated in the left panel of Figure 2.4. Each tetrad basis vector γ_m is a 4-dimensional object, with both magnitude and direction. The basis vectors γ_m are introduced so that vectors in spacetime can be expressed in an abstract coordinate-independent fashion. The prototypical vector is an infinitesimal interval $d\xi^m$ of spacetime, which can be expressed in coordinate-independent fashion as the abstract vector interval $d\mathbf{x}$ defined by

$$d\mathbf{x} \equiv \gamma_m d\xi^m = \gamma_0 d\xi^0 + \gamma_1 d\xi^1 + \gamma_2 d\xi^2 + \gamma_3 d\xi^3 . \quad (2.19)$$

The interval $d\xi^m$ transforms under a Lorentz transformation of the locally inertial coordinates as a contravariant Lorentz vector. To make the abstract vector interval $d\mathbf{x}$ invariant under Lorentz transformation, the basis vectors γ_m must transform as a covariant Lorentz vector.

The scalar length squared of the abstract vector interval $d\mathbf{x}$ is

$$ds^2 = d\mathbf{x} \cdot d\mathbf{x} = \gamma_m \cdot \gamma_n d\xi^m d\xi^n . \quad (2.20)$$

Since this must reproduce the locally inertial metric (2.10), the scalar products of the tetrad vectors γ_m must form the Minkowski metric

$$\gamma_m \cdot \gamma_n = \eta_{mn} . \quad (2.21)$$

A basis of tetrad vectors whose scalar products form the Minkowski metric is called **orthonormal**.

Tetrads are explored in depth in Chapter 11.

2.7 Basis of coordinate tangent vectors e_μ

In general relativity, coordinates can be chosen arbitrarily, subject to differentiability conditions. In an arbitrary system of coordinates x^μ , the **coordinate tangent vectors** e_μ at each point,

$$e_\mu \equiv \{e_0, e_1, e_2, e_3\} , \quad (2.22)$$

are defined to satisfy

$$d\mathbf{x} \equiv e_\mu dx^\mu = \gamma_m d\xi^m . \quad (2.23)$$

The letter e derives from the German word *einheit*, meaning unity. The relation (2.11) between coordinate intervals dx^μ and locally inertial coordinate intervals $d\xi^m$ implies that the coordinate tangent vectors e_μ must be related to the orthonormal tetrad vectors γ_m by

$$e_\mu = \gamma_m \frac{\partial \xi^m}{\partial x^\mu} . \quad (2.24)$$

Like the tetrad axes γ_m , each coordinate tangent axis e_μ is a 4-dimensional vector object, with both magnitude and direction, as illustrated in the right panel of Figure 2.4.

The scalar length squared of the abstract vector interval $d\mathbf{x}$ is

$$ds^2 = d\mathbf{x} \cdot d\mathbf{x} = e_\mu \cdot e_\nu dx^\mu dx^\nu, \quad (2.25)$$

from which it follows that the scalar products of the coordinate tangent axes e_μ must equal the coordinate metric $g_{\mu\nu}$,

$$\boxed{g_{\mu\nu} = e_\mu \cdot e_\nu}. \quad (2.26)$$

Like the orthonormal tetrad vectors γ_m , the coordinate tangent vectors e_μ form a basis for the 4-dimensional tangent space at each point. The tangent space has three basic mathematical properties. First, the tangent space is a vector space, that is, it has the properties of linearity that define a vector space. Second, the tangent space has an inner (or scalar) product, defined by the metric (2.26). That scalar product is a consequence of the postulated locally inertial, or Lorentz, structure of spacetime, which asserts that the metric is Minkowski η_{mn} with respect to locally inertial coordinates ξ^m . Third, vectors e_μ in the tangent space can be differentiated with respect to coordinates x^ν , as will be elucidated in §2.9.3.

Some texts represent the tangent vectors e_μ with the notation ∂_μ , on the grounds that e_μ transforms like the coordinate derivatives $\partial_\mu \equiv \partial/\partial x^\mu$. This notation is **not** used in this book, to avoid the potential confusion between ∂_μ as a derivative and ∂_μ as a vector.

2.8 4-vectors and tensors

2.8.1 Contravariant coordinate 4-vector

Under a general coordinate transformation

$$x^\mu \rightarrow x'^\mu, \quad (2.27)$$

a coordinate interval dx^μ transforms as

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} dx^\nu. \quad (2.28)$$

In general relativity, a **coordinate 4-vector** is defined to be a quantity $A^\mu = \{A^0, A^1, A^2, A^3\}$ that transforms under a coordinate transformation (2.27) like a coordinate interval

$$\boxed{A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu}. \quad (2.29)$$

Just because something has an index on it does not make it a 4-vector. The essential property of a contravariant coordinate 4-vector is that it transforms like a coordinate interval, equation (2.29).

2.8.2 Abstract 4-vector

A 4-vector may be written in coordinate-independent fashion as

$$\mathbf{A} = \mathbf{e}_\mu A^\mu . \quad (2.30)$$

The quantity $\mathbf{A}$ is an **abstract 4-vector**. Although $\mathbf{A}$ is a 4-vector, it is by construction unchanged by a coordinate transformation, and is therefore a coordinate scalar. See §2.8.7 for commentary on the distinction between abstract and coordinate vectors.

2.8.3 Lowering and raising indices

Define $g^{\mu\nu}$ to be the **inverse metric**, satisfying

$$g_{\lambda\mu} g^{\mu\nu} = \delta_\lambda^\nu = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad (2.31)$$

The metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ provide the means of lowering and raising coordinate indices. The components of a coordinate 4-vector A^μ with raised index are called its **contravariant** components, while those A_μ with lowered indices are called its **covariant** components,

$$\boxed{A_\mu = g_{\mu\nu} A^\nu} , \quad (2.32)$$

$$\boxed{A^\mu = g^{\mu\nu} A_\nu} . \quad (2.33)$$

2.8.4 Dual basis $\mathbf{e}^\mu$

The contravariant dual basis elements $\mathbf{e}^\mu$ are defined by raising the indices of the covariant tangent basis elements $\mathbf{e}_\nu$,

$$\mathbf{e}^\mu \equiv g^{\mu\nu} \mathbf{e}_\nu . \quad (2.34)$$

You can check that the dual vectors $\mathbf{e}^\mu$ transform as a contravariant coordinate 4-vector. The dot products of the dual basis elements $\mathbf{e}^\mu$ with each other are

$$\mathbf{e}^\mu \cdot \mathbf{e}^\nu = g^{\mu\nu} . \quad (2.35)$$

The dot products of the dual and tangent basis elements are

$$\mathbf{e}^\mu \cdot \mathbf{e}_\nu = \delta_\nu^\mu . \quad (2.36)$$

2.8.5 Covariant coordinate 4-vector

Under a general coordinate transformation (2.27), the covariant components A_μ of a coordinate 4-vector transform as

$$A'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu . \quad (2.37)$$

You can check that the transformation law (2.37) for the covariant components A_μ is consistent with the transformation law (2.29) for the contravariant components A^μ .

You can check that the tangent vectors e_μ transform as a covariant coordinate 4-vector.

2.8.6 Scalar product

If A^μ and B^μ are coordinate 4-vectors, then their scalar product is

$$A_\mu B^\mu = A^\mu B_\mu = g_{\mu\nu} A^\mu B^\nu . \quad (2.38)$$

This is a coordinate scalar, a quantity that remains invariant under general coordinate transformations. The ability to form a scalar by contracting over paired indices, always one raised and one lowered, is what makes the introduction of two species of vector, contravariant (raised index) and covariant (lowered index), so advantageous.

In abstract vector formalism, the scalar product of two 4-vectors $\mathbf{A} = e_\mu A^\mu$ and $\mathbf{B} = e_\mu B^\mu$ is

$$\mathbf{A} \cdot \mathbf{B} = e_\mu \cdot e_\nu A^\mu B^\nu = g_{\mu\nu} A^\mu B^\nu . \quad (2.39)$$

2.8.7 Comment on vector naming and notation

Different texts follow different conventions for naming and notating vectors and tensors.

This book follows the convention of calling both A^μ (with a dummy index μ) and $\mathbf{A} \equiv A^\mu e_\mu$ vectors. Although A^μ and $\mathbf{A}$ are both vectors, they are mathematically different objects.

If the index on a vector indicates a specific coordinate, then the indexed vector is the component of the vector; for example A^0 (or A^t) is the x^0 (or time t) component of the coordinate 4-vector A^μ .

In this book, the different species of vector are distinguished by an adjective:

1. A **coordinate vector** A^μ , identified by greek (brown) indices μ , is one that changes in a prescribed way under coordinate transformations. A coordinate transformation is one that changes the coordinates of the spacetime without actually changing the spacetime or whatever lies in it.
2. An **abstract vector** $\mathbf{A}$, identified by boldface, is the thing itself, and is unchanged by the choice of coordinates. Since the abstract vector is unchanged by a coordinate transformation, it is a coordinate scalar.

All the types of vector have the properties of linearity (additivity, multiplication by scalars) that identify them mathematically as belonging to vector spaces. The important distinction between the types of vector is how they behave under transformations.

In referring to both A^μ and $\mathbf{A}$ as vectors, this book follows the standard physics practice of mentally

regarding A^μ and $\mathbf{A}$ as equivalent objects. You are familiar with the advantages of treating a vector in 3-dimensional Euclidean space either as an abstract vector $\mathbf{A}$, or as a coordinate vector A_a . Depending on the problem, sometimes the abstract notation $\mathbf{A}$ is more convenient, and sometimes the coordinate notation A_a is more convenient. Sometimes it's convenient to switch between the two in the middle of a calculation. Likewise in general relativity it is convenient to have the flexibility to work in either coordinate or abstract notation, whatever suits the problem of the moment.

2.8.8 Coordinate tensor

In general, a **coordinate tensor** $A_{\mu\nu\dots}^{\kappa\lambda\dots}$ is an object that transforms under general coordinate transformations (2.27) as

$$A_{\mu\nu\dots}^{\kappa\lambda\dots} = \frac{\partial x'^\kappa}{\partial x^\pi} \frac{\partial x'^\lambda}{\partial x^\rho} \dots \frac{\partial x^\sigma}{\partial x'^\mu} \frac{\partial x^\tau}{\partial x'^\nu} \dots A_{\sigma\tau\dots}^{\pi\rho\dots}. \quad (2.40)$$

You can check that the metric tensor $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ are indeed coordinate tensors, transforming like (2.40).

The **rank** of a tensor is the number of indices of its expansion $A_{\mu\nu\dots}^{\kappa\lambda\dots}$ in components. A scalar is a tensor of rank 0. A 4-vector is a tensor of rank 1. The metric and its inverse are tensors of rank 2. The rank of a tensor with n contravariant (upstairs) and m covariant (downstairs) indices is sometimes written $\begin{pmatrix} n \\ m \end{pmatrix}$.

2.9 Covariant derivatives

2.9.1 Derivative of a coordinate scalar

Suppose that Φ is a coordinate scalar. Then the coordinate derivative of Φ is a coordinate 4-vector

$$\boxed{\frac{\partial \Phi}{\partial x^\mu}} \quad \text{a coordinate tensor} \quad (2.41)$$

transforming like equation (2.37).

As a shorthand, the ordinary partial derivative is often denoted in the literature with a comma

$$\frac{\partial \Phi}{\partial x^\mu} = \Phi_{,\mu}. \quad (2.42)$$

For the most part this book does not use the comma notation.

2.9.2 Derivative of a coordinate 4-vector

The ordinary partial derivative of a contravariant coordinate 4-vector A^μ is not a tensor

$$\frac{\partial A^\mu}{\partial x^\nu} \quad \text{not a coordinate tensor} \quad (2.43)$$

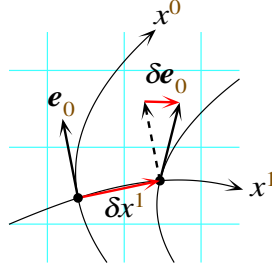


Figure 2.5 The change δe_0 in the tangent vector e_0 over a small interval δx^1 of spacetime is defined to be the difference between the tangent vector $e_0(x^1 + \delta x^1)$ at the shifted position $x^1 + \delta x^1$ and the tangent vector $e_0(x^1)$ at the original position x^1 , parallel-transported to the shifted position. The parallel-transported vector is shown as a dashed arrowed line. The parallel transport is defined with respect to a locally inertial frame, shown as a background square grid.

because it does not transform like a coordinate tensor.

However, the 4-vector $\mathbf{A} = e_\mu A^\mu$, being by construction invariant under coordinate transformations, is a coordinate scalar, and its partial derivative is a coordinate 4-vector

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial x^\nu} &= \frac{\partial e_\mu A^\mu}{\partial x^\nu} \\ &= e_\mu \frac{\partial A^\mu}{\partial x^\nu} + \frac{\partial e_\mu}{\partial x^\nu} A^\mu \quad \text{a coordinate tensor.} \end{aligned} \quad (2.44)$$

The last line of equation (2.44) assumes that it is legitimate to differentiate the tangent vectors e_μ , but what does that mean? The partial derivatives of basis vectors e_μ are defined in the usual way by

$$\frac{\partial e_\mu}{\partial x^\nu} \equiv \lim_{\delta x^\nu \rightarrow 0} \frac{e_\mu(x^0, \dots, x^\nu + \delta x^\nu, \dots, x^3) - e_\mu(x^0, \dots, x^\nu, \dots, x^3)}{\delta x^\nu}. \quad (2.45)$$

This definition relies on being able to compare the vectors $e_\mu(x)$ at some point x with the vectors $e_\mu(x + \delta x)$ at another point $x + \delta x$ a small distance away. The comparison between two vectors a small distance apart is made possible by the existence of locally inertial frames. In a locally inertial frame, two vectors a small distance apart can be compared by **parallel-transporting** one vector to the location of the other along the small interval between them, that is, by transporting the vector without accelerating or precessing with respect to the locally inertial frame. Thus the right hand side of equation (2.45) should be interpreted as $e_\mu(x + \delta x)$ minus the value of $e_\mu(x)$ parallel-transported from position x to position $x + \delta x$ along the small interval δx between them, as illustrated in Figure 2.5.

The notion of the tangent space at a point on a manifold was introduced in §2.6. Parallel transport allows the tangent spaces at neighbouring points to be adjoined in a well-defined fashion to form the **tangent manifold**, whose dimension is twice that of the underlying spacetime. Coordinates for the tangent manifold are provided by a combination $\{x^\mu, \xi^m\}$ of coordinates x^μ on the parent manifold and tangent space coordinates ξ^m extrapolated from a locally inertial frame about each point. The tangent space coordinates ξ^m vary smoothly over the manifold provided that the locally inertial frames are chosen to vary smoothly.

2.9.3 Coordinate connection coefficients

The partial derivatives of the basis vectors $\mathbf{e}_\mu$ that appear on the right hand side of equation (2.44) define the **coordinate connection coefficients** $\Gamma_{\mu\nu}^\kappa$,

$$\boxed{\frac{\partial \mathbf{e}_\mu}{\partial x^\nu} \equiv \Gamma_{\mu\nu}^\kappa \mathbf{e}_\kappa} \quad \text{not a coordinate tensor .} \quad (2.46)$$

The definition (2.46) shows that the connection coefficients express how each tangent vector $\mathbf{e}_\mu$ changes, relative to parallel-transport, when shifted along an interval δx^ν .

2.9.4 Covariant derivative of a contravariant 4-vector

Expression (2.44) along with the definition (2.46) of the connection coefficients implies that

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial x^\nu} &= \mathbf{e}_\mu \frac{\partial A^\mu}{\partial x^\nu} + \Gamma_{\mu\nu}^\kappa \mathbf{e}_\kappa A^\mu \\ &= \mathbf{e}_\kappa \left(\frac{\partial A^\kappa}{\partial x^\nu} + \Gamma_{\mu\nu}^\kappa A^\mu \right) \quad \text{a coordinate tensor .} \end{aligned} \quad (2.47)$$

The expression in parentheses is a coordinate tensor, and defines the **covariant derivative** $D_\nu A^\kappa$ of the contravariant coordinate 4-vector A^κ

$$\boxed{D_\nu A^\kappa \equiv \frac{\partial A^\kappa}{\partial x^\nu} + \Gamma_{\mu\nu}^\kappa A^\mu} \quad \text{a coordinate tensor .} \quad (2.48)$$

As a shorthand, the covariant derivative is often denoted in the literature with a semi-colon

$$D_\nu A^\kappa = A^\kappa_{;\nu} . \quad (2.49)$$

For the most part this book does not use the semi-colon notation.

2.9.5 Covariant derivative of a covariant coordinate 4-vector

Similarly,

$$\frac{\partial \mathbf{A}}{\partial x^\nu} = \mathbf{e}^\kappa D_\nu A_\kappa \quad \text{a coordinate tensor} \quad (2.50)$$

where $D_\nu A_\kappa$ is the covariant derivative of the covariant coordinate 4-vector A_κ

$$\boxed{D_\nu A_\kappa \equiv \frac{\partial A_\kappa}{\partial x^\nu} - \Gamma_{\kappa\nu}^\mu A_\mu} \quad \text{a coordinate tensor .} \quad (2.51)$$

2.9.6 Covariant derivative of a coordinate tensor

In general, the covariant derivative of a coordinate tensor is

$$D_\pi A_{\mu\nu\dots}^{\kappa\lambda\dots} = \frac{\partial A_{\mu\nu\dots}^{\kappa\lambda\dots}}{\partial x^\pi} + \Gamma_{\rho\pi}^\kappa A_{\mu\nu\dots}^{\rho\lambda\dots} + \Gamma_{\rho\pi}^\lambda A_{\mu\nu\dots}^{\kappa\rho\dots} + \dots - \Gamma_{\mu\pi}^\rho A_{\rho\nu\dots}^{\kappa\lambda\dots} - \Gamma_{\nu\pi}^\rho A_{\mu\rho\dots}^{\kappa\lambda\dots} - \dots \quad (2.52)$$

with a positive Γ term for each contravariant index, and a negative Γ term for each covariant index.

Concept question 2.3 Does covariant differentiation commute with the metric? **Answer.** Yes, essentially by construction. The covariant derivative of a tangent basis vector e_μ ,

$$D_\nu e_\mu = \frac{\partial e_\mu}{\partial x^\nu} - \Gamma_{\mu\nu}^\kappa e_\kappa = 0 , \quad (2.53)$$

vanishes by definition of the coordinate connections, equation (2.46). Consequently the covariant derivative of the metric $g_{\mu\nu} \equiv e_\mu \cdot e_\nu$ also vanishes. As a corollary, covariant differentiation commutes with the operations of raising and lowering indices, and of contraction.

2.10 Torsion

2.10.1 No-torsion condition

The postulate of the existence of locally inertial frames requires that it must be possible to arrange not only that the tangent axes e_μ are orthonormal at a point, but also that they remain orthonormal to first order in a Taylor expansion about the point. That is, it must be possible to choose the coordinates such that the tangent axes e_μ are orthonormal, and unchanged to linear order:

$$e_\mu \cdot e_\nu = \eta_{\mu\nu} , \quad (2.54a)$$

$$\frac{\partial e_\mu}{\partial x^\nu} = 0 . \quad (2.54b)$$

In view of the definition (2.46) of the connection coefficients, the second condition (2.54b) is equivalent to the vanishing of all the connection coefficients:

$$\Gamma_{\mu\nu}^\kappa = 0 . \quad (2.55)$$

Under a general coordinate transformation $x^\mu \rightarrow x'^\mu$, the tangent axes transform as $e_\mu = \partial x'^\kappa / \partial x^\mu e'_\kappa$. The 4×4 matrix $\partial x'^\kappa / \partial x^\mu$ of partial derivatives provides 16 degrees of freedom in choosing the tangent axes at a point. The 16 degrees of freedom are enough — more than enough — to accomplish the orthonormality condition (2.54a), which is a symmetric 4×4 matrix equation with 10 degrees of freedom. The additional $16 - 10 = 6$ degrees of freedom are Lorentz transformations, which rotate the tangent axes e_μ , but leave the metric $\eta_{\mu\nu}$ unchanged.

Just as it is possible to reorient the tangent axes e_μ at a point by adjusting the matrix $\partial x'^\kappa / \partial x^\mu$ of first

partial derivatives of the coordinate transformation $x^\mu \rightarrow x'^\mu$, so also it is possible to reorient the derivatives $\partial \mathbf{e}_\mu / \partial x^\nu$ of the tangent axes by adjusting the matrix $\partial^2 x'^\kappa / \partial x^\mu \partial x^\nu$ of second partial derivatives of the coordinate transformation. The second partial derivatives comprise a set of 4 symmetric 4×4 matrices, for a total of $4 \times 10 = 40$ degrees of freedom. However, there are $4 \times 4 \times 4 = 64$ connection coefficients $\Gamma_{\mu\nu}^\kappa$, all of which the condition (2.55) requires to vanish. The matrix of second derivatives is thus $64 - 40 = 24$ degrees of freedom short of being able to make all the connections vanish. The resolution of the problem is that, as shown below, equation (2.58), there are 24 combinations of the connections that form a tensor, the torsion tensor. If a tensor is zero in one frame, then it is automatically zero in any other frame. Thus the requirement that all the connections vanish requires that the torsion tensor vanish. This requires, from the expression (2.58) for the torsion tensor, the no-torsion condition that the connection coefficients are symmetric in their last two indices

$$\boxed{\Gamma_{\mu\nu}^\kappa = \Gamma_{\nu\mu}^\kappa} . \quad (2.56)$$

It should be emphasized that the condition of vanishing torsion is an assumption of general relativity, not a mathematical necessity. It has been shown in this section that torsion vanishes if and only if spacetime is locally flat, meaning that at any point coordinates can be found such that conditions (2.54) are true. The assumption of local flatness is central to the idea of the principle of equivalence. But it is an assumption, not a consequence, of the theory.

Concept question 2.4 Parallel transport when torsion is present. If torsion does not vanish, then there is no locally inertial frame. What does parallel-transport mean in such a case? **Answer.** A general coordinate transformation can always be found such that the connection coefficients $\Gamma_{\mu\nu}^\kappa$ vanish along any one direction ν . Parallel-transport along that direction can be defined relative to such a frame. For any given direction ν , there are 16 second partial derivatives $\partial^2 x'^\kappa / \partial x^\mu \partial x^\nu$, just enough to make vanish the $4 \times 4 = 16$ coefficients $\Gamma_{\mu\nu}^\kappa$.

2.10.2 Torsion tensor

General relativity assumes no torsion, but it is possible to consider generalizations to theories with torsion. The **torsion tensor** $S_{\kappa\lambda}^\mu$ is defined by the commutator of the covariant derivative acting on a scalar Φ

$$\boxed{[D_\kappa, D_\lambda] \Phi = S_{\kappa\lambda}^\mu \frac{\partial \Phi}{\partial x^\mu}} \quad \text{a coordinate tensor} . \quad (2.57)$$

Note that the covariant derivative of a scalar is just the ordinary derivative, $D_\lambda \Phi = \partial \Phi / \partial x^\lambda$. The expression (2.51) for the covariant derivatives shows that the torsion tensor is

$$\boxed{S_{\kappa\lambda}^\mu = \Gamma_{\kappa\lambda}^\mu - \Gamma_{\lambda\kappa}^\mu} \quad \text{a coordinate tensor} \quad (2.58)$$

which is evidently antisymmetric in the indices $\kappa\lambda$.

In Einstein-Cartan theory, the torsion tensor is related to the spin content of spacetime. Since this vanishes

in empty space, Einstein-Cartan theory is indistinguishable from general relativity in experiments carried out in vacuum. See §16.11 for more on Einstein-Cartan theory.

2.11 Connection coefficients in terms of the metric

The connection coefficients have been defined, equation (2.46), as derivatives of the tangent basis vectors $\mathbf{e}_\mu$. However, the connection coefficients can be expressed purely in terms of the (first derivatives of the) metric, without reference to the individual basis vectors. The partial derivatives of the metric are

$$\begin{aligned} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} &= \frac{\partial \mathbf{e}_\lambda \cdot \mathbf{e}_\mu}{\partial x^\nu} \\ &= \mathbf{e}_\lambda \cdot \frac{\partial \mathbf{e}_\mu}{\partial x^\nu} + \mathbf{e}_\mu \cdot \frac{\partial \mathbf{e}_\lambda}{\partial x^\nu} \\ &= \mathbf{e}_\lambda \cdot \mathbf{e}_\kappa \Gamma_{\mu\nu}^\kappa + \mathbf{e}_\mu \cdot \mathbf{e}_\kappa \Gamma_{\lambda\nu}^\kappa \\ &= g_{\lambda\kappa} \Gamma_{\mu\nu}^\kappa + g_{\mu\kappa} \Gamma_{\lambda\nu}^\kappa \\ &= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} , \end{aligned} \quad (2.59)$$

which is a sum of two connection coefficients. Here $\Gamma_{\lambda\mu\nu}$ with all indices lowered is defined to be $\Gamma_{\mu\nu}^\kappa$ with the first index lowered by the metric,

$$\Gamma_{\lambda\mu\nu} \equiv g_{\lambda\kappa} \Gamma_{\mu\nu}^\kappa . \quad (2.60)$$

Combining the metric derivatives in the following fashion yields an expression for a single connection,

$$\begin{aligned} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} &= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} + \Gamma_{\lambda\nu\mu} + \Gamma_{\nu\lambda\mu} - \Gamma_{\mu\nu\lambda} - \Gamma_{\nu\mu\lambda} \\ &= 2\Gamma_{\lambda\mu\nu} - S_{\lambda\mu\nu} - S_{\mu\nu\lambda} - S_{\nu\mu\lambda} , \end{aligned} \quad (2.61)$$

with $S_{\lambda\mu\nu} \equiv g_{\lambda\kappa} S_{\mu\nu}^\kappa$, which shows that, in the presence of torsion,

$$\Gamma_{\lambda\mu\nu} = \frac{1}{2} \left(\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} + S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda} \right) \quad \text{not a coordinate tensor} . \quad (2.62)$$

If torsion vanishes, as general relativity assumes, then

$$\boxed{\Gamma_{\lambda\mu\nu} = \frac{1}{2} \left(\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} \right)} \quad \text{not a coordinate tensor} . \quad (2.63)$$

This is the formula that allows connection coefficients to be calculated from the metric.

2.12 Torsion-free covariant derivative

Einstein's principle of equivalence postulates that a locally inertial frame exists at each point of spacetime, and this implies that torsion vanishes. However, torsion is of special interest as a generalization of general

relativity because, as discussed in §2.19.2, the torsion tensor and the Riemann curvature tensor can be regarded as fields associated with local gauge groups of respectively displacements and Lorentz transformations. Together displacements and Lorentz transformations form the Poincaré group of symmetries of spacetime. Spinor (spin- $\frac{1}{2}$) fields inevitably generate torsion, Exercise 16.7, but torsion is local and non-propagating, and cancels between oppositely aligned spins, so in practice is negligible in almost all circumstances, §16.11.

The torsion-free part of the covariant derivative is a covariant derivative even when torsion is present (that is, it yields a tensor when acting on a tensor). The torsion-free covariant derivative is important, even when torsion is present, for several reasons. Firstly, as will be discovered from an action principle in Chapter 4, the covariant derivative that goes in the geodesic equation (2.88) is the torsion-free covariant derivative, equation (2.90). Secondly, the torsion-free covariant curl defines the exterior derivative in the theory of differential forms, §15.6. The exterior derivative has the property that it is inverse to integration over curved hypersurfaces. Integration is central to various aspects of general relativity, such as the development of Lagrangian and Hamiltonian mechanics. Thirdly, the Lie derivative, §7.34, is a covariant derivative defined in terms of torsion-free covariant derivatives. Finally, Yang-Mills gauge symmetries, such as the U(1) gauge symmetry of electromagnetism, require the gauge field to be defined in terms of the torsion-free covariant derivative, in order to preserve the gauge symmetry.

When torsion is present and it is desirable to make the torsion part explicit, it is convenient to distinguish torsion-free quantities with a $\circ$ overscript. The torsion-free part $\mathring{\Gamma}_{\lambda\mu\nu}$ of the connection, also called the **Levi-Civita connection**, is given by the right hand side of equation (2.63). When expressed in a coordinate frame (as opposed to a tetrad frame, §11.15), the components of the torsion-free connections $\mathring{\Gamma}_{\lambda\mu\nu}$ are also called **Christoffel symbols**. Sometimes, the components $\mathring{\Gamma}_{\lambda\mu\nu}$ with all indices lowered are called Christoffel symbols of the first kind, while components $\mathring{\Gamma}_{\mu\nu}^{\lambda}$ with first index raised are called Christoffel symbols of the second kind. There is no need to remember the jargon, but it is useful to know what it means if you meet it.

The torsion-full connection $\Gamma_{\lambda\mu\nu}$ is a sum of the torsion-free connection $\mathring{\Gamma}_{\lambda\mu\nu}$ and a tensor called the **contortion tensor** (not contorsion!) $K_{\lambda\mu\nu}$,

$$\Gamma_{\lambda\mu\nu} = \mathring{\Gamma}_{\lambda\mu\nu} + K_{\lambda\mu\nu} \quad \text{not a coordinate tensor} . \quad (2.64)$$

From equation (2.62), the contortion tensor $K_{\lambda\mu\nu}$ is related to the torsion tensor $S_{\lambda\mu\nu}$ by

$$K_{\lambda\mu\nu} = \frac{1}{2} (S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda}) = -S_{\nu\lambda\mu} + \frac{3}{2} S_{[\lambda\mu\nu]} \quad \text{a coordinate tensor} . \quad (2.65)$$

The contortion $K_{\lambda\mu\nu}$ is antisymmetric in its first two indices,

$$K_{\lambda\mu\nu} = -K_{\mu\lambda\nu} , \quad (2.66)$$

and thus like the torsion tensor $S_{\lambda\mu\nu}$ has $6 \times 4 = 24$ degrees of freedom. The torsion tensor $S_{\lambda\mu\nu}$ can be expressed in terms of the contortion tensor $K_{\lambda\mu\nu}$,

$$S_{\lambda\mu\nu} = K_{\lambda\mu\nu} - K_{\lambda\nu\mu} = -K_{\mu\nu\lambda} + 3K_{[\lambda\mu\nu]} \quad \text{a coordinate tensor} . \quad (2.67)$$

The torsion-full covariant derivative D_{ν} differs from the torsion-free covariant derivative $\mathring{D}_{\nu}$ by the contortion,

$$D_{\nu} A^{\kappa} \equiv \mathring{D}_{\nu} A^{\kappa} + K_{\mu\nu}^{\kappa} A^{\mu} \quad \text{a coordinate tensor} . \quad (2.68)$$

In this book torsion will not be assumed automatically to vanish, and thus by default the symbol D_ν will denote the torsion-full covariant derivative. When torsion is assumed to vanish, or when D_ν denotes the torsion-free covariant derivative, it will be explicitly stated so.

Concept question 2.5 Does torsion vanish in reality? General relativity postulates that torsion vanishes. Does it vanish in the actual Universe? **Answer.** This is a deep question, which ultimately can only be answered by observation or experiment. As remarked in §2.12 above, general relativity is premised on two local symmetries, Lorentz transformations and translations. The Riemann tensor is the field associated with Lorentz transformations, while the torsion tensor is the field associated with translations. My own opinion is that torsion probably does not vanish. But as discussed in §16.11, torsion is non-propagating, so necessarily vanishes in empty space, and it is sourced only by fermionic (spinor) fields, not by bosonic (gauge) fields, and it cancels between fermions of opposite spin, so as a practical matter torsion is expected to be undetectably tiny under almost all circumstances.

Concept question 2.6 Can the metric be Minkowski in the presence of torsion? In §2.10.1 it was argued that the postulate of the existence of locally inertial frames implies that torsion vanishes. The basis of the argument was the proposition that derivatives of the tangent axes vanish, equation (2.54b). Impose instead the weaker condition that the derivatives of the metric (i.e. scalar products of tangent axes) vanish,

$$\frac{\partial g_{\lambda\mu}}{\partial x^\nu} = 0 . \quad (2.69)$$

Can torsion be non-vanishing under this weaker condition? **Answer.** Yes. In fact torsion may exist even in flat (Minkowski) space, where the metric is everywhere Minkowski, $g_{\lambda\mu} = \eta_{\lambda\mu}$. The torsion-free connection is defined by equation (2.63), so vanishes if metric derivatives all vanish. On the other hand the condition (2.69) of vanishing metric derivatives constrains only the torsion-free connections,

$$\frac{1}{2} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} = \Gamma_{(\lambda\mu)\nu} - \mathring{\Gamma}_{(\lambda\mu)\nu} + K_{(\lambda\mu)\nu} - \mathring{\Gamma}_{(\lambda\mu)\nu} = 0 . \quad (2.70)$$

Thus the condition (2.69) of vanishing metric derivatives imposes no condition on torsion.

Exercise 2.7 Covariant curl and coordinate curl. Show that the covariant curl of a covariant vector A_λ is

$$D_\kappa A_\lambda - D_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} + S_{\kappa\lambda}^\mu A_\mu . \quad (2.71)$$

Conclude that the coordinate curl of a vector equals its torsion-free covariant curl,

$$\mathring{D}_\kappa A_\lambda - \mathring{D}_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} . \quad (2.72)$$

Of course, if torsion vanishes as general relativity assumes, then the covariant curl is the torsion-free covariant curl. Note that since both $D_\kappa A_\lambda - D_\lambda A_\kappa$ on the left hand side and $S_{\kappa\lambda}^\mu A_\mu$ on the right hand side of equation (2.71) are both tensors, it follows that the coordinate curl $\partial A_\lambda / \partial x^\kappa - \partial A_\kappa / \partial x^\lambda$ is a tensor even in the presence of torsion.

Exercise 2.8 Covariant divergence and coordinate divergence. Show that the covariant divergence of a contravariant vector A^μ is

$$D_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu} + S^\nu_{\mu\nu} A^\mu, \quad (2.73)$$

where $g \equiv |g_{\mu\nu}|$ is the determinant of the metric matrix. Conclude that the torsion-free covariant divergence is

$$\mathring{D}_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu}. \quad (2.74)$$

Of course, if torsion vanishes as general relativity assumes, then the covariant divergence is the torsion-free covariant divergence. Note that since both the covariant divergence on the left hand side of equation (2.73) and the torsion term on the right hand side of equation (2.73) are both tensors, the torsion-free covariant divergence (2.74) is a tensor even in the presence of torsion.

Solution. The covariant divergence is

$$D_\mu A^\mu = \frac{\partial A^\mu}{\partial x^\mu} + \Gamma^\nu_{\mu\nu} A^\mu. \quad (2.75)$$

From equation (2.62),

$$\begin{aligned} \Gamma^\nu_{\mu\nu} &= \frac{1}{2} g^{\nu\lambda} \frac{\partial g_{\lambda\nu}}{\partial x^\mu} + S^\nu_{\mu\nu} \\ &= \frac{\partial \ln |\sqrt{-g}|}{\partial x^\mu} + S^\nu_{\mu\nu}. \end{aligned} \quad (2.76)$$

The second line of equations (2.76) follows because for any matrix M , the variation of the logarithm of its determinant is

$$\begin{aligned} \delta \ln |M| &= \ln |M + \delta M| - \ln |M| \\ &= \ln |M^{-1}(M + \delta M)| \\ &= \ln |1 + M^{-1} \delta M| \\ &= \ln(1 + \text{Tr } M^{-1} \delta M) \\ &= \text{Tr } M^{-1} \delta M. \end{aligned} \quad (2.77)$$

The torsion-free covariant divergence is

$$\mathring{D}_\mu A^\mu = \frac{\partial A^\mu}{\partial x^\mu} + \mathring{\Gamma}^\nu_{\mu\nu} A^\mu, \quad (2.78)$$

where the torsion-free coordinate connection is

$$\mathring{\Gamma}^\nu_{\mu\nu} = \frac{1}{2} g^{\nu\lambda} \frac{\partial g_{\lambda\nu}}{\partial x^\mu} = \frac{\partial \ln |\sqrt{-g}|}{\partial x^\mu}. \quad (2.79)$$

Concept question 2.9 If torsion does not vanish, does torsion-free covariant differentiation

commute with the metric? Answer. Yes. Unlike the torsion-full covariant derivative, Concept question 2.3, the torsion-free covariant derivative of the tangent basis vectors e_κ does *not* vanish, but rather depends on the contortion $K_{\kappa\mu}^\nu e_\nu$,

$$\mathring{D}_\mu e_\kappa = D_\mu e_\kappa + K_{\kappa\mu}^\nu e_\nu = K_{\kappa\mu}^\nu e_\nu . \quad (2.80)$$

However, the torsion-free covariant derivative of the metric, that is, of scalar products of the tangent basis vectors, does vanish,

$$\mathring{D}_\mu g_{\kappa\lambda} = \mathring{D}_\mu (e_\kappa \cdot e_\lambda) = K_{\kappa\mu}^\nu e_\nu \cdot e_\lambda + K_{\lambda\mu}^\nu e_\kappa \cdot e_\nu = K_{\lambda\kappa\mu} + K_{\kappa\lambda\mu} = 0 , \quad (2.81)$$

thanks to the antisymmetry of the contortion tensor in its first two indices. As a corollary, torsion-free covariant differentiation commutes with the operations of raising and lowering indices, and of contraction.

2.13 Mathematical aside: What if there is no metric?

General relativity is a metric theory. Many of the structures introduced above can be defined mathematically without a metric. For example, it is possible to define the tangent space of vectors with basis e_μ , and to define a dual vector space with basis e^μ such that $e^\mu \cdot e_\nu = \delta_\nu^\mu$, equation (2.36). Elements of the dual vector space are called covectors. Similarly it is possible to define connections and covariant derivatives without a metric. However, this book follows general relativity in assuming that spacetime has a metric.

2.14 Coordinate 4-velocity

Consider a particle following a worldline

$$x^\mu(\tau) , \quad (2.82)$$

where τ is the particle's proper time. The proper time along any interval of the worldline is $d\tau \equiv \sqrt{-ds^2}$.

Define the **coordinate 4-velocity** u^μ by

$$\boxed{u^\mu \equiv \frac{dx^\mu}{d\tau}} \quad \text{a coordinate 4-vector} . \quad (2.83)$$

The magnitude squared of the 4-velocity is constant

$$u_\mu u^\mu = g_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = \frac{ds^2}{d\tau^2} = -1 . \quad (2.84)$$

The negative sign arises from the choice of metric signature: with the signature $-+++$ adopted here, there is a $-$ sign between ds^2 and $d\tau^2$. Equation (2.84) can be regarded as an integral of motion associated with conservation of particle rest mass.

2.15 Geodesic equation

Let $\mathbf{u} \equiv \mathbf{e}_\mu u^\mu$ be the 4-velocity in coordinate-independent notation. The principle of equivalence (which imposes vanishing torsion) implies that the **geodesic equation**, the equation of motion of a freely-falling particle, is

$$\boxed{\frac{d\mathbf{u}}{d\tau} = 0} . \quad (2.85)$$

Why? Because $d\mathbf{u}/d\tau = 0$ in the particle's own free-fall frame, and the equation is coordinate-independent. In the particle's own free-fall frame, the particle's 4-velocity is $u^\mu = \{1, 0, 0, 0\}$, and the particle's locally inertial axes $\mathbf{e}_\mu = \{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are constant.

What does the equation of motion look like in coordinate notation? The acceleration is

$$\begin{aligned} \frac{d\mathbf{u}}{d\tau} &= \frac{dx^\nu}{d\tau} \frac{\partial \mathbf{u}}{\partial x^\nu} \\ &= u^\nu \mathbf{e}_\kappa D_\nu u^\kappa \\ &= u^\nu \mathbf{e}_\kappa \left(\frac{\partial u^\kappa}{\partial x^\nu} + \Gamma_{\mu\nu}^\kappa u^\mu \right) \\ &= \mathbf{e}_\kappa \left(\frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu \right) . \end{aligned} \quad (2.86)$$

The geodesic equation is then

$$\boxed{\frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu = 0} . \quad (2.87)$$

Another way of writing the geodesic equation is

$$\frac{Du^\kappa}{D\tau} = 0 , \quad (2.88)$$

where $D/D\tau$ is the covariant proper time derivative

$$\frac{D}{D\tau} \equiv u^\nu D_\nu . \quad (2.89)$$

The above derivation of the geodesic equation invoked the principle of equivalence, which postulates that locally inertial frames exist, and thus that torsion vanishes. What happens if torsion does not vanish? In Chapter 4, equation (4.15), it will be shown from an action principle that in the presence of torsion, the covariant derivative in the geodesic equation should simply be replaced by the torsion-free covariant derivative $\mathring{D}/D\tau = u^\mu \mathring{D}_\mu$,

$$\boxed{\frac{\mathring{D}u^\kappa}{D\tau} = 0} . \quad (2.90)$$

Thus the geodesic motion of particles is unaffected by the presence of torsion.

2.16 Coordinate 4-momentum

The coordinate 4-momentum of a particle of rest mass m is defined to be

$$\boxed{p^\mu \equiv m u^\mu = m \frac{dx^\mu}{d\tau}} \quad \text{a coordinate 4-vector .} \quad (2.91)$$

The momentum squared is, from equation (2.84),

$$p_\mu p^\mu = m^2 u_\mu u^\mu = -m^2 \quad (2.92)$$

minus the square of the rest mass. Again, the minus sign arises from the choice $-+++$ of metric signature.

2.17 Affine parameter

For photons, the rest mass is zero, $m = 0$, but the 4-momentum p^μ remains finite. Define the **affine parameter** λ by

$$\boxed{\lambda \equiv \frac{\tau}{m}} \quad \text{a coordinate scalar} \quad (2.93)$$

which remains finite in the limit $m \rightarrow 0$. The affine parameter λ is unique up to an overall linear transformation (that is, $\alpha\lambda + \beta$ is also an affine parameter, for constant α and β), because of the freedom in the choice of mass m and the zero point of proper time τ . In terms of the affine parameter, the 4-momentum is

$$\boxed{p^\mu = \frac{dx^\mu}{d\lambda}} . \quad (2.94)$$

The geodesic equation is then in coordinate-independent notation

$$\frac{d\mathbf{p}}{d\lambda} = 0 , \quad (2.95)$$

or in component form

$$\frac{dp^\kappa}{d\lambda} + \Gamma_{\mu\nu}^\kappa p^\mu p^\nu = 0 , \quad (2.96)$$

which works for massless as well as massive particles.

Another way of writing this is

$$\frac{Dp^\kappa}{D\lambda} = 0 , \quad (2.97)$$

where $D/D\lambda$ is the covariant affine derivative

$$\frac{D}{D\lambda} \equiv p^\nu D_\nu . \quad (2.98)$$

In the presence of torsion, the connection in the geodesic equation (2.96) should be interpreted as the

torsion-free connection $\overset{\circ}{\Gamma}_{\mu\nu}^{\kappa}$, and the covariant derivative in equations (2.97) and (2.98) are torsion-free covariant derivatives.

2.18 Affine distance

The freedom in the overall scaling of the affine parameter can be removed by setting it equal to the proper distance near the observer in the observer's locally inertial rest frame. With the scaling fixed in this fashion, the affine parameter is called the **affine distance**, so called because it provides a measure of distance along null geodesics. When you look at a scene with your eyes, you are looking along null geodesics, and the natural measure of distance to objects that you see is the affine distance (Hamilton and Polhemus, 2010).

In special relativity, the affine distance coincides with the perceived (e.g. binocular) distance to objects.

Exercise 2.10 Gravitational redshift in a stationary metric. Let $x^\mu \equiv \{t, x^\alpha\}$ constitute time t and spatial coordinates x^α of a spacetime. The metric $g_{\mu\nu}$ is said to be stationary if it is independent of the coordinate t . A comoving observer in the spacetime is one that is at rest in the spatial coordinates, $dx^\alpha/d\tau = 0$.

1. Argue that the coordinate 4-velocity $u^\nu \equiv dx^\nu/d\tau$ of a comoving observer in a stationary spacetime is

$$u^\nu = \{\gamma, 0, 0, 0\}, \quad \gamma \equiv \frac{1}{\sqrt{-g_{tt}}}. \quad (2.99)$$

2. Argue that the proper energy E of a particle, massless or massive, with energy-momentum 4-vector p^ν seen by a comoving observer with 4-velocity u^ν , equation (2.99), is

$$E = -u^\nu p_\nu. \quad (2.100)$$

3. Consider a particle, massless or massive, that follows a geodesic between two comoving observers. Since the metric is independent of the time coordinate t , the covariant momentum p_t is a constant of motion, equation (4.50). Argue that the ratio $E_{\text{obs}}/E_{\text{em}}$ of the observed to emitted energies between two comoving observers is

$$\frac{E_{\text{obs}}}{E_{\text{em}}} = \frac{\gamma_{\text{obs}}}{\gamma_{\text{em}}}. \quad (2.101)$$

4. Can comoving observers exist where g_{tt} is positive?

Exercise 2.11 Gravitational redshift in Rindler space. Rindler space is Minkowski space expressed in the coordinates of uniformly accelerating observers, called Rindler observers. Rindler observers are precisely the observers in the right quadrant of the spacetime wheel, Figure 1.14.

1. Start with Minkowski space in a Cartesian coordinate system $\{t, x, y, z\}$. Define Rindler coordinates α, l by

$$t = l \sinh \alpha, \quad x = l \cosh \alpha. \quad (2.102)$$

Show that the line-element in Rindler coordinates is

$$ds^2 = -l^2 d\alpha^2 + dl^2 + dy^2 + dz^2 . \quad (2.103)$$

2. A Rindler observer is a comoving observer in Rindler space, one who follows a worldline of constant l , y , and z . Since Rindler spacetime is stationary, conclude that the ratio $E_{\text{obs}}/E_{\text{em}}$ of the observed to emitted energies between two Rindler observers is, equation (2.101),

$$\frac{E_{\text{obs}}}{E_{\text{em}}} = \frac{l_{\text{em}}}{l_{\text{obs}}} . \quad (2.104)$$

3. Do Rindler observers all accelerate at the same rate? [Hint: No. An observer accelerating along a worldline of constant l accelerates at rate $1/l$.]
 4. Can Rindler space be considered equivalent to a spacetime containing a uniform gravitational field?

Exercise 2.12 Gravitational redshift in a uniformly rotating space. Start with Minkowski space in cylindrical coordinates $\{t, r, \phi, z\}$,

$$ds^2 = -dt^2 + dr^2 + r^2 d\phi^2 + dz^2 . \quad (2.105)$$

Define a uniformly rotating azimuthal angle χ by

$$\chi \equiv \phi - \omega t , \quad (2.106)$$

which is constant for observers who are at rest in a system rotating uniformly at angular velocity ω . The line-element in uniformly rotating coordinates is

$$ds^2 = -dt^2 + dr^2 + r^2(d\chi + \omega dt)^2 + dz^2 . \quad (2.107)$$

1. A comoving observer in the uniformly rotating system follows a worldline at constant r , χ , and z . Since the uniformly rotating spacetime is stationary, conclude that the ratio $E_{\text{obs}}/E_{\text{em}}$ of the observed to emitted energies between two comoving observers is, equation (2.101),

$$\frac{E_{\text{obs}}}{E_{\text{em}}} = \frac{\gamma_{\text{em}}}{\gamma_{\text{obs}}} , \quad (2.108)$$

where

$$\gamma = \frac{1}{\sqrt{1 - v^2}} , \quad v = \omega r . \quad (2.109)$$

2. What happens where $v > 1$?

Concept question 2.13 Can Minkowski space rotate? Exercise 2.12 considered Minkowski space in rotating coordinates. Can Minkowski space rotate globally? **Answer.** No. General relativity allows arbitrary choices of coordinates, including choices that allow physical objects to move through the coordinates faster than light. However, the choice of coordinates does not affect physical observables in any way. The metric encodes locally inertial frames, determining what intervals are timelike, lightlike, or spacelike (ds^2 less than, equal to, or greater than zero). That locally inertial structure is independent of the choice of coordinates.

Objects cannot move through locally inertial frames than light. Thus Minkowski spacetime does not rotate globally, regardless of the choice of coordinates.

2.19 Riemann tensor

2.19.1 Riemann curvature tensor

The **Riemann curvature tensor** $R_{\kappa\lambda\mu\nu}$ is defined by the commutator of the covariant derivative acting on a 4-vector. In the presence of torsion,

$$[D_\kappa, D_\lambda] A_\mu = S_{\kappa\lambda}^\nu D_\nu A_\mu + R_{\kappa\lambda\mu\nu} A^\nu \quad \text{a coordinate tensor .} \quad (2.110)$$

If torsion vanishes, as general relativity assumes, then the definition (2.110) reduces to

$$\boxed{[D_\kappa, D_\lambda] A_\mu = R_{\kappa\lambda\mu\nu} A^\nu} \quad \text{a coordinate tensor .} \quad (2.111)$$

The expression (2.51) for the covariant derivative yields the following formula for the Riemann tensor in terms of connection coefficients

$$\boxed{R_{\kappa\lambda\mu\nu} = \frac{\partial \Gamma_{\mu\nu\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{\mu\nu\kappa}}{\partial x^\lambda} + \Gamma_{\mu\lambda}^\pi \Gamma_{\pi\nu\kappa} - \Gamma_{\mu\kappa}^\pi \Gamma_{\pi\nu\lambda}} \quad \text{a coordinate tensor .} \quad (2.112)$$

This is the formula that allows the Riemann tensor to be calculated from the connection coefficients. The same formula (2.112) remains valid if torsion does not vanish, but the connection coefficients $\Gamma_{\lambda\mu\nu}$ themselves are given by (2.62) in place of (2.63).

In flat (Minkowski) space, covariant derivatives reduce to partial derivatives, $D_\kappa \rightarrow \partial/\partial x^\kappa$, and

$$[D_\kappa, D_\lambda] \rightarrow \left[\frac{\partial}{\partial x^\kappa}, \frac{\partial}{\partial x^\lambda} \right] = 0 \quad \text{in flat space} \quad (2.113)$$

so that $R_{\kappa\lambda\mu\nu} = 0$ in flat space.

Exercise 2.14 Derivation of the Riemann tensor. Confirm expression (2.112) for the Riemann tensor. This is an exercise that any serious student of general relativity should do. However, you might like to defer this rite of passage to Chapter 11, where Exercises 11.4–11.7 take you through the derivation and properties of the tetrad-frame Riemann tensor.

2.19.2 Commutator of the covariant derivative acting on a general tensor

The commutator of the covariant derivative is of fundamental importance because it defines what is meant by the field in gauge theories.

It has been seen above that the commutator of the covariant derivative acting on a scalar defined the torsion tensor, equation (2.57), which general relativity assumes vanishes, while the commutator of the covariant

derivative acting on a vector defined the Riemann tensor, equation (2.111). Does the commutator of the covariant derivative acting on a general tensor introduce any other distinct tensor? No: the torsion and Riemann tensors completely define the action of the commutator of the covariant derivative on any tensor. Acting on a general tensor, the commutator of the covariant derivative is

$$[D_\kappa, D_\lambda] A_{\mu\nu\dots}^{\pi\rho\dots} = S_{\kappa\lambda}^\sigma D_\sigma A_{\mu\nu\dots}^{\pi\rho\dots} + R_{\kappa\lambda\mu}{}^\sigma A_{\sigma\nu\dots}^{\pi\rho\dots} + R_{\kappa\lambda\nu}{}^\sigma A_{\mu\sigma\dots}^{\pi\rho\dots} - R_{\kappa\lambda\sigma}{}^\pi A_{\mu\nu\dots}^{\sigma\rho\dots} - R_{\kappa\lambda\sigma}{}^\rho A_{\mu\nu\dots}^{\pi\sigma\dots} . \quad (2.114)$$

In more abstract notation, the commutator of the covariant derivative is the operator

$$[D_\kappa, D_\lambda] = S_{\kappa\lambda}^\mu D_\mu + \hat{R}_{\kappa\lambda} \quad (2.115)$$

where the Riemann curvature operator $\hat{R}_{\kappa\lambda}$ is an operator whose action on any tensor is specified by equation (2.114). The action of the operator $\hat{R}_{\kappa\lambda}$ is analogous to that of the covariant derivative (2.52): there's a positive R term for each covariant index, and a negative R term for each contravariant index. The action of $\hat{R}_{\kappa\lambda}$ on a scalar is zero, which reflects the fact that a scalar is unchanged by a Lorentz transformation.

The general expression (2.114) for the commutator of the covariant derivative reveals the meaning of the torsion and Riemann tensors. The torsion and Riemann tensors describe respectively the displacement and the Lorentz transformation experienced by an object when parallel-transported around a curve. Displacements and Lorentz transformations together constitute the Poincaré group, the complete group of symmetries of flat spacetime.

How can an object detect a displacement when parallel-transported around a curve? If you go around a curve back to the same coordinate in spacetime where you began, won't you necessarily be at the same position? This is a question that goes to heart of the meaning of spacetime. To answer the question, you have to consider how fundamental particles are able to detect position, orientation, and velocity. Classically, particles may be structureless points, but quantum mechanically, particles possess frequency, wavelength, spin, and (in the relativistic theory) boost, and presumably it is these properties that allow particles to “measure” the properties of the spacetime in which they live. For example, a Dirac spinor (relativistic spin- $\frac{1}{2}$ particle) Lorentz transforms under the fundamental (spin- $\frac{1}{2}$) representation of the Lorentz group, and is thus endowed with precisely the properties that allow it to “measure” boost and rotation, §14.10. The Dirac wave equation shows that a Dirac spinor propagating through spacetime varies as $\sim e^{ip_\mu x^\mu}$, whose phase encodes the displacement of the Dirac spinor. Thus a Dirac spinor could potentially detect a displacement through a change in its phase when parallel-transported around a curve back to the same point in spacetime. Since a change in phase is indistinguishable from a spatial rotation about the spin axis of the Dirac spinor, operationally torsion rotates particles, whence the name torsion.

2.19.3 No torsion

In the remainder of this Chapter, torsion will be assumed to vanish, as general relativity postulates. A decomposition of the Riemann tensor into torsion-free and contortion parts is deferred to §11.18.

2.19.4 Symmetries of the Riemann tensor

In a locally inertial frame (necessarily, with vanishing torsion), the connection coefficients all vanish, $\Gamma_{\lambda\mu\nu} = 0$, but their partial derivatives, which are proportional to second derivatives of the metric tensor, equation (2.63), do not vanish. Thus in a locally inertial frame the Riemann tensor is

$$\begin{aligned} R_{\kappa\lambda\mu\nu} &= \frac{\partial \Gamma_{\mu\nu\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{\mu\nu\kappa}}{\partial x^\lambda} \\ &= \frac{1}{2} \left(\frac{\partial^2 g_{\mu\nu}}{\partial x^\kappa \partial x^\lambda} + \frac{\partial^2 g_{\mu\lambda}}{\partial x^\kappa \partial x^\nu} - \frac{\partial^2 g_{\nu\lambda}}{\partial x^\kappa \partial x^\mu} - \frac{\partial^2 g_{\mu\nu}}{\partial x^\lambda \partial x^\kappa} - \frac{\partial^2 g_{\mu\kappa}}{\partial x^\lambda \partial x^\nu} + \frac{\partial^2 g_{\nu\kappa}}{\partial x^\lambda \partial x^\mu} \right) \\ &= \frac{1}{2} \left(\frac{\partial^2 g_{\mu\lambda}}{\partial x^\kappa \partial x^\nu} - \frac{\partial^2 g_{\nu\lambda}}{\partial x^\kappa \partial x^\mu} - \frac{\partial^2 g_{\mu\kappa}}{\partial x^\lambda \partial x^\nu} + \frac{\partial^2 g_{\nu\kappa}}{\partial x^\lambda \partial x^\mu} \right) . \end{aligned} \quad (2.116)$$

You can check that the bottom line of equation (2.116):

1. is antisymmetric in $\kappa \leftrightarrow \lambda$,
2. is antisymmetric in $\mu \leftrightarrow \nu$,
3. is symmetric in $\kappa\lambda \leftrightarrow \mu\nu$,
4. has the property that the sum of the cyclic permutations of the last three (or first three, or indeed any three) indices vanishes

$$R_{\kappa\lambda\mu\nu} + R_{\kappa\nu\lambda\mu} + R_{\kappa\mu\nu\lambda} = 0 . \quad (2.117)$$

Actually, as shown in Exercise 11.7, the third, symmetric, symmetry is a consequence of the fourth, cyclic symmetry. The first three of the four symmetries can be expressed compactly

$$R_{\kappa\lambda\mu\nu} = R_{([\kappa\lambda][\mu\nu])} , \quad (2.118)$$

in which $[]$ denotes antisymmetrization and $()$ symmetrization, as in

$$A_{[\kappa\lambda]} \equiv \frac{1}{2} (A_{\kappa\lambda} - A_{\lambda\kappa}) , \quad A_{(\kappa\lambda)} \equiv \frac{1}{2} (A_{\kappa\lambda} + A_{\lambda\kappa}) . \quad (2.119)$$

The symmetries (2.118) imply that the Riemann tensor is a symmetric matrix of antisymmetric matrices. An antisymmetric tensor is also known as a **bivector**, much more about which you can discover in Chapter 13 on the geometric algebra. An antisymmetric matrix, or bivector, in 4 dimensions has 6 degrees of freedom. A symmetric matrix of bivectors is a 6×6 symmetric matrix, which has 21 degrees of freedom. The final, cyclic symmetry of the Riemann tensor, equation (2.117), which can be abbreviated

$$R_{\kappa[\lambda\mu\nu]} = 0 , \quad (2.120)$$

removes 1 further degree of freedom. Thus the Riemann tensor has a net 20 degrees of freedom.

Although the above symmetries were derived in a locally inertial frame, the fact that the Riemann tensor is a tensor means that the symmetries hold in any frame. If you prefer, you can add back the products of connection coefficients in equation (2.112), and check that the claimed symmetries remain.

Some of the symmetries of the Riemann tensor persist when torsion is present, and others do not. The relation between symmetries of the Riemann tensor and torsion is deferred to Exercises 11.5–11.7.

2.20 Ricci tensor, Ricci scalar

The Ricci tensor $R_{\kappa\mu}$ and Ricci scalar R are the essentially unique contractions of the Riemann curvature tensor. The **Ricci tensor**, the compressive part of the Riemann tensor, is

$$\boxed{R_{\kappa\mu} \equiv g^{\lambda\nu} R_{\kappa\lambda\mu\nu}} \quad \text{a coordinate tensor .} \quad (2.121)$$

If torsion vanishes as general relativity assumes, then the Ricci tensor is symmetric

$$R_{\kappa\mu} = R_{\mu\kappa} \quad (2.122)$$

and therefore has 10 independent components.

The **Ricci scalar** is

$$\boxed{R \equiv g^{\kappa\mu} R_{\kappa\mu}} \quad \text{a coordinate tensor (a scalar) .} \quad (2.123)$$

2.21 Einstein tensor

The **Einstein tensor** $G_{\kappa\mu}$ is defined by

$$\boxed{G_{\kappa\mu} \equiv R_{\kappa\mu} - \frac{1}{2} g_{\kappa\mu} R} \quad \text{a coordinate tensor .} \quad (2.124)$$

For vanishing torsion, the symmetry of the Ricci and metric tensors imply that the Einstein tensor is likewise symmetric

$$G_{\kappa\mu} = G_{\mu\kappa} , \quad (2.125)$$

and thus has 10 independent components.

2.22 Bianchi identities

The Jacobi identity

$$[D_{\kappa}, [D_{\lambda}, D_{\mu}]] + [D_{\lambda}, [D_{\mu}, D_{\kappa}]] + [D_{\mu}, [D_{\kappa}, D_{\lambda}]] = 0 \quad (2.126)$$

implies the **Bianchi identities** which, for vanishing torsion, are

$$D_{\kappa} R_{\lambda\mu\nu\pi} + D_{\lambda} R_{\mu\kappa\nu\pi} + D_{\mu} R_{\kappa\lambda\nu\pi} = 0 . \quad (2.127)$$

The torsion-free Bianchi identities can be written in shorthand

$$\boxed{D_{[\kappa} R_{\lambda\mu]\nu\pi} = 0} . \quad (2.128)$$

The Bianchi identities constitute a set of differential relations between the components of the Riemann tensor, which are distinct from the algebraic symmetries of the Riemann tensor. There are 4 ways to pick

$[\kappa\lambda\mu]$, and 6 ways to pick antisymmetric $\nu\pi$, giving $4 \times 6 = 24$ Bianchi identities, but 4 of the identities, $D_{[\kappa}R_{\lambda\mu\nu]\pi} = 0$, are implied by the cyclic symmetry (2.120), which is a consequence of vanishing torsion. Thus there are $24 - 4 = 20$ non-trivial torsion-free Bianchi identities on the 20 components of the torsion-free Riemann tensor.

Exercise 2.15 Jacobi identity. Prove the Jacobi identity (2.126).

2.23 Covariant conservation of the Einstein tensor

The most important consequence of the torsion-free Bianchi identities (2.127) is obtained from the double contraction

$$g^{\kappa\nu}g^{\lambda\pi}(D_{\kappa}R_{\lambda\mu\nu\pi} + D_{\lambda}R_{\mu\kappa\nu\pi} + D_{\mu}R_{\kappa\lambda\nu\pi}) = -D^{\kappa}R_{\kappa\mu} - D^{\lambda}R_{\lambda\mu} + D_{\mu}R = 0, \quad (2.129)$$

or equivalently

$$D^{\kappa}G_{\kappa\mu} = 0, \quad (2.130)$$

where $G_{\kappa\mu}$ is the Einstein tensor, equation (2.124). Equation (2.130) is a primary motivation for the form of the Einstein equations, since it implies energy-momentum conservation, equation (2.132). It is worth remarking that the derivation of the contracted Bianchi identities (2.129) holds in arbitrarily many spacetime dimensions, so the factor of $\frac{1}{2}$ multiplying the Ricci scalar R in the definition (2.124) of the Einstein tensor holds in arbitrarily many spacetime dimensions, not just 4.

2.24 Einstein equations

Einstein's equations are

$$\boxed{G_{\kappa\mu} = 8\pi GT_{\kappa\mu}} \quad \text{a coordinate tensor equation.} \quad (2.131)$$

What motivates the form of Einstein's equations?

1. The equation is generally covariant.
2. For vanishing torsion, the contracted Bianchi identities (2.129) guarantee covariant conservation of the Einstein tensor, equation (2.130), which in turn guarantees covariant conservation of energy-momentum,

$$\boxed{D^{\kappa}T_{\kappa\mu} = 0}. \quad (2.132)$$

3. The Einstein tensor depends on the lowest (second) order derivatives of the metric tensor that do not vanish in a locally inertial frame.

In Chapter 16, the Einstein equations will be derived from an action principle. Although Einstein derived his equations from considerations of theoretical elegance, the real justification for them is that they reproduce observation.

Einstein's equations (2.131) constitute a complete set of gravitational equations, generalizing Poisson's equation of Newtonian gravity. However, Einstein's equations by themselves do not constitute a closed set of equations: in general, other equations, such as Maxwell's equations of electromagnetism, and equations describing the microphysics of the energy-momentum, must be adjoined to form a closed set.

Exercise 2.16 Einstein tensor in 3 or more dimensions. What is the Einstein tensor in $N \geq 3$ spacetime dimensions?

Solution. The Einstein tensor must be covariantly conserved to ensure that its source, energy-momentum, is covariantly conserved. The doubly-contracted Bianchi identities (2.129) hold as long as there are at least 3 spacetime dimensions. In $N = 2$ spacetime dimensions, there are zero Bianchi identities (2.128), since there are zero ways of picking 3 distinct indices. Thus the expression (2.124) for the Einstein tensor holds in any number $N \geq 3$ of spacetime dimensions. See §11.19 for general relativity in 2 spacetime dimensions.

2.25 Summary of the path from metric to the energy-momentum tensor

1. Start by defining the metric $g_{\mu\nu}$.
2. Compute the connection coefficients $\Gamma_{\lambda\mu\nu}$ from equation (2.63).
3. Compute the Riemann tensor $R_{\kappa\lambda\mu\nu}$ from equation (2.112).
4. Compute the Ricci tensor $R_{\kappa\mu}$ from equation (2.121), the Ricci scalar R from equation (2.123), and the Einstein tensor $G_{\kappa\mu}$ from equation (2.124).
5. The Einstein equations (2.131) then imply the energy-momentum tensor $T_{\kappa\mu}$.

The path from metric to energy-momentum tensor is straightforward to program on a computer, but the results are typically messy and complicated, even for fairly simple spacetimes. Inverting the path to recover the metric from a given energy-momentum content is typically highly non-trivial, the subject of a vast literature.

The great majority of metrics $g_{\mu\nu}$ yield an energy-momentum tensor $T_{\kappa\mu}$ that cannot be achieved with normal matter.

2.26 Energy-momentum tensor of a perfect fluid

The simplest non-trivial energy-momentum tensor is that of a **perfect fluid**. In this case $T^{\mu\nu}$ is taken to be isotropic in the locally inertial rest frame of the fluid, taking the form

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} \quad (2.133)$$

where

$$\begin{aligned} \rho & \text{ is the proper mass-energy density ,} \\ p & \text{ is the proper pressure .} \end{aligned} \quad (2.134)$$

The expression (2.133) is valid only in the locally inertial rest frame of the fluid. An expression that is valid in any frame is

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu} , \quad (2.135)$$

where u^μ is the 4-velocity of the fluid. Equation (2.135) is valid because it is a tensor equation, and it is true in the locally inertial rest frame, where $u^\mu = \{1, 0, 0, 0\}$.

2.27 Cosmological constant

A possible modification of Einstein's equations (2.131), introduced by Einstein (1917), is to adjoin a **cosmological constant** term $\Lambda g_{\kappa\mu}$ to the Einstein tensor,

$$G_{\kappa\mu} + \Lambda g_{\kappa\mu} = 8\pi G T_{\kappa\mu} \quad \text{a coordinate tensor equation .} \quad (2.136)$$

The cosmological constant Λ is constant everywhere in spacetime. The modification is possible because the metric is (by definition of the covariant derivative) covariantly conserved, $D^\kappa g_{\kappa\mu} = 0$. The standard modern interpretation is that the cosmological constant term should be taken from the left to the right hand side of equation (2.136) and reinterpreted as a species of energy-momentum,

$$T_{\kappa\mu}^\Lambda = -\rho_\Lambda g_{\kappa\mu} , \quad \rho_\Lambda = \frac{\Lambda}{8\pi G} . \quad (2.137)$$

The energy-momentum is that of a perfect fluid, equation (2.133), with pressure equal to minus density, $p_\Lambda = -\rho_\Lambda$. Einstein introduced a cosmological constant to source a gravitation repulsion that would counterbalance the attraction of matter, allowing the Universe to be static, but that motivation disappeared with the discovery that the Universe is expanding, §10.1.1. The modern discovery that the Universe is accelerating, §10.1.2, lends credence to the idea that our Universe has a nonzero positive cosmological constant. Astrophysicists agnostically refer to the repulsive source as **dark energy**, since its origin in fundamental physics remains uncertain.

2.28 Newtonian limit

The Newtonian limit is obtained in the limit of a weak gravitational field and nonrelativistic (pressureless) matter. In Cartesian coordinates, the metric in the Newtonian limit is (see Chapter 27)

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)(dx^2 + dy^2 + dz^2) , \quad (2.138)$$

in which

$$\Phi(x, y, z) = \text{Newtonian potential} \quad (2.139)$$

is a function only of the spatial coordinates x, y, z , not of time t .

For this metric, to first order in the potential Φ the only non-vanishing component of the Einstein tensor is the time-time component

$$G_{tt} = 2\nabla^2\Phi , \quad (2.140)$$

where $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the usual 3-dimensional Laplacian operator. This component (2.140) of the Einstein tensor plugged into Einstein's equations (2.131) implies Poisson's equation (2.4).

Exercise 2.17 Special and general relativistic corrections for clocks on satellites. The metric just above the surface of the Earth is well-approximated by

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) , \quad (2.141)$$

where

$$\Phi(r) = -\frac{GM}{r} \quad (2.142)$$

is the familiar Newtonian gravitational potential.

1. **Proper time.** Consider an object at fixed radius r , moving along the equator $\theta = \pi/2$ with constant nonrelativistic velocity $r d\phi/dt = v$. Compare the proper time of this object with that at rest at infinity. [Hint: Work to first order in the potential Φ . Regard v^2 as first order in Φ . Why is that reasonable?]
2. **Orbits.** Consider a satellite in orbit about the Earth. The conservation of energy E per unit mass, angular momentum L per unit mass, and rest mass per unit mass are expressed by (§4.8)

$$u_t = -E , \quad u_\phi = L , \quad u_\mu u^\mu = -1 . \quad (2.143)$$

For equatorial orbits, $\theta = \pi/2$, show that the radial component u^r of the 4-velocity satisfies

$$u^r = \sqrt{2(\Delta E - U)} , \quad (2.144)$$

where ΔE is the energy per unit mass of the particle excluding its rest mass energy,

$$\Delta E = E - 1 , \quad (2.145)$$

and the effective potential U is

$$U = \Phi + \frac{L^2}{2r^2} . \quad (2.146)$$

[Hint: Neglect air resistance. Remember to work to first order in Φ . Treat ΔE and L^2 as first order in Φ . Why is that reasonable?]

3. **Circular orbits.** From the condition that the potential U be an extremum, find the circular orbital velocity $v = r d\phi/dt$ of a satellite at radius r .
4. **Special and general relativistic corrections for satellites.** Compare the proper time of a satellite in circular orbit to that of a person at rest at infinity. Express your answer in the form

$$\frac{d\tau_{\text{satellite}}}{dt} - 1 = -\Phi_{\oplus} (f_{\text{GR}} + f_{\text{SR}}) , \quad (2.147)$$

where f_{GR} and f_{SR} are the general relativistic and special relativistic corrections, and $\Phi_{\oplus}$ is the dimensionless gravitational potential at the surface of the Earth,

$$\Phi_{\oplus} = -\frac{GM_{\oplus}}{c^2 R_{\oplus}} . \quad (2.148)$$

What is the value of $\Phi_{\oplus}$ in milliseconds per year?

5. **Special and general relativistic corrections for satellites vs. Earth observer.** Compare the proper time of a satellite in circular orbit to that of a person on Earth at one of the poles (so the person has no motion from the Earth's rotation). Express your answer in the form

$$\frac{d\tau_{\text{satellite}}}{dt} - \frac{d\tau_{\text{person}}}{dt} = -\Phi_{\oplus} (f_{\text{GR}} + f_{\text{SR}}) . \quad (2.149)$$

At what satellite radius r , in units of Earth radius $R_{\oplus}$, do the special and general relativistic corrections cancel?

6. **Special and general relativistic corrections for ISS and GPS satellites.** What are the corrections (be careful to get the sign right!) in units of $\Phi_{\oplus}$, and in units of ms yr^{-1} , for (i) a satellite in low Earth orbit, such as the International Space Station; (ii) a nearly geostationary satellite, such as a GPS satellite? Google the numbers that you may need.

Exercise 2.18 Equations of motion in weak gravity. Take the metric to be the Newtonian metric (2.138) with the Newtonian potential $\Phi(x, y, z)$ a function only of the spatial coordinates x, y, z , not of time t , equation (2.139).

1. Confirm that the non-zero connection coefficients are (coefficients as below but with the last two indices swapped are the same by the no-torsion condition $\Gamma_{\mu\nu}^{\kappa} = \Gamma_{\nu\mu}^{\kappa}$)

$$\Gamma_{t\alpha}^t = \Gamma_{tt}^{\alpha} = \Gamma_{\beta\beta}^{\alpha} = -\Gamma_{\beta\alpha}^{\beta} = -\Gamma_{\alpha\alpha}^{\alpha} = \frac{\partial\Phi}{\partial x^{\alpha}} \quad (\alpha \neq \beta = x, y, z) . \quad (2.150)$$

[Hint: Work to linear order in Φ .]

2. Consider a massive, nonrelativistic particle moving with 4-velocity $u^{\mu} \equiv dx^{\mu}/d\tau = \{u^t, u^x, u^y, u^z\}$. Show that $u_{\mu} u^{\mu} = -1$ implies that

$$u^t = 1 + \frac{1}{2}u^2 - \Phi , \quad (2.151)$$

whereas

$$u_t = - \left(1 + \frac{1}{2} u^2 + \Phi \right) \quad (2.152)$$

where $u \equiv [(u^x)^2 + (u^y)^2 + (u^z)^2]^{1/2}$. One of u^t or u_t is constant. Which one? [Hint: Work to linear order in Φ . Note that u^2 is of linear order in Φ .]

3. **Equation of motion of a massive particle.** From the geodesic equation

$$\frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu = 0 \quad (2.153)$$

show that

$$\frac{du^\alpha}{dt} = - \frac{\partial \Phi}{\partial x^\alpha} \quad \alpha = x, y, z. \quad (2.154)$$

Why is it legitimate to replace $d\tau$ by dt ? Show further that

$$\frac{du^t}{dt} = - 2 u^\alpha \frac{\partial \Phi}{\partial x^\alpha} \quad (2.155)$$

with implicit summation over $\alpha = x, y, z$. Does the result agree with what you would expect from equation (2.151)?

4. For a massless particle, the proper time along a geodesic is zero, and the affine parameter λ must be used instead of the proper time. The 4-velocity of a massless particle can be defined to be (and really this is just the 4-momentum p^μ up to an arbitrary overall factor) $v^\mu \equiv dx^\mu/d\lambda = \{v^t, v^x, v^y, v^z\}$. Show that $v_\mu v^\mu = 0$ implies that

$$v^t = (1 - 2\Phi)v, \quad (2.156)$$

whereas

$$v_t = -v, \quad (2.157)$$

where $v \equiv [(v^x)^2 + (v^y)^2 + (v^z)^2]^{1/2}$. One of v^t or v_t is constant. Which one?

5. **Equation of motion of a massless particle.** From the geodesic equation

$$\frac{dv^\kappa}{d\lambda} + \Gamma_{\mu\nu}^\kappa v^\mu v^\nu = 0 \quad (2.158)$$

show that the spatial components $\mathbf{v} \equiv \{v^x, v^y, v^z\}$ satisfy

$$\frac{d\mathbf{v}}{d\lambda} = 2\mathbf{v} \times (\mathbf{v} \times \nabla\Phi), \quad (2.159)$$

where boldface symbols represent 3D vectors, and in particular $\nabla\Phi$ is the spatial 3D gradient $\nabla\Phi \equiv \partial\Phi/\partial x^\alpha = \{\partial\Phi/\partial x, \partial\Phi/\partial y, \partial\Phi/\partial z\}$.

6. Interpret your answer, equation (2.159). In what ways does this equation for the acceleration of photons differ from the equation governing the acceleration of massive particles? [Hint: Without loss of generality, the affine parameter can be normalized so that the photon speed is one, $v = 1$, so that $\mathbf{v}$ is a unit vector representing the direction of the photon.]

7. Consider an observer who happens to be at rest in the Newtonian metric, so that $u^x = u^y = u^z = 0$. Argue that the energy of a photon observed by this observer, relative to an observer at rest at zero potential, is

$$-u^\mu v_\mu = 1 - \Phi . \quad (2.160)$$

Does the observed photon have higher or lower energy in a deeper potential well?

Exercise 2.19 Deflection of light by the Sun.

1. Consider light that passes by a spherical mass M sufficiently far away that the potential Φ is always weak. The potential at distance r from the spherical mass can be approximated by the Newtonian potential

$$\Phi = -\frac{GM}{r} . \quad (2.161)$$

Approximate the unperturbed path of light past the mass as a straight line. The plan is to calculate the deflection as a perturbation to the straight line (physicists call this the Born approximation). For definiteness, take the light to be moving in the x -direction, offset by a constant amount y away from the mass in the y -direction (so y is the impact parameter, or periapsis). Argue that equation (2.159) becomes

$$\frac{dv^y}{d\lambda} = v^x \frac{dv^y}{dx} = -2(v^x)^2 \frac{\partial \Phi}{\partial y} . \quad (2.162)$$

Integrate this equation to show that

$$\frac{\Delta v^y}{v^x} = -\frac{4GM}{y} . \quad (2.163)$$

Argue that this equals the deflection angle $\Delta\phi$.

2. Calculate the predicted deflection angle $\Delta\phi$ in arcseconds for light that just grazes the limb of the Sun.

Exercise 2.20 Shapiro time delay. The three classic tests of general relativity are the gravitational redshift (Exercise 2.10), the gravitational bending of light around the Sun (Exercise 2.19), and the precession of Mercury (Exercise 7.9). Shapiro (1964) pointed out a fourth test, that the round-trip time for a light beam bounced off a planet or spacecraft would be lengthened slightly by the passage of the light through the gravitational potential of the Sun. The experiment could be done with radio signals, since the Sun does not overwhelm a radio signal passing near its limb. In Exercise 2.18 you showed that the time component of the 4-velocity $v^\mu \equiv dx^\mu/d\lambda$ of a massless particle moving through a weak gravitational potential Φ is (units $c = 1$)

$$v^\mu \equiv \left\{ \frac{dt}{d\lambda}, \frac{d\mathbf{x}}{d\lambda} \right\} = \{v^t, \mathbf{v}\} = \{1 - 2\Phi, \mathbf{v}\} , \quad (2.164)$$

where $\mathbf{v}$ is a 3-vector of unit magnitude. Equation (2.164) implies that

$$\frac{dt}{dl} = 1 - 2\Phi , \quad (2.165)$$

where $dl \equiv |d\mathbf{x}|$ is the magnitude of the 3-vector interval $d\mathbf{x}$. The Shapiro time delay comes from the 2Φ correction.

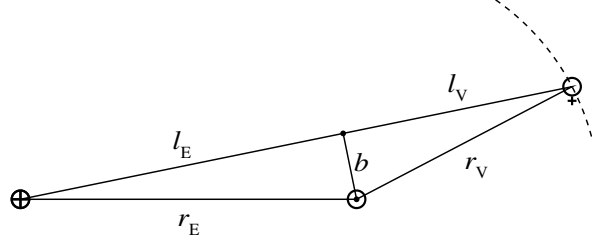


Figure 2.6 A person on Earth sends out a radio signal that passes by the Sun, bounces off the planet Venus, and returns to Earth.

1. **Time delay.** The potential Φ at distance r from the Sun is

$$\Phi = -\frac{GM_{\odot}}{r} . \quad (2.166)$$

Assume that the path of the light can be well-approximated as a straight line, as illustrated in Figure 2.6. Show that the round-trip time Δt is, with units of c restored,

$$\Delta t = \frac{2}{c}(l_E + l_V) + \frac{4GM_{\odot}}{c^3} \ln \left[\frac{(r_E + l_E)(r_V + l_V)}{b^2} \right] , \quad (2.167)$$

where, as illustrated in Figure 2.6, r_E and r_V are the distances of Earth and Venus from the Sun, b is the impact parameter, and l_E and l_V are the distances of Earth and Venus from the point of closest approach. The first term in equation (2.167) is the Newtonian expectation, while the last term in equation (2.167) is the Shapiro term.

2. **Shapiro time delay for the Earth-Venus-Sun system.** Evaluate the Shapiro time delay, in milliseconds, for the Earth-Venus-Sun system when the radio signal just grazes the limb of the Sun, with $b = R_{\odot}$. [Hint: The Earth-Sun distance is $r_E = 1.496 \times 10^{11}$ m, while the Venus-Sun distance is $r_V = 1.082 \times 10^{11}$ m.]
3. **Change in the time delay as the planets orbit.** Assume that Earth and Venus are in circular orbit about the Sun (so r_E and r_V are constant). What are the derivatives dl_E/db and dl_V/db , in terms of l_E , l_V , and b ? Deduce an expression for $c d\Delta t/db$. Identify which is the Newtonian contribution, and which the Shapiro contribution. Among the terms in the Shapiro contribution, which one term dominates for small impact parameters, where $b \ll r_E$ and $b \ll r_V$?
4. **Relative sizes of Newtonian and Shapiro terms.** From your results in part (c), calculate approximately the relative sizes of the Newtonian and Shapiro contributions to the variation $c d\Delta t/db$ of the time delay when the radio signal just grazes the limb of the Sun, $b = R_{\odot}$. Comment.

Exercise 2.21 Gravitational lensing. In Exercise 2.19 you found that, in the weak field limit, light

passing a spherical mass M at impact parameter y is deflected by angle

$$\Delta\phi = \frac{4GM}{yc^2} . \quad (2.168)$$

1. **Lensing equation.** Argue that the deflection angle $\Delta\phi$ is related to the angles α and β illustrated in

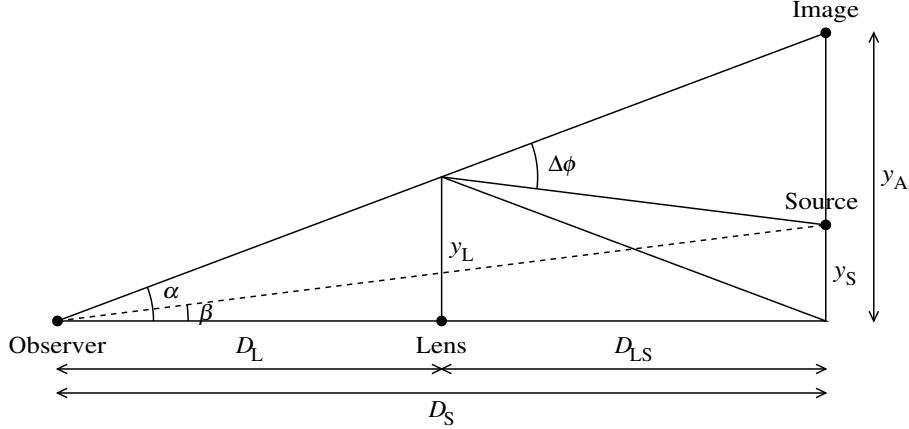


Figure 2.7 Lensing diagram.

the lensing diagram in Figure 2.7 by

$$\alpha D_S = \beta D_S + \Delta\phi D_{LS} . \quad (2.169)$$

Hence or otherwise obtain the “lensing equation” in the form commonly used by astronomers

$$\beta = \alpha - \frac{\alpha_E^2}{\alpha} , \quad (2.170)$$

where

$$\alpha_E = \left(\frac{4GM}{c^2} \frac{D_{LS}}{D_L D_S} \right)^{1/2} . \quad (2.171)$$

2. **Solutions.** Equation (2.170) has two solutions for the apparent angles α in terms of β . What are they? Sketch both solutions on a lensing diagram similar to Figure 2.7.
3. **Magnification.** Figure 2.8 illustrates the appearance of a finite-sized source lensed by a point gravitational lens. If the source is far from the lens, then the source redshift is unchanged by the gravitational lensing. But the distortion changes the apparent brightness of the source by a magnification μ equal to the ratio of the apparent area of the lensed source to that of the unlensed source. For a small source, the ratio of areas is

$$\mu = \frac{y_A dy_A}{y_S dy_S} . \quad (2.172)$$

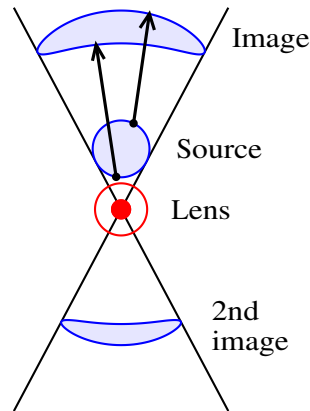


Figure 2.8 The appearance of a source lensed by a point lens. The lens in this case is a black hole, whose physical size is the filled circle, and whose apparent (lensed) size is the surrounding unfilled circle. However, any mass, not just a black hole, will lens a background source.

What is the magnification of a small source in terms of α and α_E ? When is the magnification largest?

4. **Einstein ring around the Sun?** The case $\alpha = \alpha_E$ evidently corresponds to the case where the source is exactly behind the lens, $\beta = 0$. In this case the lensed source appears as an “Einstein ring” of light around the lens. Could there be an Einstein ring around the Sun, as seen from Earth?
 5. **Einstein ring around Sgr A*.** What is the maximum possible angular size of an Einstein ring around the $4 \times 10^6 M_\odot$ black hole at the center of our Milky Way, 8 kpc away? Might this be observable?
-

3

More on the coordinate approach

3.1 Weyl tensor

The trace-free, or tidal, part of the Riemann curvature tensor defines the **Weyl tensor** $C_{\kappa\lambda\mu\nu}$

$$C_{\kappa\lambda\mu\nu} \equiv R_{\kappa\lambda\mu\nu} - \frac{1}{2} (g_{\kappa\mu} R_{\lambda\nu} - g_{\kappa\nu} R_{\lambda\mu} + g_{\lambda\nu} R_{\kappa\mu} - g_{\lambda\mu} R_{\kappa\nu}) + \frac{1}{6} (g_{\kappa\mu} g_{\lambda\nu} - g_{\kappa\nu} g_{\lambda\mu}) R \quad \text{a coordinate tensor .} \quad (3.1)$$

The Weyl tensor is by construction trace-free, meaning that it vanishes on contraction of any two indices, which is true with or without torsion.

If torsion vanishes as general relativity assumes, then the Weyl tensor has 10 independent components, which together with the 10 components of the Ricci tensor account for the 20 distinct components of the Riemann tensor. The Weyl tensor $C_{\kappa\lambda\mu\nu}$ inherits the symmetries (2.118) of the Riemann tensor, which for vanishing torsion are

$$C_{\kappa\lambda\mu\nu} = C_{([\kappa\lambda][\mu\nu])} . \quad (3.2)$$

Whereas the Einstein tensor $G_{\kappa\mu}$ necessarily vanishes in a region of spacetime where there is no energy-momentum, $T_{\kappa\mu} = 0$, the Weyl tensor does not. The Weyl tensor expresses the presence of tidal gravitational forces, and of gravitational waves.

If torsion does not vanish, then the Weyl tensor has 20 independent components, which together with the 16 components of the Ricci tensor account for the 36 distinct components of the Riemann tensor with torsion. The 6 antisymmetric components $G_{[\kappa\mu]}$ of the Einstein tensor vanish if torsion vanishes, and likewise the 10 antisymmetric components $C_{[[\kappa\lambda][\mu\nu]]}$ of the Weyl tensor vanish if torsion vanishes. With or without torsion, the 10 symmetric components $C_{([\kappa\lambda][\mu\nu])}$ of the Weyl tensor encode gravitational waves that propagate in empty space.

Exercise 3.1 Weyl tensor in arbitrary dimensions. What is the Weyl tensor in N spacetime dimensions?

Solution. The Weyl tensor is the trace-free part of the Riemann tensor. In N spacetime dimensions it is

given by the same expression (3.1) but with different coefficients,

$$C_{\kappa\lambda\mu\nu} \equiv R_{\kappa\lambda\mu\nu} - \frac{1}{N-2} (g_{\kappa\mu} R_{\lambda\nu} - g_{\kappa\nu} R_{\lambda\mu} + g_{\lambda\nu} R_{\kappa\mu} - g_{\lambda\mu} R_{\kappa\nu}) + \frac{1}{(N-1)(N-2)} (g_{\kappa\mu} g_{\lambda\nu} - g_{\kappa\nu} g_{\lambda\mu}) R . \quad (3.3)$$

The Weyl tensor vanishes identically in $N = 2$ and 3 spacetime dimensions.

Exercise 3.2 Number of components of the Riemann, Ricci, and Weyl tensors in arbitrary dimensions. How many components do the Riemann, Ricci, and Weyl tensors have in N spacetime dimensions?

Solution. The number of components depends on the total number N of spacetime dimensions, regardless of how many of those dimensions are timelike or spacelike. With torsion, the Riemann tensor is a matrix of bivectors. If torsion vanishes, the cyclic symmetry (2.120) imposes $\frac{1}{6}N^2(N-1)(N-2)$ conditions. Thus the number of components components of the Riemann tensor with and without torsion is

$$\text{Riemann torsion-full: } \left(\frac{1}{2}N(N-1)\right)^2 , \quad (3.4a)$$

$$\text{Riemann torsion-free: } \frac{1}{12}(N+1)N^2(N-1) . \quad (3.4b)$$

The Ricci tensor is the trace-full part of the Riemann tensor. In $N \geq 3$ spacetime dimensions, the Ricci tensor with torsion is a matrix of vectors, and without torsion is a symmetric matrix of vectors. Thus the number of components of the Ricci tensor with and without torsion is

$$\text{Ricci torsion-full: } N^2 , \quad (3.5a)$$

$$\text{Ricci torsion-free: } \frac{1}{2}(N+1)N . \quad (3.5b)$$

The Weyl tensor is the trace-free part of the Riemann tensor. The number of Weyl components is the difference between the number of Riemann and Ricci components, which with and without torsion is, in $N \geq 3$ spacetime dimensions,

$$\text{Weyl torsion-full: } \frac{1}{4}(N+1)N^2(N-3) , \quad (3.6a)$$

$$\text{Weyl torsion-free: } \frac{1}{12}(N+2)(N+1)N(N-3) . \quad (3.6b)$$

Equations (3.5) and (3.6) hold only for $N \geq 3$. For $N = 2$, the Riemann tensor has 1 component, the Ricci tensor 1 component, and the Weyl tensor 0 components, equation (11.93).

3.2 Evolution equations for the Weyl tensor, and gravitational waves

This section shows how the evolution equations for the Weyl tensor resemble Maxwell's equations for the electromagnetic field, and how the Weyl tensor encodes gravitational waves. In this section, torsion is taken to vanish, as general relativity assumes.

Contracted on one index, the torsion-free Bianchi identities (2.127) are

$$D_{[\kappa} R_{\lambda\mu]\nu}{}^{\kappa} = D_{\kappa} R_{\lambda\mu\nu}{}^{\kappa} + D_{\lambda} R_{\mu\nu}{}^{\kappa} - D_{\mu} R_{\lambda\nu}{}^{\kappa} = 0 . \quad (3.7)$$

In 4-dimensional spacetime, there are 20 such independent contracted identities, consisting of 4 trace identities obtained by contracting over $\lambda\nu$, and 16 trace-free identities. Since this is the same as the number 20 of independent torsion-free Bianchi identities, it follows that the singly-contracted Bianchi identities (3.7) are equivalent to the full set of Bianchi identities (2.128). An explicit expression for the Bianchi identities in terms of the contracted Bianchi identities is, in 4-dimensions (in 5 or higher dimensions there are additional terms),

$$D_{[\kappa} R_{\lambda\mu]}{}^{\nu\pi} = (18 \delta_{[\kappa}^{\rho} \delta_{\lambda}^{\sigma} \delta_{\mu]}^{[\nu} \delta_{\tau}^{\pi]} + 9 \delta_{\tau}^{\rho} \delta_{[\kappa}^{\sigma} \delta_{\lambda}^{\nu} \delta_{\mu]}^{\pi]} D_{[\nu} R_{\rho\sigma]}{}^{\tau\pi} \quad (4\text{D spacetime}) . \quad (3.8)$$

If the Riemann tensor is separated into its trace (Ricci) and traceless (Weyl) parts, equation (3.1), then the contracted Bianchi identities (3.7) become the Weyl evolution equations

$$D^{\kappa} C_{\kappa\lambda\mu\nu} = J_{\lambda\mu\nu} , \quad (3.9)$$

where $J_{\lambda\mu\nu}$ is the Weyl current

$$J_{\lambda\mu\nu} \equiv \frac{1}{2} (D_{\mu} G_{\lambda\nu} - D_{\nu} G_{\lambda\mu}) - \frac{1}{6} (g_{\lambda\nu} D_{\mu} G - g_{\lambda\mu} D_{\nu} G) . \quad (3.10)$$

The Weyl evolution equations (3.9) can be regarded as the gravitational analogue of Maxwell's equations of electromagnetism.

The Weyl current $J_{\lambda\mu\nu}$ is a vector of bivectors, which would suggest that it has $4 \times 6 = 24$ components, but it loses 4 of those components because of the cyclic identity (2.117), valid for vanishing torsion, which implies the cyclic symmetry

$$J_{[\lambda\mu\nu]} = 0 . \quad (3.11)$$

Thus the torsion-free Weyl current $J_{\lambda\mu\nu}$ has 20 independent components, in agreement with the above assertion that there are 20 independent torsion-free contracted Bianchi identities. Since the Weyl tensor is traceless, contracting the Weyl evolution equations (3.9) on $\lambda\mu$ yields zero on the left hand side, so that the contracted Weyl current satisfies

$$J^{\lambda}{}_{\lambda\nu} = 0 . \quad (3.12)$$

This doubly-contracted Bianchi identity, which is the same as equation (2.130), enforces conservation of energy-momentum. Unlike the cyclic symmetry (3.11), which follows from the cyclic symmetry of the Riemann tensor and is not a differential condition on the Riemann tensor, equations (3.12) constitute a non-trivial set of 4 differential conditions on the Einstein tensor. Besides the algebraic relations (3.11) and (3.12), the Weyl current satisfies 6 differential identities comprising the conservation law

$$D^{\lambda} J_{\lambda\mu\nu} = 0 \quad (3.13)$$

in view of equation (3.9) and the antisymmetry of $C_{\kappa\lambda\mu\nu}$ with respect to the indices $\kappa\lambda$. The Weyl current conservation law (3.13) follows from the form (3.10) of the Weyl current, coupled with covariant conservation of the Einstein tensor, equation (2.130), so does not impose any additional non-trivial conditions on the Riemann tensor. The Weyl current conservation law (3.13) is the gravitational analogue of the conservation law for electric current that follows from Maxwell's equations.

Whereas the Einstein equations relating the Einstein tensor to the energy-momentum tensor are postulated equations of general relativity, the evolution equations (3.9) for the Weyl tensor, and the equations enforcing covariant conservation of the Einstein tensor, follow mathematically from the Bianchi identities, and do not represent additional assumptions of the theory.

Exercise 3.3 Number of Bianchi identities. Confirm the counting of degrees of freedom.

3.3 Geodesic deviation

This section on geodesic deviation is included not because the equation of geodesic deviation is crucial to everyday calculations in general relativity, but rather for two reasons. First, the equation offers insight into the physical meaning of the Riemann tensor. Second, the derivation of the equation offers a fine illustration of the fact that in general relativity, whenever you take differences at infinitesimally separated points in space or time, you should always take covariant differences.

Consider two objects that are free-falling along two infinitesimally separated geodesics. In flat space the acceleration between the two objects would be zero, but in curved space the curvature induces a finite acceleration between the two objects. This is how an observer can measure curvature, at least in principle: set up an ensemble of objects initially at rest a small distance away from the observer in the observer's locally inertial frame, and watch how the objects begin to move. The equation (3.20) that describes this acceleration between objects an infinitesimal distance apart is called the **equation of geodesic deviation**.

The covariant difference in the velocities of two objects an infinitesimal distance δx^μ apart is

$$\frac{D\delta x^\mu}{D\tau} = \delta u^\mu . \quad (3.14)$$

In general relativity, the ordinary difference between vectors at two points a small interval apart is not a physically meaningful thing, because the frames of reference at the two points are different. The only physically meaningful difference is the covariant difference, which is the difference in the two vectors parallel-transported across the gap between them. It is only this covariant difference that is independent of the frame of reference. On the left hand side of equation (3.14), the proper time derivative must be the covariant proper time derivative, $D/D\tau = u^\lambda D_\lambda$. On the right hand side of equation (3.14), the difference in the 4-velocity at two points δx^κ apart must be the covariant difference $\delta = \delta x^\kappa D_\kappa$. Thus equation (3.14) means explicitly the covariant equation

$$u^\lambda D_\lambda \delta x^\mu = \delta x^\kappa D_\kappa u^\mu . \quad (3.15)$$

To derive the equation of geodesic deviation, first vary the geodesic equation $Du_\mu/D\tau = 0$ (the index μ is put downstairs so that the final equation (3.20) looks cosmetically better, but of course since everything is

covariant the μ index could just as well be put upstairs everywhere):

$$\begin{aligned}
 0 &= \delta \frac{Du_\mu}{D\tau} \\
 &= \delta x^\kappa D_\kappa (u^\lambda D_\lambda u_\mu) \\
 &= \delta u^\lambda D_\lambda u_\mu + \delta x^\kappa u^\lambda D_\kappa D_\lambda u_\mu .
 \end{aligned} \tag{3.16}$$

On the second line, the covariant difference δ between quantities a small distance δx^κ apart has been set equal to $\delta x^\kappa D_\kappa$, while $D/D\tau$ has been set equal to the covariant time derivative $u^\lambda D_\lambda$ along the geodesic. On the last line, $\delta x^\kappa D_\kappa u^\lambda$ has been replaced by δu^μ . Next, consider the covariant acceleration of the interval δx_μ , which is the covariant proper time derivative of the covariant velocity difference δu_μ :

$$\begin{aligned}
 \frac{D^2 \delta x_\mu}{D\tau^2} &= \frac{D\delta u_\mu}{D\tau} \\
 &= u^\lambda D_\lambda (\delta x^\kappa D_\kappa u_\mu) \\
 &= \delta u^\kappa D_\kappa u_\mu + \delta x^\kappa u^\lambda D_\lambda D_\kappa u_\mu .
 \end{aligned} \tag{3.17}$$

As in the previous equation (3.16), on the second line $D/D\tau$ has been set equal to $u^\lambda D_\lambda$, while δ has been set equal to $\delta x^\kappa D_\kappa$. On the last line, $u^\lambda D_\lambda \delta x^\kappa$ has been set equal to δu^μ , equation (3.15). Subtracting (3.16) from (3.17) gives

$$\frac{D^2 \delta x_\mu}{D\tau^2} = \delta x^\kappa u^\lambda [D_\lambda, D_\kappa] u_\mu , \tag{3.18}$$

or equivalently

$$\frac{D^2 \delta x_\mu}{D\tau^2} + S_{\kappa\lambda}^\nu \delta x^\kappa u^\lambda D_\nu u_\mu + R_{\kappa\lambda\mu\nu} \delta x^\kappa u^\lambda u^\nu = 0 . \tag{3.19}$$

If torsion vanishes as general relativity assumes, then

$$\boxed{\frac{D^2 \delta x_\mu}{D\tau^2} + R_{\kappa\lambda\mu\nu} \delta x^\kappa u^\lambda u^\nu = 0} , \tag{3.20}$$

which is the desired equation of geodesic deviation.

Action principle for point particles

This Chapter describes the action principle for point particles in a prescribed gravitational field. The action principle provides a powerful way to obtain equations of motion for particles in a given spacetime, such as a black hole, or a cosmological spacetime. An action principle for the gravitational field itself is deferred to Chapter 16, after development of the tetrad formalism in Chapter 11.

Hamilton's principle of least action postulates that any dynamical system is characterized by a scalar action S , which has the property that when the system evolves from one specified state to another, the path by which it gets between the two states is such as to minimize the action. The action need not be a global minimum, just a local minimum with respect to small variations in the path between fixed initial and final states.

That nature appears to respect a principle of such simplicity and power is quite remarkable, and a deep mystery. But it works, and in modern physics, the principle of least action has become a basic building block with which physicists construct theories.

From a practical perspective, the principle of least action, in either Lagrangian or Hamiltonian form, provides the most powerful way to solve equations of motion. For example, integrals of motion associated with symmetries of the spacetime emerge automatically in the Lagrangian or Hamiltonian formalisms.

4.1 Principle of least action for point particles

The path of a point particle through spacetime is specified by its coordinates $x^\mu(\lambda)$ as a function of some arbitrary parameter λ . In nonrelativistic mechanics it is usual to take the parameter λ to be the time t , and the path of a particle through space is then specified by three spatial coordinates $x^a(t)$. In relativity however it is more natural to treat the time and space coordinates on an equal footing, and to regard the path of a particle as being specified by four spacetime coordinates $x^\mu(\lambda)$ as a function of an arbitrary parameter λ , as illustrated in Figure 4.1. The parameter λ is simply a differentiable parameter that labels points along the path, and has no physical significance (for example, it is not necessarily an affine parameter).

The path of a system of N point particles through spacetime is specified by $4N$ coordinates $x^\mu(\lambda)$. The action principle postulates that, for a system of N point particles, the **action** S is an integral of a **Lagrangian**

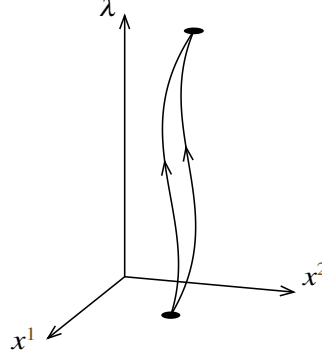


Figure 4.1 The action principle considers various paths through spacetime between fixed initial and final conditions, and chooses that path that minimizes the action.

$L(x^\mu, dx^\mu/d\lambda)$ which is a function of the $4N$ coordinates $x^\mu(\lambda)$ together with the $4N$ velocities $dx^\mu/d\lambda$ with respect to the arbitrary parameter λ . The action from an initial state at λ_i to a final state at λ_f is thus

$$S = \int_{\lambda_i}^{\lambda_f} L \left(x^\mu, \frac{dx^\mu}{d\lambda} \right) d\lambda . \quad (4.1)$$

The principle of least action demands that the actual path taken by the system between given initial and final coordinates x_i^μ and x_f^μ is such as to minimize the action. Thus the variation δS of the action must be zero under any change δx^μ in the path, subject to the constraint that the coordinates at the endpoints are fixed, $\delta x_i^\mu = 0$ and $\delta x_f^\mu = 0$,

$$\delta S = \int_{\lambda_i}^{\lambda_f} \left(\frac{\partial L}{\partial x^\mu} \delta x^\mu + \frac{\partial L}{\partial (dx^\mu/d\lambda)} \delta(dx^\mu/d\lambda) \right) d\lambda = 0 . \quad (4.2)$$

Linearity of the derivative,

$$\frac{d}{d\lambda} (x^\mu + \delta x^\mu) = \frac{dx^\mu}{d\lambda} + \frac{d(\delta x^\mu)}{d\lambda} , \quad (4.3)$$

shows that the change in the velocity along the path equals the velocity of the change, $\delta(dx^\mu/d\lambda) = d(\delta x^\mu)/d\lambda$. Integrating the second term in the integrand of equation (4.2) by parts yields

$$\delta S = \left[\frac{\partial L}{\partial (dx^\mu/d\lambda)} \delta x^\mu \right]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left(\frac{\partial L}{\partial x^\mu} - \frac{d}{d\lambda} \frac{\partial L}{\partial (dx^\mu/d\lambda)} \right) \delta x^\mu d\lambda = 0 . \quad (4.4)$$

The surface term in equation (4.4) vanishes, since by hypothesis the coordinates are held fixed at the endpoints, so $\delta x^\mu = 0$ at the endpoints. Therefore the integral in equation (4.4) must vanish. Indeed, least action requires the integral to vanish for all possible variations δx^μ in the path. The only way this can happen

is that the integrand must be identically zero. The result is the **Euler-Lagrange equations of motion**

$$\boxed{\frac{d}{d\lambda} \frac{\partial L}{\partial(dx^\mu/d\lambda)} - \frac{\partial L}{\partial x^\mu} = 0} . \quad (4.5)$$

It might seem that the Euler-Lagrange equations (4.5) are inadequately specified, since they depend on some arbitrary unknown parameter λ . But in fact the Euler-Lagrange equations are the same regardless of the choice of λ . An example of the arbitrariness of λ will be seen in §4.3. Since λ can be chosen arbitrarily, it is common to choose it in some convenient fashion. For a massive particle, λ can be taken equal to the proper time τ of the particle. For a massless particle, whose proper time never progresses, λ can be taken equal to an affine parameter.

Concept question 4.1 Redundant time coordinates? How can it be possible to treat the time coordinate t for each particle as an independent coordinate? Isn't the time coordinate t the same for all N particles? **Answer.** Different particles follow different trajectories in spacetime. One is free to choose $t(\lambda)$ to be a different function of the parameter λ for each particle, in the same way that the spatial coordinate $x^\alpha(\lambda)$ may be a different function for each particle.

4.2 Generalized momentum

The left hand side of the Euler-Lagrange equations of motion (4.5) involves the partial derivative of the Lagrangian with respect to the velocity $dx^\mu/d\lambda$. This quantity plays a fundamental role in the Hamiltonian formulation of the action principle, §4.10, and is called the **generalized momentum** π_μ conjugate to the coordinate x^μ ,

$$\boxed{\pi_\mu \equiv \frac{\partial L}{\partial(dx^\mu/d\lambda)}} . \quad (4.6)$$

4.3 Lagrangian for a test particle

According to the principle of equivalence, a test particle in a gravitating system moves along a geodesic, a straight line relative to local free-falling frames. A geodesic is the shortest distance between two points. In relativity this translates, for a massive particle, into the longest proper time between two points. The proper time along any path is $d\tau = \sqrt{-ds^2} = \sqrt{-g_{\mu\nu}dx^\mu dx^\nu}$. Thus the action S_m of a test particle of constant rest mass m in a gravitating system is

$$S_m = -m \int_{\lambda_i}^{\lambda_f} d\tau = -m \int_{\lambda_i}^{\lambda_f} \sqrt{-g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}} d\lambda . \quad (4.7)$$

The factor of rest mass m brings the action, which has units of angular momentum, to standard normalization. The overall minus sign comes from the fact that the action is a minimum whereas the proper time is a maximum along the path. The action principle requires that the Lagrangian $L(x^\mu, dx^\mu/d\lambda)$ be written as a function of the coordinates x^μ and velocities $dx^\mu/d\lambda$, and it is seen that the integrand in the last expression of equation (4.7) has the desired form, the metric $g_{\mu\nu}$ being considered a given function of the coordinates. Thus the Lagrangian L_m of a test particle of mass m is

$$L_m = -m \sqrt{-g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}} . \quad (4.8)$$

The partial derivatives that go in the Euler-Lagrange equations (4.5) are then

$$\frac{\partial L_m}{\partial(dx^\kappa/d\lambda)} = -m \frac{-g_{\kappa\nu} \frac{dx^\nu}{d\lambda}}{\sqrt{-g_{\pi\rho}(dx^\pi/d\lambda)(dx^\rho/d\lambda)}} , \quad (4.9a)$$

$$\frac{\partial L_m}{\partial x^\kappa} = -m \frac{-\frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}}{\sqrt{-g_{\pi\rho}(dx^\pi/d\lambda)(dx^\rho/d\lambda)}} . \quad (4.9b)$$

The denominators in the expressions (4.9) for the partial derivatives of the Lagrangian are $\sqrt{-g_{\pi\rho}(dx^\pi/d\lambda)(dx^\rho/d\lambda)} = d\tau/d\lambda$. It was not legitimate to make this substitution before taking the partial derivatives, since the Euler-Lagrange equations require that the Lagrangian be expressed in terms of x^μ and $dx^\mu/d\lambda$, but it is fine to make the substitution now that the partial derivatives have been obtained. The partial derivatives (4.9) thus simplify to

$$\frac{\partial L_m}{\partial(dx^\kappa/d\lambda)} = m g_{\kappa\nu} \frac{dx^\nu}{d\lambda} \frac{d\lambda}{d\tau} = m u_\kappa , \quad (4.10a)$$

$$\frac{\partial L_m}{\partial x^\kappa} = \frac{1}{2} m \frac{\partial g_{\mu\nu}}{\partial x^\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} \frac{d\lambda}{d\tau} = m \Gamma_{\mu\nu\kappa} u^\mu u^\nu \frac{d\tau}{d\lambda} , \quad (4.10b)$$

in which $u^\kappa \equiv dx^\kappa/d\tau$ is the usual 4-velocity, and the derivative of the metric has been replaced by connections in accordance with equation (2.59). The generalized momentum π_κ , equation (4.6), of the test particle coincides with its ordinary momentum p_κ :

$$\pi_\kappa = p_\kappa \equiv m u_\kappa . \quad (4.11)$$

The resulting Euler-Lagrange equations of motion (4.5) are

$$\frac{dm u_\kappa}{d\lambda} = m \Gamma_{\mu\nu\kappa} u^\mu u^\nu \frac{d\tau}{d\lambda} . \quad (4.12)$$

As remarked in §4.1, the choice of the arbitrary parameter λ has no effect on the equations of motion. With a factor of $m d\tau/d\lambda$ cancelled, equation (4.12) becomes

$$\frac{du_\kappa}{d\tau} = \Gamma_{\mu\nu\kappa} u^\mu u^\nu . \quad (4.13)$$

Splitting the connection $\Gamma_{\mu\nu\kappa}$ into its torsion-free part $\mathring{\Gamma}_{\mu\nu\kappa}$ and the contortion $K_{\mu\nu\kappa}$, equation (2.64), gives

$$\frac{du_\kappa}{d\tau} = (\mathring{\Gamma}_{\mu\nu\kappa} + K_{\mu\nu\kappa})u^\mu u^\nu = \mathring{\Gamma}_{\mu\kappa\nu}u^\mu u^\nu, \quad (4.14)$$

where the last step follows from the symmetry of the torsion-free connection $\mathring{\Gamma}_{\mu\nu\kappa}$ in its last two indices, and the antisymmetry of the contortion tensor $K_{\mu\nu\kappa}$ in its first two indices. With or without torsion, equation (4.14) yields the torsion-free geodesic equation of motion,

$$\boxed{\frac{\mathring{D}u_\kappa}{D\tau} = 0}. \quad (4.15)$$

Equation (4.15) shows that presence of torsion does not affect the geodesic motion of particles.

Concept question 4.2 Throw a clock up in the air.

1. This question is posed by Rovelli (2007). Standing on the surface of the Earth, you throw a clock up in the air, and catch it. Which clock shows more time elapsed, the one you threw up in the air, or the one on your wrist?
2. Suppose you throw the clock so hard that it goes around the Moon. Which clock shows more time elapsed?

4.4 Massless test particle

The equation of motion for a massless test particle is obtained from that for a massive particle in the limit of zero mass, $m \rightarrow 0$. The proper time τ along the path of a massless particle is zero, but an affine parameter $\lambda \equiv \tau/m$ proportional to proper time can be defined, equation (2.93), which remains finite in the limit $m \rightarrow 0$. In terms of the affine parameter λ , the momentum p^κ of a particle can be written

$$p^\kappa \equiv m u^\kappa = \frac{dx^\kappa}{d\lambda}, \quad (4.16)$$

and the equation of motion (4.15) becomes

$$\frac{\mathring{D}p_\kappa}{D\lambda} = 0, \quad (4.17)$$

which works for massless as well as massive particles.

The action for a test particle in terms of the affine parameter λ defined by equation (2.93) is

$$S = -m^2 \int d\lambda, \quad (4.18)$$

which vanishes for $m \rightarrow 0$. One might be worried that the action seemingly vanishes identically for a massless particle. An alternative nice action is given below, equation (4.30), that vanishes in the massless limit only after the equations of motion are imposed.

Concept question 4.3 Conventional Lagrangian. In the conventional Lagrangian approach, the parameter λ is set equal to the time coordinate t , and the Lagrangian $L(t, x^\alpha, dx^\alpha/dt)$ of a system of N particles is considered to be a function of the time t , the $3N$ spatial coordinates x^α , and the $3N$ spatial velocities dx^α/dt . Compare the conventional and covariant Lagrangian approaches for a point particle. **Answer.** The Euler-Lagrange equations in the conventional Lagrangian approach are

$$\frac{d}{dt} \frac{\partial L}{\partial (dx^\alpha/dt)} - \frac{\partial L}{\partial x^\alpha} = 0 . \quad (4.19)$$

For a point particle, the Euler-Lagrange equations (4.19) yield the spatial components of the geodesic equation of motion (4.17),

$$\frac{\dot{D}p_\alpha}{D\lambda} = 0 . \quad (4.20)$$

What about the time component of the geodesic equation of motion? The geodesic equation for the time component is a consequence of the geodesic equations for the spatial components, coupled with conservation of rest mass m ,

$$p^0 \frac{\dot{D}p_0}{D\lambda} = \frac{1}{2} \frac{\dot{D}p^0 p_0}{D\lambda} = -\frac{1}{2} \frac{\dot{D}(p^\alpha p_\alpha + m^2)}{D\lambda} = -p^\alpha \frac{\dot{D}p_\alpha}{D\lambda} = 0 . \quad (4.21)$$

Put another way, the covariant Lagrangian approach applied to a point particle enforces conservation of the rest mass m of the particle, a conservation law that the conventional Lagrangian approach simply assumes. Invariance of the action with respect to reparametrization of λ implies conservation of rest mass.

4.5 Effective Lagrangian for a test particle

A drawback of the test particle Lagrangian (4.8) is that it involves a square root. This proves to be problematic for various reasons, among which is that it is an obstacle to deriving a satisfactory super-Hamiltonian, §4.12. This section describes an alternative approach that gets rid of the square root, making the test particle Lagrangian quadratic in velocities $dx^\mu/d\lambda$, equation (4.25).

After equations of motion are imposed, the Lagrangian (4.8) for a test particle of constant rest mass m is

$$L_m = -m \frac{d\tau}{d\lambda} . \quad (4.22)$$

If the parameter λ is chosen such that $d\tau/d\lambda$ is constant,

$$\frac{d\tau}{d\lambda} = \text{constant} , \quad (4.23)$$

so that the Lagrangian L_m is constant after equations of motion are imposed, then the Euler-Lagrange equations of motion (4.5) are unchanged if the Lagrangian is replaced by any function of it,

$$L'_m = f(L_m) . \quad (4.24)$$

A convenient choice of alternative Lagrangian L'_m , also called an **effective Lagrangian**, is

$$L'_m = -\frac{L_m^2}{2m^2} = \frac{1}{2}g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} . \quad (4.25)$$

For the effective Lagrangian (4.25), the partial derivatives (4.9) are

$$\frac{\partial L'_m}{\partial(dx^\kappa/d\lambda)} = g_{\kappa\nu} \frac{dx^\nu}{d\lambda} , \quad (4.26a)$$

$$\frac{\partial L'_m}{\partial x^\kappa} = \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} = \Gamma_{\mu\nu\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} . \quad (4.26b)$$

The Euler-Lagrange equations of motion (4.5) are then

$$\frac{d}{d\lambda} \left(g_{\kappa\nu} \frac{dx^\nu}{d\lambda} \right) = \Gamma_{\mu\nu\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} . \quad (4.27)$$

Equations (4.27) are valid subject to the condition (4.23), which asserts that $d\lambda \propto d\tau$. The constant of proportionality does not affect the equations of motion (4.27), which thus reproduce the earlier equations of motion in either of the forms (4.15) or (4.17).

If the test particle is moving in a prescribed gravitational field and there are no other fields, then the equations of motion are unchanged by the normalization of the effective Lagrangian L'_m . But if there are other fields that affect the particle's motion, such as an electromagnetic field, §4.7, then the effective Lagrangian L'_m must be normalized correctly if it is to continue to recover the correct equations of motion. The correct normalization is such that the generalized momentum of the test particle, defined by equation (4.26a), equal its ordinary momentum p_μ , in agreement with equation (4.11),

$$g_{\kappa\nu} \frac{dx^\nu}{d\lambda} = p_\kappa \equiv g_{\kappa\nu} m \frac{dx^\nu}{d\tau} . \quad (4.28)$$

This requires that the constant in equation (4.23) must equal the rest mass m ,

$$\frac{d\tau}{d\lambda} = m . \quad (4.29)$$

This is just the definition of the affine parameter λ , equation (2.93). Thus the λ in the definition (4.25) of the effective Lagrangian L'_m should be interpreted as the affine parameter.

Notice that the value of the effective Lagrangian L'_m after condition (4.29) is applied (*after* equations of motion are imposed) is $-m^2/2$, which is half the value of the original Lagrangian L_m (4.8).

4.6 Nice Lagrangian for a test particle

The effective Lagrangian (4.25) has the advantage that it does not involve a square root, but this advantage was achieved at the expense of imposing the condition (4.29) ad hoc after the equations of motion are derived. It is possible to retain the advantage of a Lagrangian quadratic in velocities, but get rid of the ad

hoc condition, by modifying the Lagrangian so that the ad hoc condition essentially emerges as an equation of motion. I call the resulting Lagrangian (4.31) the “nice” Lagrangian.

As seen in §4.1, the equations of motion are independent of the choice of the arbitrary parameter λ that labels the path of the particle between its fixed endpoints. The equations of motion are said to be **reparametrization independent**. Introduce, therefore, a scale factor $a(\lambda)$, an arbitrary function of λ , that rescales the parameter λ , and let the action for a test particle of mass m be

$$S_m = \int \frac{1}{2} \left(g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} - m^2 \right) a d\lambda , \quad (4.30)$$

with nice Lagrangian

$$\boxed{L_m = \frac{a}{2} \left(g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} - m^2 \right)} . \quad (4.31)$$

Variation of the action (4.30) with respect to x^μ and $dx^\mu/d\lambda$ yields the Euler-Lagrange equations in the form

$$\frac{d}{a d\lambda} \left(g_{\kappa\nu} \frac{dx^\nu}{a d\lambda} \right) = \Gamma_{\mu\nu\kappa} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} . \quad (4.32)$$

Variation of the action (4.30) with respect to the parameter a gives

$$\delta S_m = \int \frac{1}{2} \left(-g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} - m^2 \right) \delta a d\lambda , \quad (4.33)$$

and requiring that this be an extremum imposes

$$g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} = -m^2 . \quad (4.34)$$

Equation (4.34) is equivalent to

$$a d\lambda = \frac{d\tau}{m} , \quad (4.35)$$

where the sign has been taken positive without loss of generality. Substituting equation (4.35) into the equations of motion (4.32) recovers the usual equations of motion (4.15).

Condition (4.35) substituted into the action (4.30) recovers the standard test particle action (4.7) with the correct sign and normalization.

4.7 Action for a charged test particle in an electromagnetic field

The equations of motion for a test particle of charge q in a prescribed gravitational and electromagnetic field can be obtained by adding to the test particle action S_m an interaction action S_q that characterizes the interaction between the charge and the electromagnetic field,

$$S = S_m + S_q . \quad (4.36)$$

In flat (Minkowski) space, experiment shows that the required equation of motion is the classical Lorentz force law (4.45). The Lorentz force law is recovered with the interaction action

$$S_q = q \int_{\lambda_i}^{\lambda_f} A_\mu dx^\mu = q \int_{\lambda_i}^{\lambda_f} A_\mu \frac{dx^\mu}{d\lambda} d\lambda, \quad (4.37)$$

where A_μ is the electromagnetic 4-vector potential. The interaction Lagrangian L_q corresponding to the action (4.37) is

$$L_q = q A_\mu \frac{dx^\mu}{d\lambda}. \quad (4.38)$$

If the electromagnetic potential A_μ is taken to be a prescribed function of the coordinates x^μ along the path of the particle, then the Lagrangian L_q (4.38) is a function of coordinates x^μ and velocities $dx^\mu/d\lambda$ as required by the action principle. The partial derivatives of the interaction Lagrangian L_q with respect to velocities and coordinates are

$$\frac{\partial L_q}{\partial(dx^\kappa/d\lambda)} = q A_\kappa, \quad (4.39a)$$

$$\frac{\partial L_q}{\partial x^\kappa} = q \frac{\partial A_\mu}{\partial x^\kappa} \frac{dx^\mu}{d\lambda} = q \frac{\partial A_\mu}{\partial x^\kappa} u^\mu \frac{d\tau}{d\lambda}. \quad (4.39b)$$

The generalized momentum π_κ , equation (4.6), of the test particle of mass m and charge q in the electromagnetic field of potential A_μ is, from equations (4.10a) and (4.39a),

$$\pi_\kappa \equiv \frac{\partial(L_m + L_q)}{\partial(dx^\kappa/d\lambda)} = m u_\kappa + q A_\kappa. \quad (4.40)$$

Applied to the Lagrangian $L = L_m + L_q$, the Euler-Lagrange equations (4.5) are

$$\frac{d}{d\lambda} (m u_\kappa + q A_\kappa) = \left(m \Gamma_{\mu\nu\kappa} u^\mu u^\nu + q \frac{\partial A_\mu}{\partial x^\kappa} u^\mu \right) \frac{d\tau}{d\lambda}, \quad (4.41)$$

which rearranges to

$$\frac{dm u_\kappa}{d\tau} = m \Gamma_{\mu\nu\kappa} u^\mu u^\nu + q F_{\kappa\mu} u^\mu, \quad (4.42)$$

where the antisymmetric electromagnetic field tensor $F_{\kappa\mu}$ is defined to be the torsion-free covariant curl of the electromagnetic potential A_μ ,

$$F_{\kappa\mu} \equiv \frac{\partial A_\mu}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\mu}. \quad (4.43)$$

The definition (4.43) of the electromagnetic field holds even in the presence of torsion (see §16.5). Splitting the connection in equation (4.42) into its torsion-free part and the contortion, as done previously in equation (4.14), yields the Lorentz force law for a test particle of mass m and charge q moving in a prescribed gravitational and electromagnetic field, with or without torsion,

$$\boxed{\frac{\mathring{D}m u_\kappa}{D\tau} = q F_{\kappa\mu} u^\mu}. \quad (4.44)$$

Equation (4.44), which involves the torsion-free covariant derivative $\mathring{D}/D\tau$, shows that the Lorentz force law is unaffected by the presence of torsion.

In flat (Minkowski) space, the spatial components of equation (4.44) reduce to the classical special relativistic Lorentz force law

$$\frac{d\mathbf{p}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) . \quad (4.45)$$

In equation (4.45), $\mathbf{p}$ is the 3-momentum and $\mathbf{v}$ is the 3-velocity, related to the 4-momentum and 4-velocity by $p^k = \{p^t, \mathbf{p}\} = mu^k = mu^t\{1, \mathbf{v}\}$ (note that $d/dt = (1/u^t) d/d\tau$). In flat space, the components of the electric and magnetic fields $\mathbf{E} = \{E_x, E_y, E_z\}$ and $\mathbf{B} = \{B_x, B_y, B_z\}$ are related to the electromagnetic field tensor F_{mn} by (the signs in the expression (4.46) are arranged precisely so as to agree with the classical law (4.45))

$$F_{mn} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}, \quad F^{mn} = \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{pmatrix}. \quad (4.46)$$

If the electromagnetic 4-potential A^m is written in terms of a classical electric potential ϕ and electric 3-vector potential $\mathbf{A} \equiv \{A_x, A_y, A_z\}$,

$$A^m = \{\phi, \mathbf{A}\}, \quad (4.47)$$

then in flat space equation (4.43) reduces to the traditional relations for the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ in terms of the potentials ϕ and $\mathbf{A}$,

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}, \quad (4.48)$$

where $\nabla \equiv \{\partial/\partial x, \partial/\partial y, \partial/\partial z\}$ is the spatial 3-gradient.

4.8 Symmetries and constants of motion

If a spacetime possesses a symmetry of some kind, then a test particle moving in that spacetime possesses an associated constant of motion. The Lagrangian formalism makes it transparent how to relate symmetries to constants of motion.

Suppose that the Lagrangian of a particle has some spacetime symmetry, such as time translation symmetry, or spatial translation symmetry, or rotational symmetry. In a suitable coordinate system, the symmetry is expressed by the condition that the Lagrangian L is independent of some coordinate, call it ξ . In the case of time translation symmetry, for example, the coordinate would be a suitable time coordinate t . Coordinate independence requires that the metric $g_{\mu\nu}$, along with any other field, such as an electromagnetic field, that may affect the particle's motion, is independent of the coordinate ξ . Then the Euler-Lagrangian equations

of motion (4.5) imply that the derivative of the covariant ξ -component π_ξ of the conjugate momentum of the particle vanishes along the trajectory of the particle,

$$\frac{d\pi_\xi}{d\lambda} = \frac{\partial L}{\partial \xi} = 0 . \quad (4.49)$$

Thus the covariant momentum π_ξ is a constant of motion,

$$\boxed{\pi_\xi = \text{constant}} . \quad (4.50)$$

4.9 Conformal symmetries

Sometimes the Lagrangian possesses a weaker kind of symmetry, called **conformal symmetry**, in which the Lagrangian L depends on a coordinate ξ only through an overall scaling of the Lagrangian,

$$L = e^{2\xi} \tilde{L} , \quad (4.51)$$

where the conformal Lagrangian $\tilde{L}$ is independent of ξ . The factor e^ξ is called a conformal factor. The Euler-Lagrangian equation of motion (4.5) for the conformal coordinate ξ is then

$$\frac{d\pi_\xi}{d\lambda} = \frac{\partial L}{\partial \xi} = 2L . \quad (4.52)$$

As an example, consider a test particle moving in a spacetime with conformally symmetric metric

$$g_{\mu\nu} = e^{2\xi} \tilde{g}_{\mu\nu} , \quad (4.53)$$

where the conformal metric $\tilde{g}_{\mu\nu}$ is independent of the coordinate ξ . The effective Lagrangian L'_m of the test particle is given by equation (4.25). The equation of motion (4.52) becomes

$$\frac{dp_\xi}{d\lambda} = 2L'_m = -m^2 . \quad (4.54)$$

If the test particle is massive, $m \neq 0$, then equation (4.54) integrates to

$$\boxed{p_\xi = -m\tau} , \quad (4.55)$$

where a constant of integration has been absorbed, without loss of generality, into a shift of the zero point of the proper time τ of the particle. If the test particle is massless, $m = 0$, then equation (4.54) implies that

$$p_\xi = \text{constant} . \quad (4.56)$$

Exercise 4.4 Geodesics in Rindler space. The Rindler line-element (2.103) can be written

$$ds^2 = e^{2\xi} (-d\alpha^2 + d\xi^2) + dy^2 + dz^2 , \quad (4.57)$$

where the Rindler coordinates α and ξ are related to Minkowski coordinates t and x by

$$t = e^\xi \sinh \alpha , \quad x = e^\xi \cosh \alpha . \quad (4.58)$$

What are the constants of motion of a test particle? Integrate the Euler-Lagrange equations of motion.

Solution. The Rindler metric is independent of the coordinates α , y , and z . The three corresponding constants of motion are

$$p_\alpha , \quad p_y , \quad p_z . \quad (4.59)$$

A fourth integral of motion follows from conservation of rest mass

$$p^\nu p_\nu = -m^2 . \quad (4.60)$$

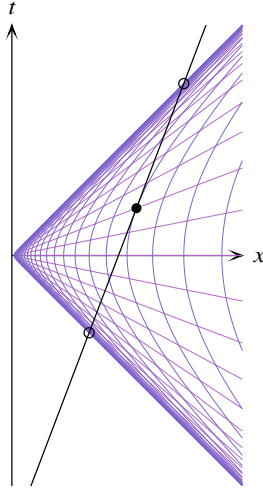


Figure 4.2 Rindler wedge of Minkowski space. Purple and blue lines are lines of constant Rindler time α and constant Rindler spatial coordinate ξ respectively. The grid of lines is equally spaced by 0.2 in each of α and ξ . The Rindler coordinates α and ξ , each extending over the interval $(-\infty, \infty)$, cover only the $x > |t|$ quadrant of Minkowski space. The fact that the Rindler metric is conformally Minkowski in α and ξ (the line-element is proportional to $-d\alpha^2 + d\xi^2$, equation (4.57)) shows up in the fact that small areal elements of the α - ξ grid are rhombi with null (45°) diagonals. The straight black line is a representative geodesic. The solid dot marks the point where the geodesic goes through $\{\alpha_0, \xi_0\}$. Open circles mark $\alpha = \mp\infty$, where the geodesic passes through the null boundaries $t = \mp x$ of the Rindler wedge.

Equation (4.60) rearranges to give

$$\frac{d\xi}{d\lambda} \equiv p^\xi = e^{-\xi} \sqrt{(e^{-\xi} p_\alpha)^2 - \mu^2} , \quad (4.61)$$

where μ is the positive constant

$$\mu \equiv \sqrt{p_y^2 + p_z^2 + m^2} . \quad (4.62)$$

Equation (4.61) integrates to give ξ as a function of λ ,

$$e^{2\xi} = \frac{p_\alpha^2}{\mu^2} - \mu^2 \lambda^2 , \quad (4.63)$$

where a constant of integration has been absorbed without loss of generality into a shift of the zero point of the affine parameter λ along the trajectory of the particle. The coordinate ξ passes through its maximum value ξ_0 where $\lambda = 0$, at which point

$$e^{\xi_0} = -\frac{p_\alpha}{\mu} , \quad (4.64)$$

the sign coming from the fact that $p_\alpha = g_{\alpha\alpha} p^\alpha = -e^{2\xi} d\alpha/d\lambda$ must be negative, since the particle must move forward in Rindler time α . The trajectory is illustrated in Figure 4.2; the trajectory is of course a straight line in the parent Minkowski space.

The evolution equation (4.63) for $\xi(\lambda)$ can be derived alternatively from the Euler-Lagrange equation for ξ ,

$$\frac{dp_\xi}{d\lambda} = -\mu^2 . \quad (4.65)$$

The Euler-Lagrange equation (4.65) integrates to

$$p_\xi = -\mu^2 \lambda , \quad (4.66)$$

where a constant of integration has again been absorbed into a shift of the zero point of the affine parameter λ (this choice is consistent with the previous one). Given that $p_\xi = g_{\xi\xi} p^\xi = e^{2\xi} d\xi/d\lambda$, equation (4.66) integrates to yield the same result (4.63), the constant of integration being established by the rest-mass relation (4.60).

The evolution of Rindler time α along the particle's trajectory follows from integrating $p_\alpha = g_{\alpha\alpha} p^\alpha = -e^{2\xi} d\alpha/d\lambda$, which gives

$$\alpha - \alpha_0 = -\frac{1}{2} \ln \left(\frac{e^{\xi_0} + \mu\lambda}{e^{\xi_0} - \mu\lambda} \right) , \quad (4.67)$$

where α_0 is the value of α for $\lambda = 0$, where ξ takes its maximum ξ_0 . The Rindler time coordinate α varies between limits $\mp\infty$ at $\mu\lambda = \mp e^{\xi_0}$.

4.10 (Super-)Hamiltonian

The Lagrangian approach characterizes the paths of particles through spacetime in terms of their $4N$ coordinates x^μ and corresponding velocities $dx^\mu/d\lambda$ along those paths. The Hamiltonian approach on the other hand characterizes the paths of particles through spacetime in terms of $4N$ coordinates x^μ and the $4N$ generalized momenta π_μ , which are treated as independent from the coordinates. In the Hamiltonian approach, the **Hamiltonian** $H(x^\mu, \pi_\mu)$ is considered to be a function of coordinates and generalized momenta, and the action is minimized with respect to independent variations of those coordinates and momenta. In the Hamiltonian approach, the coordinates and momenta are treated essentially on an equal footing.

The Hamiltonian H can be defined in terms of the Lagrangian L by

$$H \equiv \pi_\mu \frac{dx^\mu}{d\lambda} - L . \quad (4.68)$$

If desired, the indices on the momentum term can be flipped, $\pi_\mu dx^\mu/d\lambda = \pi^\mu dx_\mu/d\lambda$, where it is to be understood that $dx_\mu = g_{\mu\nu} dx^\nu$ (and not $dx_\mu = d(g_{\mu\nu} x^\nu)$), since the interval dx^μ is a tensor, but the coordinate x^μ itself is not. Here, as previously in §4.1, the parameter λ is to be regarded as an arbitrary parameter that labels the path of the system through the $8N$ -dimensional phase space of coordinates and momenta of the N particles. Misner, Thorne, and Wheeler (1973) call the Hamiltonian (4.68) the **super-Hamiltonian**, to distinguish it from the conventional Hamiltonian, equation (4.74), where the parameter λ is taken equal to the time coordinate t . Here however the super-Hamiltonian (4.68) is simply referred to as the Hamiltonian, for brevity.

In terms of the Hamiltonian (4.68), the action (4.1) is

$$S = \int_{\lambda_i}^{\lambda_f} \left(\pi_\mu \frac{dx^\mu}{d\lambda} - H \right) d\lambda . \quad (4.69)$$

In accordance with Hamilton's principle of least action, the action must be varied with respect to the coordinates and momenta along the path. The variation of the first term in the integrand of equation (4.69) can be written

$$\delta \left(\pi_\mu \frac{dx^\mu}{d\lambda} \right) = \delta \pi_\mu \frac{dx^\mu}{d\lambda} + \pi_\mu \frac{d\delta x^\mu}{d\lambda} = \delta \pi_\mu \frac{dx^\mu}{d\lambda} + \frac{d}{d\lambda} (\pi_\mu \delta x^\mu) - \frac{d\pi_\mu}{d\lambda} \delta x^\mu . \quad (4.70)$$

The middle term on the right hand side of equation (4.70) yields a surface term on integration. Thus the variation of the action is

$$\delta S = [\pi_\mu \delta x^\mu]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left\{ - \left(\frac{d\pi_\mu}{d\lambda} + \frac{\partial H}{\partial x^\mu} \right) \delta x^\mu + \left(\frac{dx^\mu}{d\lambda} - \frac{\partial H}{\partial \pi_\mu} \right) \delta \pi_\mu \right\} d\lambda , \quad (4.71)$$

which takes into account that the Hamiltonian is to be considered a function $H(x^\mu, \pi_\mu)$ of coordinates and momenta. The principle of least action requires that the action is a minimum with respect to variations of the coordinates and momenta along the paths of particles, the coordinates and momenta at the endpoints λ_i and λ_f of the integration being held fixed. Since the coordinates are fixed at the endpoints, $\delta x^\mu = 0$, the

surface term in equation (4.71) vanishes. Minimization of the action with respect to arbitrary independent variations of the coordinates and momenta then yields **Hamilton's equations of motion**

$$\boxed{\frac{dx^\mu}{d\lambda} = \frac{\partial H}{\partial \pi_\mu}, \quad \frac{d\pi_\mu}{d\lambda} = -\frac{\partial H}{\partial x^\mu}}. \quad (4.72)$$

4.11 Conventional Hamiltonian

The conventional Hamiltonian of classical mechanics is not the same as the super-Hamiltonian (4.68). In the conventional approach, the parameter λ is set equal to the time coordinate t . The Lagrangian is taken to be a function $L(t, x^\alpha, dx^\alpha/dt)$ of time t and of the $3N$ spatial coordinates x^α and $3N$ spatial velocities dx^α/dt . The generalized momenta are defined to be, analogously to (4.6),

$$\pi_\alpha \equiv \frac{\partial L}{\partial(dx^\alpha/dt)}. \quad (4.73)$$

The conventional Hamiltonian is taken to be a function $H(t, x^\alpha, \pi_\alpha)$ of time t and of the $3N$ spatial coordinates x^α and corresponding $3N$ generalized momenta π_α . The conventional Hamiltonian is related to the conventional Lagrangian by

$$H \equiv \pi_\alpha \frac{dx^\alpha}{dt} - L. \quad (4.74)$$

The conventional Hamilton's equations are

$$\frac{dx^\alpha}{dt} = \frac{\partial H}{\partial \pi_\alpha}, \quad \frac{d\pi_\alpha}{dt} = -\frac{\partial H}{\partial x^\alpha}. \quad (4.75)$$

The advantage of the super-Hamiltonian (4.68) over the conventional Hamiltonian (4.74) in general relativity will become apparent in the sections following.

4.12 Conventional Hamiltonian for a test particle

The test-particle Lagrangian (4.8) is

$$L_m = -m \sqrt{-g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}}. \quad (4.76)$$

The corresponding test-particle Hamiltonian is supposedly given by equation (4.68). However, one runs into a difficulty. The Hamiltonian is supposed to be expressed in terms of coordinates x^μ and momenta p_μ . But the expression (4.68) for the Hamiltonian depends on the arbitrary parameter λ , whereas as seen in §4.3 the coordinates x^μ and momenta p_μ are (before the least action principle is applied) independent of the choice of λ . For example, the square of the momentum (4.11) derived from the Lagrangian (4.8) is $p_\mu p^\mu = -m^2$, which is independent of the choice of λ . There is no way to express the Hamiltonian in the prescribed form

without imposing some additional constraint on λ . Two ways to achieve this are described in the next two sections, §4.13 and §4.14.

A third approach is to revert to the conventional approach of fixing the arbitrary parameter λ equal to coordinate time t . This choice eliminates the time coordinate and corresponding generalized momentum as parameters to be determined by the least action principle. It also breaks manifest covariance, by singling out the time coordinate for special treatment. For simplicity, consider flat space, where the metric is Minkowski η_{mn} . The Lagrangian (4.76) becomes

$$L_m = -m \sqrt{-\eta_{mn} \frac{dx^m}{dt} \frac{dx^n}{dt}} = -m \sqrt{1 - v^2} , \quad (4.77)$$

where $v \equiv \sqrt{\eta_{ab} v^a v^b}$ is the magnitude of the 3-velocity v^a ,

$$v^a \equiv \frac{dx^a}{dt} . \quad (4.78)$$

The generalized momentum π_a defined by (4.73) equals the ordinary momentum p_a ,

$$\pi_a = p_a \equiv \frac{mv_a}{\sqrt{1 - v^2}} . \quad (4.79)$$

The Hamiltonian (4.74) is

$$H = p_a v^a - L = \frac{m}{\sqrt{1 - v^2}} . \quad (4.80)$$

Expressed in terms of the spatial momenta p_a , the Hamiltonian is

$$H = \sqrt{p^2 + m^2} , \quad (4.81)$$

where $p \equiv \sqrt{\eta^{ab} p_a p_b}$ is the magnitude of the 3-momentum p_a . Hamilton's equations (4.75) are

$$\frac{dx^a}{dt} = \frac{p^a}{\sqrt{p^2 + m^2}} , \quad \frac{dp^a}{dt} = 0 . \quad (4.82)$$

The Hamiltonian (4.81) can be recognized as the energy of the particle, or minus the covariant time component of the 4-momentum,

$$H = -p_0 . \quad (4.83)$$

A similar, more complicated, analysis in curved space leads to the same conclusion, that the conventional Hamiltonian H is minus the covariant time component of the 4-momentum,

$$H = -p_t . \quad (4.84)$$

The expression for the Hamiltonian in terms of spatial coordinates x^α and momenta p_α can be inferred from conservation of rest mass,

$$g^{\mu\nu} p_\mu p_\nu + m^2 = 0 . \quad (4.85)$$

Explicitly, the conventional Hamiltonian is

$$H = -p_t = \frac{1}{g^{tt}} \left[g^{t\alpha} p_\alpha + \sqrt{(g^{t\alpha} g^{t\beta} - g^{tt} g^{\alpha\beta}) p_\alpha p_\beta - g^{tt} m^2} \right] . \quad (4.86)$$

In the presence of an electromagnetic field, replace the momenta p_t and p_α in equation (4.86) by $p_\mu = \pi_\mu - qA_\mu$, and set the Hamiltonian equal to $-\pi_t$,

$$H = -\pi_t . \quad (4.87)$$

The super-Hamiltonians (4.90) and (4.96) derived in the next two sections are more elegant than the conventional Hamiltonian (4.86). All lead to the same equations of motion, but the super-Hamiltonian exhibits general covariance more clearly.

4.13 Effective (super-)Hamiltonian for a test particle with electromagnetism

In the effective approach, the condition (4.29) on the parameter λ is applied *after* equations of motion are derived. The effective test-particle Lagrangian (4.25), coupled to electromagnetism, is

$$L = L_m + L_q = \frac{1}{2} g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} + q A_\mu \frac{dx^\mu}{d\lambda} , \quad (4.88)$$

where the metric $g_{\mu\nu}$ and electromagnetic potential A_μ are considered to be given functions of the coordinates x^μ . The corresponding generalized momentum (4.6) is

$$\pi_\mu = g_{\mu\nu} \frac{dx^\nu}{d\lambda} + q A_\mu . \quad (4.89)$$

The (super-)Hamiltonian (4.68) expressed in terms of coordinates x^μ and momenta π_μ as required is

$$H = \frac{1}{2} g^{\mu\nu} (\pi_\mu - q A_\mu) (\pi_\nu - q A_\nu) . \quad (4.90)$$

Hamilton's equations (4.72) are

$$\frac{dx^\mu}{d\lambda} = p^\mu , \quad \frac{dp_\kappa}{d\lambda} = \Gamma_{\mu\nu\kappa} p^\mu p^\nu + q F_{\kappa\mu} p^\mu , \quad (4.91)$$

where p_μ is defined by

$$p_\mu \equiv \pi_\mu - q A_\mu . \quad (4.92)$$

The equations of motion (4.91) having been derived from the Hamiltonian (4.90), the parameter λ is set equal to the affine parameter in accordance with condition (4.29). In particular, the first of equations (4.91) together with condition (4.29) implies that $p^\mu = m dx^\mu/d\tau$, as it should be. The equations of motion (4.91) thus reproduce the equations (4.42) derived in Lagrangian approach. The value of the Hamiltonian (4.90) after the equations of motion and condition (4.29) are imposed is constant,

$$H = -\frac{m^2}{2} . \quad (4.93)$$

Recall that the super-Hamiltonian H is a scalar, associated with rest mass, to be distinguished from the conventional Hamiltonian, which is the time component of a vector, associated with energy. The minus sign in equation (4.93) is associated with the choice of metric signature $-+++$, where scalar products of timelike quantities are negative. The negative Hamiltonian (4.93) signifies that the particle is propagating along a timelike direction. If the particle is massless, $m = 0$, then the Hamiltonian is zero (after equations of motion are imposed), signifying that the particle is propagating along a null direction.

4.14 Nice (super-)Hamiltonian for a test particle with electromagnetism

The nice test-particle Lagrangian (4.31), coupled to electromagnetism, is

$$L = \frac{a}{2} \left(g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} - m^2 \right) + q A_\mu \frac{dx^\mu}{d\lambda} . \quad (4.94)$$

The corresponding generalized momentum (4.6) is

$$\pi_\mu = g_{\mu\nu} \frac{dx^\nu}{a d\lambda} + q A_\mu . \quad (4.95)$$

The associated nice (super-)Hamiltonian (4.68) expressed in terms of coordinates x^μ and momenta π_μ as required is

$$H = \frac{a}{2} [g^{\mu\nu} (\pi_\mu - q A_\mu) (\pi_\nu - q A_\nu) + m^2] . \quad (4.96)$$

The nice Hamiltonian H , equation (4.96), depends on the auxiliary scale factor a as well as on x^μ and π_μ , and the action must be varied with respect to all of these to obtain all the equations of motion. Compared to the variation (4.71), the variation of the action contains an additional term proportional to δa :

$$\delta S = [\pi_\mu \delta x^\mu]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left\{ - \left(\frac{d\pi_\mu}{d\lambda} + \frac{\partial H}{\partial x^\mu} \right) \delta x^\mu + \left(\frac{dx^\mu}{d\lambda} - \frac{\partial H}{\partial \pi_\mu} \right) \delta \pi_\mu - \frac{\partial H}{\partial a} \delta a \right\} d\lambda . \quad (4.97)$$

Requiring that the variation (4.97) of the action vanish under arbitrary variations of the coordinates x^μ and momenta π_μ yields Hamilton's equations (4.72), which here are

$$\frac{dx^\mu}{a d\lambda} = p^\mu , \quad \frac{dp_\kappa}{a d\lambda} = \Gamma_{\mu\nu\kappa} p^\mu p^\nu + q F_{\kappa\mu} p^\mu , \quad (4.98)$$

with p_μ defined by

$$p_\mu \equiv \pi_\mu - q A_\mu . \quad (4.99)$$

The condition (4.103) found below, substituted into the first of Hamilton's equations (4.98), implies that p^μ coincides with the usual ordinary momentum $p^\mu = m dx^\mu / d\tau$, as it should. Requiring that the variation (4.97) of the action vanish under arbitrary variation of the parameter a yields the additional equation of motion

$$\frac{\partial H}{\partial a} = 0 . \quad (4.100)$$

The additional equation of motion (4.100) applied to the Hamiltonian (4.96) implies that

$$g^{\mu\nu}(\pi_\mu - qA_\mu)(\pi_\nu - qA_\nu) = -m^2 . \quad (4.101)$$

From the first of the equations of motion (4.98) along with the definition (4.99), equation (4.101) is the same as

$$g_{\mu\nu} \frac{dx^\mu}{a d\lambda} \frac{dx^\nu}{a d\lambda} = -m^2 , \quad (4.102)$$

which in turn is equivalent to

$$a d\lambda = \frac{d\tau}{m} , \quad (4.103)$$

recovering equation (4.35) derived using the Lagrangian formalism. Inserting the condition (4.103) into Hamilton's equations (4.98) recovers the equations of motion (4.42) for a test particle in a prescribed gravitational and electromagnetic field. The value of the Hamiltonian (4.96) after the equation of motion (4.101) is imposed is zero,

$$H = 0 . \quad (4.104)$$

4.15 Derivatives of the action

Besides being a scalar whose minimum value between fixed endpoints defines the path between those points, the action S can also be treated as a function of its endpoints along the actual path. Along the actual path, the equations of motion are satisfied, so the integral in the variation (4.4) or (4.71) of the action vanishes identically. The surface term in the variation (4.4) or (4.71) then implies that $\delta S = \pi_\mu \delta x^\mu$. This means that the partial derivatives of the action with respect to the coordinates are equal to the generalized momenta,

$$\frac{\partial S}{\partial x^\mu} = \pi_\mu . \quad (4.105)$$

This is the basis of the Hamilton-Jacobi method for solving equations of motion, §4.16.

By definition, the total derivative of the action S with respect to the arbitrary parameter λ along the actual path equals the Lagrangian L . In addition to being a function of the coordinates x^μ along the actual path, the action may also be an explicit function $S(\lambda, x^\mu)$ of the parameter λ . The total derivative of the action along the path may thus be expressed

$$\frac{dS}{d\lambda} = L = \frac{\partial S}{\partial \lambda} + \frac{\partial S}{\partial x^\mu} \frac{dx^\mu}{d\lambda} . \quad (4.106)$$

Comparing equation (4.106) to the definition (4.68) of the Hamiltonian shows that the partial derivative of the action with respect to the parameter λ is minus the Hamiltonian

$$\frac{\partial S}{\partial \lambda} = -H . \quad (4.107)$$

In the conventional approach where the parameter λ is fixed equal to the time coordinate t , equations (4.105) and (4.107) together show that

$$\frac{\partial S}{\partial t} = \pi_t = -H , \quad (4.108)$$

in agreement with equation (4.87). In the super-Hamiltonian approach, the Hamiltonian H is constant, equal to $-m^2/2$ in the effective approach, equation (4.93), and equal to zero in the nice approach, equation (4.104).

Concept question 4.5 Action vanishes along a null geodesic, but its gradient does not. How can it be that the gradient of the action $p_\mu = \partial S / \partial x^\mu$ is non-zero along a null geodesic, yet the variation of the action $dS = -m d\tau$ is identically zero along the same null geodesic? **Answer.** This has to do with the fact that a vector can be finite yet null,

$$\frac{dS}{d\lambda} = \frac{dx^\mu}{d\lambda} \frac{\partial S}{\partial x^\mu} = \pi^\mu \pi_\mu = -m^2 = 0 \quad \text{for } m = 0 . \quad (4.109)$$

4.16 Hamilton-Jacobi equation

The Hamilton-Jacobi equation provides a powerful way to solve equations of motion. The Hamilton-Jacobi equation proves to be separable in the Kerr-Newman geometry for an ideal rotating black hole, Chapter 23. The hypothesis that the Hamilton-Jacobi equation be separable provides one way to derive the Kerr-Newman line-element, Chapter 22, and to discover other separable spacetimes.

The Hamilton-Jacobi equation is obtained by writing down the expression for the Hamiltonian H in terms of coordinates x^μ and generalized momenta π_μ , and replacing the Hamiltonian H by $-\partial S / d\lambda$ in accordance with equation (4.107), and the generalized momenta π_μ by $\partial S / \partial x^\mu$ in accordance with equation (4.105).

For the effective Hamiltonian (4.90), the resulting Hamilton-Jacobi equation is

$$H = -\frac{\partial S}{\partial \lambda} = \frac{1}{2} g^{\mu\nu} \left(\frac{\partial S}{\partial x^\mu} - q A_\mu \right) \left(\frac{\partial S}{\partial x^\nu} - q A_\nu \right) = -\frac{1}{2} m^2 , \quad (4.110)$$

the final expression, $-\frac{1}{2} m^2$, coming from equation (4.93). For the nice Hamiltonian (4.96), the resulting Hamilton-Jacobi equation is

$$\frac{H}{a} = -\frac{\partial S}{a \partial \lambda} = \frac{1}{2} \left[g^{\mu\nu} \left(\frac{\partial S}{\partial x^\mu} - q A_\mu \right) \left(\frac{\partial S}{\partial x^\nu} - q A_\nu \right) + m^2 \right] = 0 , \quad (4.111)$$

the final expression, 0, coming from equation (4.104). The Hamilton-Jacobi equations (4.110) and (4.111) agree, as they should. The Hamilton-Jacobi equation (4.110) or (4.111) is a partial differential equation for the action $S(\lambda, x^\mu)$. In spacetimes with sufficient symmetry, such as Kerr-Newman, the partial differential equation can be solved by separation of variables. This will be done in §22.3.

4.17 Canonical transformations

The Lagrangian equations of motion (4.5) take the same form regardless of the choice of coordinates x^μ of the underlying spacetime. This expresses general covariance: the form of the Lagrangian equations of motion is unchanged by general coordinate transformations.

Coordinate transformations also preserve Hamilton's equations of motion (4.72). But the Hamiltonian formalism allows a wider range of transformations that preserve the form of Hamilton's equations. Transformations of the coordinates and momenta that preserve Hamilton's equations are called **canonical transformations**. The construction of canonical transformations is addressed in §4.17.1.

The wide range of possible canonical transformations means that the coordinates and momenta lose much of their original meaning as actual spacetime coordinates and momenta of particles. For example, there is a canonical transformation (4.117) that simply exchanges coordinates and their conjugate momenta. It is common therefore to refer to general systems of coordinates and momenta that satisfy Hamilton's equations as **generalized coordinates** and **generalized momenta**, and to denote them by q^μ and p_μ ,

$$q^\mu, \quad p_\mu. \quad (4.112)$$

4.17.1 Construction of canonical transformations

Consider a canonical transformation of coordinates and momenta

$$\{q^\mu, p_\mu\} \rightarrow \{q'^\mu(q, p), p'_\mu(q, p)\}. \quad (4.113)$$

By definition of canonical transformation, both the original and transformed sets of coordinates and momenta satisfy Hamilton's equations.

For the equations of motion to take Hamiltonian form, the original and transformed actions S and S' must take the form

$$S = \int_{\lambda_i}^{\lambda_f} p_\mu dq^\mu - H d\lambda, \quad S' = \int_{\lambda_i}^{\lambda_f} p'_\mu dq'^\mu - H' d\lambda. \quad (4.114)$$

As remarked following equation (4.68), the indices on the momentum term can be flipped if desired, $\pi_\mu dx^\mu = \pi^\mu dx_\mu$. One way for the original and transformed coordinates and momenta to yield equivalent equations of motion is that the integrands of the actions differ by the total derivative dF of some function F ,

$$dF = p_\mu dq^\mu - p'_\mu dq'^\mu - (H - H') d\lambda. \quad (4.115)$$

When the actions S and S' are varied, the difference in the variations is the difference in the variation of F between the initial and final points λ_i and λ_f , which vanishes provided that whatever F depends on is held fixed on the initial and final points,

$$\delta S - \delta S' = [\delta F]_{\lambda_i}^{\lambda_f} = 0. \quad (4.116)$$

Because the variations of the actions are the same, the resulting equations of motion are equivalent. The function F is called the **generator** of the canonical transformation between the original and transformed coordinates.

Given any function $F(\lambda, q, q')$, equation (4.115) determines p_μ , $-p'_\mu$, and $H - H'$ as partial derivatives of F with respect to q^μ , q'^μ , and λ . For example, the function $F = \sum_\mu q^\mu q'_\mu$ generates a canonical transformation that simply trades coordinates and momenta,

$$p_\mu = \frac{\partial F}{\partial q^\mu} = q'_\mu, \quad p'^\mu = -\frac{\partial F}{\partial q'_\mu} = -q^\mu. \quad (4.117)$$

The generating function $F(\lambda, q, q')$ depends on q^μ and q'_μ . Other generating functions depending on either of q^μ or p_μ , and either of q'^μ or p'_μ , are obtained by subtracting $p_\mu q^\mu$ and/or adding $p'_\mu q'^\mu$ to F . For example, equation (4.115) can be rearranged as

$$dG = p_\mu dq^\mu + q'^\mu dp'_\mu - (H - H')d\lambda, \quad (4.118)$$

where $G \equiv F + p'_\nu q'^\nu$ is now some function $G(\lambda, q, p')$. For example, the function $G(q, p') = \sum_\mu f^\mu(q) p'_\mu$, in which $f^\mu(q)$ is some function of the coordinates q^ν but not of the momenta p_ν , generates the canonical transformation

$$p_\mu = \frac{\partial G}{\partial q^\mu} = \sum_\nu \frac{\partial f^\nu}{\partial q^\mu} p'_\nu, \quad q'^\mu = \frac{\partial G}{\partial p'_\mu} = f^\mu(q). \quad (4.119)$$

This is just a coordinate transformation $q^\mu \rightarrow q'^\mu = f^\mu(q)$.

If the generator of a canonical transformation does not depend on the parameter λ , then the Hamiltonians are the same in the original and transformed systems,

$$H(q^\mu, p_\mu) = H'(q'^\mu, p'_\mu). \quad (4.120)$$

In the super-Hamiltonian approach, where the parameter λ is arbitrary, the Hamiltonian is without loss of generality independent of λ , and there is no physical significance to canonical transformations generated by functions that depend on λ . The super-Hamiltonian $H(q^\mu, p_\mu)$ is then a scalar, invariant with respect to canonical transformations that do not depend explicitly on λ . This contrasts with the conventional Hamiltonian approach, where the parameter λ is set equal to the coordinate time t , and the conventional Hamiltonian is the time component of a 4-vector, which varies under canonical transformations generated by functions that depend on time t .

Exercise 4.6 Canonical transformations from a quadratic generating function. For simplicity, focus on a single coordinate q and its conjugate momentum p . Consider a quadratic generating function of the form

$$F(q, q') = \frac{1}{2} A q^2 + B q q' + \frac{1}{2} C q'^2 \quad (4.121)$$

with constant coefficients A, B, C . Show that F generates a linear transformation

$$\begin{pmatrix} q' \\ p' \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}, \quad (4.122)$$

in which the transformation matrix has unit determinant,

$$\alpha\delta - \beta\gamma = 1. \quad (4.123)$$

Solution. The partial derivatives of $F(q, q')$ with respect to q and q' define the momenta p and p' in accordance with equation (4.115),

$$p = \frac{\partial F}{\partial q} = Aq + Bq' , \quad p' = -\frac{\partial F}{\partial q'} = -Bq - Cq' , \quad (4.124)$$

which yield expressions for q' and p' in terms of q and p ,

$$\begin{pmatrix} q' \\ p' \end{pmatrix} = \frac{1}{B} \begin{pmatrix} -A & 1 \\ AC - B^2 & -C \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} , \quad (4.125)$$

whose matrix has determinant 1. The generating function (4.121) can be expressed in terms of α , β , and δ as

$$F(q, q') = \frac{1}{\beta} \left(-\frac{1}{2} \alpha q^2 + qq' - \frac{1}{2} \delta q'^2 \right) . \quad (4.126)$$

4.17.2 Evolution is a canonical transformation

The evolution of the system from some initial hypersurface $\lambda = 0$ to some final hypersurface λ is itself a canonical transformation. This is evident from the fact that Hamilton's equations (4.72) hold for any value of the parameter λ , so in particular Hamilton's equations are unchanged when initial coordinates and momenta $q^\mu(0)$ and $p_\mu(0)$ are replaced by evolved values $q^\mu(\lambda)$ and $p_\mu(\lambda)$,

$$q^\mu(0) \rightarrow q'^\mu = q^\mu(\lambda) , \quad p_\mu(0) \rightarrow p'_\mu = p_\mu(\lambda) . \quad (4.127)$$

The action varies by the total derivative $dS = p_\mu dq^\mu - H d\lambda$ along the actual path of the system, equation (4.106), so the initial and evolved actions differ by a total derivative, equation (4.115),

$$dF = p_\mu(0) dq^\mu(0) - p_\mu(\lambda) dq^\mu(\lambda) - [H(0) - H(\lambda)] d\lambda = dS(0) - dS(\lambda) . \quad (4.128)$$

Thus the canonical transformation from an initial $\lambda = 0$ to a final λ is generated by the difference in the actions along the actual path of the system,

$$F = S(0) - S(\lambda) . \quad (4.129)$$

4.18 Symplectic structure

The generalized coordinates q^μ and momenta p_μ of a dynamical system of particles have a geometrical structure that transcends the geometrical structure of the underlying spacetime manifold. For N coordinates q^μ and N momenta p_μ , the geometrical structure is a $2N$ -dimensional manifold called a **symplectic manifold**. A symplectic manifold is also called **phase space**, and the coordinates $\{q^\mu, p_\mu\}$ of the manifold are called phase-space coordinates.

A central property of a symplectic manifold is that the Hamiltonian dynamics define a scalar product with

antisymmetric **symplectic metric** ω_{ij} . Let z^i with $i = 1, \dots, 2N$ denote the combined set of $2N$ generalized coordinates and momenta $\{q^\mu, p_\mu\}$,

$$\{z^1, \dots, z^N, z^{N+1}, \dots, z^{2N}\} \equiv \{q^1, \dots, q^N, p_1, \dots, p_N\} . \quad (4.130)$$

Hamilton's equations (4.72) can be written

$$\frac{dz^i}{d\lambda} = \omega^{ij} \frac{\partial H}{\partial z^j} , \quad (4.131)$$

where ω^{ij} is the antisymmetric symplectic metric (actually the inverse symplectic metric)

$$\omega^{ij} \equiv \delta_{i+N, j} - \delta_{i, j+N} = \begin{cases} 1 & \text{if } z^i = q^\mu \text{ and } z^j = p_\mu , \\ -1 & \text{if } z^i = p_\mu \text{ and } z^j = q^\mu , \\ 0 & \text{otherwise .} \end{cases} \quad (4.132)$$

As a matrix, the symplectic metric ω^{ij} is the $2N \times 2N$ matrix

$$\omega^{ij} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} , \quad (4.133)$$

where 1 denotes the $N \times N$ unit matrix. Inverting the inverse symplectic metric ω^{ij} yields the symplectic metric ω_{ij} , which is the same matrix but flipped in sign,

$$\omega_{ij} \equiv (\omega^{ij})^{-1} = (\omega^{ij})^\top = -\omega^{ij} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} . \quad (4.134)$$

Let z'^i be another set of generalized coordinates and momenta satisfying Hamilton's equations with the same Hamiltonian H ,

$$\frac{dz'^i}{d\lambda} = \omega^{ij} \frac{\partial H}{\partial z'^j} . \quad (4.135)$$

It is being assumed here that the Hamiltonian H does not depend explicitly on the parameter λ . In the super-Hamiltonian approach, there is no loss of generality in taking the Hamiltonian H to be independent of λ , since the parameter λ is arbitrary, without physical significance. The important point about equation (4.135) is that the symplectic metric ω^{ij} is the same regardless of the choice of phase-space coordinates. Under a canonical transformation $z^i \rightarrow z'^i(z)$ of generalized coordinates and momenta, $dz'^i/d\lambda$ transforms as

$$\frac{dz'^i}{d\lambda} = \frac{\partial z'^i}{\partial z^k} \frac{dz^k}{d\lambda} = \frac{\partial z'^i}{\partial z^k} \omega^{kl} \frac{\partial H}{\partial z^l} = \frac{\partial z'^i}{\partial z^k} \omega^{kl} \frac{\partial z'^j}{\partial z^l} \frac{\partial H}{\partial z'^j} . \quad (4.136)$$

Comparing equations (4.135) and (4.136) shows that the symplectic matrix ω^{ij} is invariant under a canonical transformation,

$$\omega^{ij} = \frac{\partial z'^i}{\partial z^k} \omega^{kl} \frac{\partial z'^j}{\partial z^l} . \quad (4.137)$$

Equation (4.137) can be expressed as the invariance under canonical transformations of

$$\omega^{ij} \frac{\partial}{\partial z^i} \frac{\partial}{\partial z^j} = \omega^{ij} \frac{\partial}{\partial z'^i} \frac{\partial}{\partial z'^j} . \quad (4.138)$$

Equivalently,

$$\omega_{ij} dz^i dz^j = \omega_{ij} dz'^i dz'^j . \quad (4.139)$$

The invariance of the symplectic metric ω_{ij} under canonical transformations can be thought of as analogous to the invariance of the Minkowski metric η_{mn} under Lorentz transformations. But whereas the Minkowski metric η_{mn} is symmetric, the symplectic metric ω_{ij} is antisymmetric.

4.19 Symplectic scalar product and Poisson brackets

Let $f(z^i)$ and $g(z^i)$ be two functions of phase-space coordinates z^i . Their tangent vectors in the phase space are $\partial f / \partial z^i$ and $\partial g / \partial z^i$. The symplectic scalar product of the tangent vectors defines the **Poisson bracket** of the two functions f and g ,

$$[f, g]_P \equiv \omega^{ij} \frac{\partial f}{\partial z^i} \frac{\partial g}{\partial z^j} = \frac{\partial f}{\partial q^\mu} \frac{\partial g}{\partial p_\mu} - \frac{\partial f}{\partial p_\mu} \frac{\partial g}{\partial q^\mu} . \quad (4.140)$$

The invariance (4.138) of the symplectic metric implies that the Poisson bracket is a scalar, invariant under canonical transformations of the phase-space coordinates z^i . The Poisson bracket is antisymmetric thanks to the antisymmetry of the symplectic metric ω^{ij} ,

$$[f, g]_P = -[g, f]_P . \quad (4.141)$$

4.19.1 Poisson brackets of phase-space coordinates

The Poisson brackets of the phase-space coordinates and momenta themselves satisfy

$$[z^i, z^j]_P = \omega^{ij} . \quad (4.142)$$

Explicitly in terms of the generalized coordinates and momenta q^μ and p_μ ,

$$[q^\mu, p_\nu]_P = \delta^\mu_\nu , \quad [q^\mu, q^\nu]_P = 0 , \quad [p_\mu, p_\nu]_P = 0 . \quad (4.143)$$

Reinterpreting equations (4.143) as operator equations provides a path from classical to quantum mechanics. Experiments such as the two-slit experiment show that a fundamental particle such as a photon or an electron behaves like a particle when you look at it, but like a wave when you don't look at it. The particle's momentum p_μ , a particle-like property, proves to be equal to the reduced Planck's constant $\hbar$ multiplied by the particle's wavevector k_μ , a wave-like property, $p_\mu = \hbar k_\mu$, a relation that is exact, not uncertain. Quantum mechanics reinterprets coordinates and momenta as operators $\hat{q}^\mu$ and $\hat{p}_\mu$ that, acting

on eigenstates, yield definite values, the eigenvalue of the eigestate. The momentum is interpreted as an operator $\hat{p}_\mu = -i\hbar \partial/\partial q^\mu$ which, acting on a plane-wave eigenstate $\Phi \propto e^{ik_\mu q^\mu} = e^{(i/\hbar)p_\mu q^\mu}$, yields its measured momentum, $\hat{p}_\mu \Phi = p_\mu \Phi$. The coordinate and momentum operators of quantum mechanics satisfy

$$[\hat{q}^\mu, \hat{p}_\nu] = i\hbar \delta_\nu^\mu, \quad [\hat{q}^\mu, \hat{q}^\nu] = 0, \quad [\hat{p}_\mu, \hat{p}_\nu] = 0, \quad (4.144)$$

which look like the Poisson brackets equations (4.143) with an extra factor of $i\hbar$.

Notwithstanding the resemblance between Poisson brackets and quantum-mechanical commutators, the two should not be conflated. Classical coordinates and their conjugate momenta commute, $[q^\mu, p_\nu] = 0$, and functions $f(q, p)$ and $g(q, p)$ of classical coordinates and their momenta likewise commute, $[f, g] = 0$.

Groenewold (1946) and Moyal (1949) advocated that the Poisson bracket between two functions f and g of phase-space coordinates z^i be generalized to the Moyal bracket

$$[f, g]_M \equiv f \star g - g \star f, \quad (4.145)$$

where the Moyal star-product is defined by a complex exponential of the Poisson bracket,

$$f \star g \equiv f \exp \left(\frac{i\hbar}{2} \overleftarrow{\partial} \omega^{ij} \frac{\overrightarrow{\partial}}{\partial z^j} \right) g = fg + \frac{i\hbar}{2} \omega^{ij} \frac{\partial f}{\partial z^i} \frac{\partial g}{\partial z^j} + \frac{1}{2!} \left(\frac{i\hbar}{2} \right)^2 \omega^{ij} \omega^{kl} \frac{\partial^2 f}{\partial z^i \partial z^k} \frac{\partial^2 g}{\partial z^j \partial z^l} + \dots, \quad (4.146)$$

whose leading term is the classical product fg . The Moyal bracket (4.145) equals twice the odd (in $\hbar$) part of the product (4.146). Its leading term, linear in $\hbar$, equals the Poisson bracket. The Moyal bracket offers an entry to the mathematical discipline of noncommutative geometry. See Hiley (2021) for an assessment.

4.20 (Super-)Hamiltonian as a generator of evolution

The Poisson bracket of a function $f(z^i)$ with the Hamiltonian H is

$$[f, H]_P = \frac{\partial f}{\partial q^\mu} \frac{\partial H}{\partial p_\mu} - \frac{\partial f}{\partial p_\mu} \frac{\partial H}{\partial q^\mu}. \quad (4.147)$$

Inserting Hamilton's equations (4.72) implies

$$[f, H]_P = \frac{\partial f}{\partial q^\mu} \frac{dq^\mu}{d\lambda} + \frac{\partial f}{\partial p_\mu} \frac{dp_\mu}{d\lambda} = \frac{df}{d\lambda}. \quad (4.148)$$

That is, the evolution of a function $f(q^\mu, p_\mu)$ of generalized coordinates and momenta is its Poisson bracket with the Hamiltonian H ,

$$\frac{df}{d\lambda} = [f, H]_P. \quad (4.149)$$

Equation (4.149) shows that the (super-)Hamiltonian defined by equation (4.68) can be interpreted as generating the evolution of the system.

The same derivation holds in the conventional case where λ is taken to be time t , but generically the function $f(t, q^\alpha, p_\alpha)$ and conventional Hamiltonian $H(t, q^\alpha, p_\alpha)$ must be allowed to be explicit functions of

time t as well as of generalized spatial coordinates and momenta q^α and p_α . Equation (4.149) becomes in the conventional case

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + [f, H]_P . \quad (4.150)$$

4.21 Infinitesimal canonical transformations

A canonical transformation generated by $G = q^\mu p'_\mu$ is the identity transformation, since it leaves the coordinates and momenta unchanged. Consider a canonical transformation with generator infinitesimally shifted from the identity transformation, with ϵ an infinitesimal parameter,

$$G = q^\mu p'_\mu + \epsilon g(q, p') . \quad (4.151)$$

The resulting canonical transformation is, from equation (4.119),

$$q'^\mu = \frac{\partial G}{\partial p'_\mu} = q^\mu + \epsilon \frac{\partial g}{\partial p'_\mu} , \quad p'_\mu = \frac{\partial G}{\partial q^\mu} = p'_\mu + \epsilon \frac{\partial g}{\partial q^\mu} . \quad (4.152)$$

Because ϵ is infinitesimal, the term $\epsilon \partial g / \partial p'_\mu$ can be replaced by $\epsilon \partial g / \partial p_\mu$ to linear order, yielding

$$q'^\mu = q^\mu + \epsilon \frac{\partial g}{\partial p_\mu} , \quad p'_\mu = p_\mu - \epsilon \frac{\partial g}{\partial q^\mu} . \quad (4.153)$$

Equations (4.153) imply that the changes δp_μ and δq^μ in the coordinates and momenta under an infinitesimal canonical transformation (4.151) is their Poisson bracket with g ,

$$\delta p_\mu = \epsilon [p_\mu, g]_P , \quad \delta q^\mu = \epsilon [q^\mu, g]_P . \quad (4.154)$$

As a particular example, the evolution of the system under an infinitesimal change $\delta\lambda$ in the parameter λ is, in accordance with the evolutionary equation (4.149), generated by a canonical transformation with g in equation (4.151) set equal to the Hamiltonian H ,

$$\delta p_\mu = \delta\lambda [p_\mu, H]_P , \quad \delta q^\mu = \delta\lambda [q^\mu, H]_P . \quad (4.155)$$

4.22 Constancy of phase-space volume under canonical transformations

The invariance of the symplectic metric under canonical transformations implies the invariance of phase-space volume under canonical transformations.

The volume V of a region of $2N$ -dimensional phase space is

$$V \equiv \int dV \equiv \int dz^1 \dots dz^{2N} \equiv \int dq^1 \dots dq^N dp_1 \dots dp_N , \quad (4.156)$$

integrated over the region. Under a canonical transformation $z^i \rightarrow z'^i(z)$ of phase-space coordinates, the

phase-space volume element dV transforms by the Jacobian of the transformation, which is the determinant $|\partial z'^i / \partial z^j|$,

$$dV' = \left| \frac{\partial z'^i}{\partial z^j} \right| dV . \quad (4.157)$$

But equation (4.137) implies that

$$|\omega^{ij}| = \left| \frac{\partial z'^i}{\partial z^k} \right| |\omega^{kl}| \left| \frac{\partial z'^j}{\partial z^l} \right| , \quad (4.158)$$

so the Jacobian must be 1 in absolute magnitude,

$$\left| \frac{\partial z'^i}{\partial z^j} \right| = \pm 1 . \quad (4.159)$$

If the canonical transformation can be obtained by a continuous transformation from the identity, then the Jacobian must equal 1. As a particular case, the Jacobian equals 1 for the canonical transformation generated by evolution, §4.22.1, since evolution is continuous from initial to final conditions.

4.22.1 Constancy of phase-space volume under evolution

Since evolution is a canonical transformation, §4.17.2 and §4.21, phase-space volume V is preserved under evolution of the system. Each phase-space point inside the volume V evolves according to the equations of motion. As the system of points evolves, the region distorts, but the magnitude of the volume V of the region remains constant. The constancy of phase-space volume as it evolves was proved explicitly in 1871 by Boltzmann, who later referred to the result as “Liouville’s theorem” since the proof was based in part on a mathematical theorem proved by Liouville (see Nolte, 2010).

4.23 Poisson algebra of integrals of motion

A function $f(z^i)$ of the generalized coordinates and momenta is said to be an integral of motion if it is constant as the system evolves. In view of equation (4.149), a function $f(z^i)$ is an integral of motion if and only if its Poisson brackets with the Hamiltonian vanishes,

$$[f, H]_P = 0 . \quad (4.160)$$

As a particular example, the antisymmetry of the Poisson bracket implies that the Poisson bracket of the Hamiltonian with itself is zero,

$$[H, H]_P = 0 , \quad (4.161)$$

so the Hamiltonian H is itself a constant of motion. The super-Hamiltonian H is a constant of motion in general, while the conventional Hamiltonian H is constant provided that it does not depend explicitly on time t .

Suppose that $f(z^i)$ and $g(z^i)$ are both integrals of motion. Then their Poisson brackets with each other is also an integral of motion,

$$[[f, g]_P, H]_P = - [[g, H]_P, f]_P - [[H, f]_P, g]_P = 0, \quad (4.162)$$

the first equality of which expresses the Jacobi identity, and the last equality of which follows because the Poisson bracket of each of f and g with the Hamiltonian H vanishes. The Poisson bracket of two integrals of motion f and g may or may not yield a further distinct integral of motion. A set of linearly independent integrals of motion whose Poisson brackets close forms a Lie algebra is called a **Poisson algebra**.

Concept question 4.7 How many integrals of motion can there be? How many distinct integrals of motion can there be in a dynamical system described by N coordinates and N momenta? A distinct integral of motion is one that cannot be expressed as a function of the other integrals of motion (this is more stringent than the condition that the integrals be linearly independent). **Answer.** The dynamical motion of the system is described by a 1-dimensional line in a $2N$ -dimensional phase-space manifold consisting of the N coordinates and N momenta. Any constant of motion $f(x^\mu, \pi_\mu)$ defines a $(2N-1)$ -dimensional submanifold of the phase-space manifold. A 1-dimensional line can be the intersection of no more than $2N-1$ distinct such submanifolds, so there can be at most $2N-1$ distinct constants of motion. In the super-Hamiltonian formulation, the phase space of a single particle in 4 spacetime dimensions is 8-dimensional, and there are at most 7 distinct integrals of motion. A particle moving along a straight line in Minkowski space provides an example of a system with a full set of 7 integrals of motion: 4 integrals constitute the covariant energy-momentum 4-vector p_m , and a further 3 integrals of motion comprise $x^a - v^a t = x^a(0)$ where $v^a \equiv p^a/p^0$ is the velocity, and $x^m(0)$ is the origin of the line at $t = 0$. In the conventional Hamiltonian formulation, the phase space of a single particle is 6-dimensional, and there are at most 5 distinct integrals of motion. The apparent discrepancy in the number of integrals occurs because in the super-Hamiltonian formalism the time t and time component π_t of the generalized momentum are treated as distinct variables whose equations of motion are determined by Hamilton's equations, whereas in the conventional Hamiltonian formalism the arbitrary parameter λ is set equal to the time t , which is therefore no longer an independent variable, and the generalized momentum π_t , which equals minus the conventional Hamiltonian H , equation (4.108), is eliminated as an independent variable by re-expressing it in terms of the spatial coordinates and momenta.

Concept Questions

1. What evidence do astronomers currently accept as indicating the presence of a black hole in a system?
2. Why can astronomers measure the masses of supermassive black holes only in relatively nearby galaxies?
3. To what extent (with what accuracy) are real black holes in our Universe described by the no-hair theorem?
4. Does the no-hair theorem apply inside a black hole?
5. Black holes lose their hair on a light-crossing time. How long is a light-crossing time for a typical stellar-sized or supermassive astronomical black hole?
6. Relativists say that the metric is $g_{\mu\nu}$, but they also say that the metric is $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$. How can both statements be correct?
7. The Schwarzschild geometry is said to describe the geometry of spacetime outside the surface of the Sun or Earth. But the Schwarzschild geometry supposedly describes non-rotating masses, whereas the Sun and Earth are rotating. If the Sun or Earth collapsed to a black hole conserving their mass M and angular momentum L , roughly what would the spin $a/M = L/M^2$ of the black hole be relative to the maximal spin $a/M = 1$ of a Kerr black hole?
8. What happens at the horizon of a black hole?
9. As cold matter becomes denser, it goes through the stages of being solid/liquid like a planet, then electron degenerate like a white dwarf, then neutron degenerate like a neutron star, then finally it collapses to a black hole. Why could there not be a denser state of matter, denser than a neutron star, that brings a star to rest inside its horizon?
10. How can an observer determine whether they are “at rest” in the Schwarzschild geometry?
11. An observer outside the horizon of a black hole never sees anything pass through the horizon, even to the end of the Universe. Does the black hole then ever actually collapse, if no one ever sees it do so?
12. If nothing can ever get out of a black hole, how does its gravity get out?
13. Why did Einstein believe that black holes could not exist in nature?
14. In what sense is a rotating black hole “stationary” but not “static”?
15. What is a white hole? Do they exist?
16. Could the expanding Universe be a white hole?
17. Could the Universe be the interior of a black hole?

18. You know the Schwarzschild metric for a black hole. What is the corresponding metric for a white hole?
19. What is the best kind of black hole to fall into if you want to avoid being tidally torn apart?
20. Why do astronomers often assume that the inner edge of an accretion disk around a black hole occurs at the innermost stable orbit?
21. A collapsing star of uniform density has the geometry of a collapsing Friedmann-Lemaître-Robertson-Walker cosmology. If a spatially flat FLRW cosmology corresponds to a star that starts from zero velocity at infinity, then to what do open or closed FLRW cosmologies correspond?
22. Your friend falls into a black hole, and you watch her image freeze and redshift at the horizon. A shell of matter falls on to the black hole, increasing the mass of the black hole. What happens to the image of your friend? Does it disappear, or does it remain on the horizon?
23. Is the singularity of a Reissner-Nordström black hole gravitationally attractive or repulsive?
24. If you are a charged particle, which dominates near the singularity of the Reissner-Nordström geometry, the electrical attraction/repulsion or the gravitational attraction/repulsion?
25. Is a white hole gravitationally attractive or repulsive?
26. What happens if you fall into a white hole?
27. Which way does time go in Parallel Universes in the Reissner-Nordström geometry?
28. What does it mean that geodesics inside a black hole can have negative energy?
29. Can geodesics have negative energy outside a black hole? How about inside the ergosphere?
30. Physically, what causes mass inflation?
31. Is mass inflation likely to occur inside real astronomical black holes?
32. What happens at the X point, where the outgoing and ingoing inner horizons of the Reissner-Nordström geometry intersect?
33. Can a particle like an electron or proton, whose charge far exceeds its mass (in geometric units), be modelled as Reissner-Nordström black hole?
34. Does it make sense that a person might be at rest in the Kerr-Newman geometry? How would the Boyer-Lindquist coordinates of such a person vary along their worldline?
35. In identifying M as the mass and a the angular momentum per unit mass of the black hole in the Boyer-Lindquist metric, why is it sufficient to consider the behaviour of the metric at $r \rightarrow \infty$?
36. Does space move faster than light inside the ergosphere?
37. If space moves faster than light inside the ergosphere, why is the outer boundary of the ergosphere not a horizon?
38. Do closed timelike curves make sense?
39. What does Carter's fourth integral of motion $\mathcal{Q}$ signify physically?
40. What is special about a principal null congruence?
41. Evaluated in the locally inertial frame of a principal null congruence, the spin-0 component of the Weyl scalar of the Kerr geometry is $C = -M/(r - ia \cos \theta)^3$, which looks like the Weyl scalar $C = -M/r^3$ of the Schwarzschild geometry but with radius r replaced by the complex radius $r - ia \cos \theta$. Is there something deep here? Can the Kerr geometry be constructed from the Schwarzschild geometry by complexifying the radial coordinate r ?

What's important?

1. Astronomical evidence suggests that stellar-sized and supermassive black holes exist ubiquitously in nature.
2. The no-hair theorem, and when and why it applies.
3. The physical picture of black holes as regions of spacetime where space is falling faster than light.
4. A physical understanding of how the metric of a black hole relates to its physical properties.
5. Penrose (conformal) diagrams. In particular, the Penrose diagrams of the various kinds of vacuum black hole: Schwarzschild, Reissner-Nordström, Kerr-Newman.
6. What really happens inside black holes. Collapse of a star. Mass inflation instability.

Observational Evidence for Black Holes

It is beyond the intended scope of this book to discuss the extensive and rapidly evolving observational evidence for black holes in any detail. However, it is useful to summarize a few facts.

1. Observational evidence supports the idea that black holes occur ubiquitously in nature. They are not observed directly, but reveal themselves through their effects on their surroundings. Two kinds of black hole are observed: stellar-sized black holes in x-ray binary systems, mostly in our own Milky Way galaxy, and supermassive black holes in Active Galactic Nuclei (AGN) found at the centres of our own and other galaxies.
2. The primary evidence that astronomers accept as indicating the presence of a black hole is a lot of mass compacted into a tiny space.
 - a. In an x-ray binary system, if the mass of the compact object exceeds $3 M_{\odot}$, the maximum theoretical mass of a neutron star, then the object is considered to be a black hole. Many hundreds of x-ray binary systems are known in our Milky Way galaxy, but only tens of these have measured masses, and in about 20 the measured mass indicates a black hole (McClintock et al., 2011).
 - b. Several tens of thousands of AGN have been catalogued, identified either in the radio, optical, or x-rays. But only in nearby galaxies can the mass of a supermassive black hole be measured directly. This is because it is only in nearby galaxies that the velocities of gas or stars can be measured sufficiently close to the nuclear centre to distinguish a regime where the velocity becomes constant, so that the mass can be attribute to an unresolved central point as opposed to a continuous distribution of stars. The masses of about 40 supermassive black holes have been measured in this way (Kormendy and Gebhardt, 2001). The masses range from the $4 \times 10^6 M_{\odot}$ mass of the black hole at the centre of the Milky Way (Ghez et al., 2008; Gillessen et al., 2009) to the $6.6 \pm 0.4 \times 10^9 M_{\odot}$ mass of the black hole at the centre of the M87 galaxy at the centre of the Virgo cluster at the centre of the Local Supercluster of galaxies (Gebhardt et al., 2011; Akiyama et al., 2019).
3. Secondary evidences for the presence of a black hole are:
 - a. high luminosity;
 - b. non-stellar spectrum, extending from radio to gamma-rays;
 - c. rapid variability.
 - d. relativistic jets.

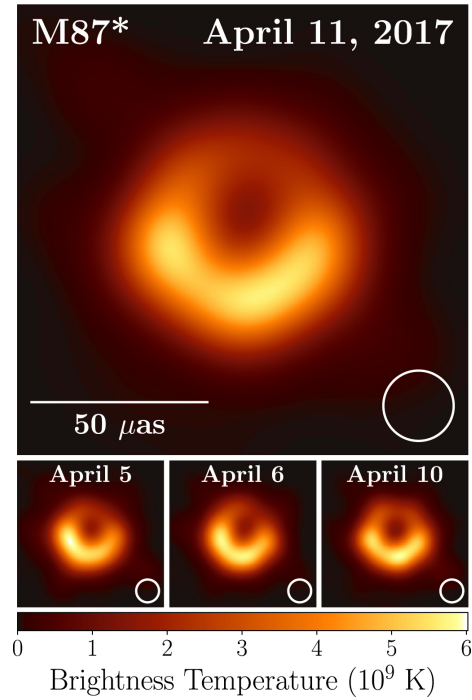


Figure 5.1 The supermassive black hole in the M87 galaxy imaged by the Event Horizon Telescope (Akiyama et al., 2019).

Jets in AGN are often one-sided, and a few that are bright enough to be resolved at high angular resolution show superluminal motion. Both evidences indicate that jets are commonly relativistic, moving at close to the speed of light. There are a few cases of jets in x-ray binary systems, sometimes called microquasars.

4. Stellar-sized black holes are thought to be created in supernovae as the result of the core-collapse of stars more massive than about $25 M_{\odot}$ (this number depends in part on uncertain computer simulations). Supermassive black holes are probably created initially in the same way, but they then grow by accretion of gas funnelled to the centre of the galaxy. The growth rates inferred from AGN luminosities are consistent with this picture.
5. Long gamma-ray bursts (lasting more than about 2 seconds) are associated observationally with supernovae. It is thought that in such bursts we are seeing the formation of a black hole. As the black hole gulps down the huge quantity of material needed to make it, it regurgitates a relativistic jet that punches through the envelope of the star. If the jet happens to be pointed in our direction, then we see it relativistically beamed as a gamma-ray burst.
6. Astronomical black holes present the only realistic prospect for testing general relativity in the strong

field regime, since such fields cannot be reproduced in the laboratory. At the present time the observational tests of general relativity from astronomical black holes are at best tentative. One test is the redshifting of 7 keV iron lines in a small number of AGN, notably MCG-6-30-15, which can be interpreted as being emitted by matter falling on to a rotating (Kerr) black hole.

7. The first direct detection of gravitational waves was with the Laser Interferometer Gravitational wave Observatory (LIGO) on 14 September 2015 (Abbott et al., 2016). The wave-form was consistent with the merger of two black holes of masses 29 and 36 $M_{\odot}$.
8. Before gravitational waves were detected directly, their existence was inferred from the gradual speeding up of the orbit of the Hulse-Taylor binary, which consists of two neutron stars, one of which, PSR1913+16, is a pulsar. The parameters of the orbit have been measured with exquisite precision, and the rate of orbital speed-up is in good agreement with the energy loss by quadrupole gravitational wave emission predicted by general relativity.

6

Ideal Black Holes

6.1 Definition of a black hole

What is a black hole? Doubtless you have heard the standard definition: It is a region whose gravity is so strong that not even light can escape.

But why can light not escape from a black hole? A standard answer, which John Michell (1784) would have found familiar, is that the escape velocity exceeds the speed of light. But that answer brings to mind a Newtonian picture of light going up, turning around, and coming back down, that is altogether different from what general relativity actually predicts.

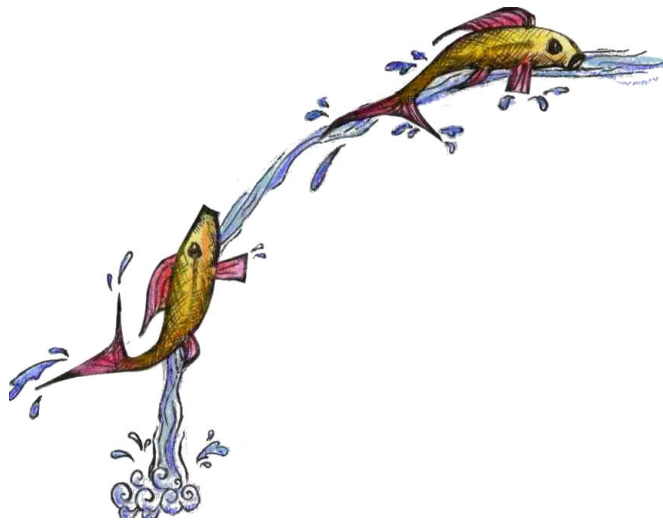


Figure 6.1 The fish upstream can make way against the current, but the fish downstream is swept to the bottom of the waterfall (Art by Wildrose Hamilton). This painting appeared on the cover of the June 2008 issue of the American Journal of Physics (Hamilton and Lisle, 2008). A similar depiction appeared in Susskind (2003).

A better definition of a black hole is that it is a

region where space is falling faster than light.

Inside the horizon, light emitted outwards is carried inward by the faster-than-light inflow of space, like a fish trying but failing to swim up a waterfall, Figure 6.1.

The definition may seem jarring. If space has no substance, how can it fall faster than light? It means that inside the horizon any locally inertial frame is compelled to fall to smaller radius as its proper time goes by. This fundamental fact is true regardless of the choice of coordinates.

A similar concept of space moving arises in cosmology. Astronomers observe that the Universe is expanding. Cosmologists find it convenient to conceptualize the expansion by saying that space itself is expanding. For example, the picture that space expands makes it more straightforward, both conceptually and mathematically, to deal with regions of spacetime beyond the horizon, the surface of infinite redshift, of an observer.

6.2 Ideal black hole

The simplest kind of black hole, an **ideal** black hole, is one that is stationary, and electrovac outside its singularity. Electrovac means that the energy-momentum tensor $T_{\mu\nu}$ is zero except for the contribution from a stationary electromagnetic field. The most important ideal black holes are those that extend to asymptotically flat empty space (Minkowski space) at infinity. There are ideal black hole solutions that do not asymptote to flat empty space, but most of these have little relevance to reality. The most important ideal black hole solutions that are not flat at infinity are those containing a non-zero cosmological constant.

The next several chapters deal with ideal black holes in asymptotically flat space. The importance of ideal black holes stems from the no-hair theorem, discussed in the next section. The no-hair theorem has the consequence that, except during their initial collapse, or during a merger, real astronomical black holes are accurately described as ideal outside their horizons.

6.3 No-hair theorem

I will state and justify the no-hair theorem, but I will not prove it mathematically, since the proof is technical.

The **no-hair** theorem states that a stationary black hole in asymptotically flat space is characterized by just three quantities:

1. Mass M ;
2. Electric charge Q ;
3. Spin, usually parameterized by the angular momentum a per unit mass.

The mechanism by which a black hole loses its hair is gravitational radiation. When initially formed, whether from the collapse of a massive star or from the merger of two black holes, a black hole will form a complicated, oscillating region of spacetime. But over the course of several light crossing times, the oscillations lose energy by gravitational radiation, and damp out, leaving a stationary black hole.

Real astronomical black holes are not isolated, and continue to accrete (cosmic microwave background photons, if nothing else). However, the timescale (a light crossing time) for oscillations to damp out by gravitational radiation is usually far shorter than the timescale for accretion, so in practice real black holes are extremely well described by no-hair solutions almost all of their lives.

The physical reason that the no-hair theorem applies is that space is falling faster than light inside the horizon. Consequently, unlike a star, no energy can bubble up from below to replace the energy lost by gravitational radiation. The loss of energy by gravitational radiation brings the black hole to a state where it can no longer radiate gravitational energy. The properties of a no-hair black hole are characterized entirely by conserved quantities.

As a corollary, the no-hair theorem does not apply from the inner horizon of a black hole inward, because space ceases to fall superluminally inside the inner horizon.

If there exist other absolutely conserved quantities, such as magnetic charge (magnetic monopoles), or various supersymmetric charges in theories where supersymmetry is not broken, then the black hole will also be characterized by those quantities.

Black holes are expected not to conserve quantities such as baryon or lepton number that are thought not to be absolutely conserved, even though they appear to be conserved in low energy physics.

It is legitimate to think of the process of reaching a stationary state as analogous to reaching a condition of thermodynamic equilibrium, in which a macroscopic system is described by a small number of parameters associated with the conserved quantities of the system.

Schwarzschild Black Hole

The Schwarzschild geometry was discovered by Karl Schwarzschild in late 1915 at essentially the same time that Einstein was arriving at his final version of the General Theory of Relativity. Schwarzschild was Director of the Astrophysical Observatory in Potsdam, perhaps the foremost astronomical position in Germany. Despite his position, he joined the German army at the outbreak of World War 1, and was serving on the front at the time of his discovery. Sadly, Schwarzschild contracted a rare skin disease on the front. Returning to Berlin, he died in May 1916 at the age of 42.

The realisation that the Schwarzschild geometry describes a collapsed object, a black hole, was not understood by Einstein and his contemporaries. Understanding did not emerge until many decades later, in the late 1950s. Thorne (1994) gives a delightful popular account of the history.

7.1 Schwarzschild metric

The **Schwarzschild metric** was discovered first by Karl Schwarzschild (1916b), and then independently by Johannes Droste (1916). In a polar coordinate system $\{t, r, \theta, \phi\}$, and in geometric units $c = G = 1$, the Schwarzschild metric is

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 do^2, \quad (7.1)$$

where do^2 (this is the Landau & Lifshitz notation) is the metric of a unit 2-sphere,

$$do^2 = d\theta^2 + \sin^2\theta d\phi^2. \quad (7.2)$$

With units restored, the time-time component g_{tt} of the Schwarzschild metric is

$$g_{tt} = - \left(1 - \frac{2GM}{c^2 r}\right). \quad (7.3)$$

The Schwarzschild geometry describes the simplest kind of black hole: a black hole with mass M , but no electric charge, and no spin.

The geometry describes not only a black hole, but also any empty space surrounding a spherically symmetric mass. Thus the Schwarzschild geometry describes to a good approximation the spacetimes outside the surfaces of the Sun and the Earth.

Comparison with the spherically symmetric Newtonian metric

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)(dr^2 + r^2 d\phi^2) \quad (7.4)$$

with Newtonian potential

$$\Phi(r) = -\frac{M}{r} \quad (7.5)$$

establishes that the M in the Schwarzschild metric is to be interpreted as the mass of the black hole (Exercise 7.1).

The Schwarzschild geometry is asymptotically flat, because the metric tends to the Minkowski metric in polar coordinates at large radius

$$ds^2 \rightarrow -dt^2 + dr^2 + r^2 d\phi^2 \quad \text{as } r \rightarrow \infty. \quad (7.6)$$

Exercise 7.1 Schwarzschild metric in isotropic form. The Schwarzschild metric (7.1) does not have the same form as the spherically symmetric Newtonian metric (7.4). By a suitable transformation of the radial coordinate r , bring the Schwarzschild metric (7.1) to the isotropic form

$$ds^2 = -\left(\frac{1 - M/2R}{1 + M/2R}\right)^2 dt^2 + (1 + M/2R)^4 (dR^2 + R^2 d\phi^2). \quad (7.7)$$

What is the relation between R and r ? Hence conclude that the identification (7.5) is correct, and therefore that M is indeed the mass of the black hole. Is the isotropic form (7.7) of the Schwarzschild metric valid inside the horizon?

7.2 Stationary, static

The Schwarzschild geometry is **stationary**. A spacetime is said to be stationary if and only if there exists a timelike coordinate t such that the metric is independent of t . In other words, the spacetime possesses time translation symmetry: the metric is unchanged by a time translation $t \rightarrow t + t_0$ where t_0 is some constant. Evidently the Schwarzschild metric (7.1) is independent of the timelike coordinate t , and is therefore stationary, time translation symmetric.

As will be found below, §7.6, the Schwarzschild time coordinate t is timelike outside the horizon, but spacelike inside. Some authors therefore refer to the spacetime inside the horizon of a stationary black hole as being homogeneous. However, I think it is less confusing to refer to time translation symmetry, which is a single symmetry of the spacetime, by a single name, stationarity, everywhere in the spacetime.

The Schwarzschild geometry is also **static**. A spacetime is static if and only if in addition to being

stationary with respect to a time coordinate t , spatial coordinates can be chosen that do not change along the direction of the tangent vector $\mathbf{e}_t$. This requires that the tangent vector $\mathbf{e}_t$ be orthogonal to all the spatial tangent vectors $\mathbf{e}_\alpha$

$$\mathbf{e}_t \cdot \mathbf{e}_\alpha = g_{t\alpha} = 0 . \quad (7.8)$$

The Kerr geometry for a rotating black hole is an example of a geometry that is stationary but not static. If time t and azimuthal ϕ coordinates are coordinates associated with time and azimuthal symmetry, then the scalar product $\mathbf{e}_t \cdot \mathbf{e}_\phi$ of their tangent vectors in the Kerr geometry is a non-vanishing scalar, §9.3. Physically, in a static geometry, a system of static observers, those who are at rest in static spatial coordinates, see each other to remain at rest as time passes. In a non-static geometry, no such system of static observers exists.

The Gullstrand-Painlevé metric for the Schwarzschild geometry, discussed in §7.12, is an example of a metric that is stationary, since the metric coefficients are independent of the free-fall time t_{ff} , but not explicitly static. Observers at rest with respect to Gullstrand-Painlevé spatial coordinates fall into the black hole, and do not see each other as remaining at rest as time goes by. The Schwarzschild geometry is nevertheless static because there exist coordinates, the Schwarzschild coordinates, with respect to which the metric is explicitly static, $g_{t\alpha} = 0$. The Schwarzschild time coordinate t is thus identified as a special one: it is the unique time coordinate with respect to which the Schwarzschild geometry is manifestly static.

7.3 Spherically symmetric

The Schwarzschild geometry is also **spherically symmetric**. This is evident from the fact that the angular part $r^2 d\phi^2$ of the metric is the metric of a 2-sphere of radius r . This can be seen as follows. Consider the metric of ordinary flat 3-dimensional Euclidean space in Cartesian coordinates $\{x, y, z\}$:

$$ds^2 = dx^2 + dy^2 + dz^2 . \quad (7.9)$$

Convert to polar coordinates $\{r, \theta, \phi\}$, defined so that

$$x = r \sin \theta \cos \phi , \quad (7.10a)$$

$$y = r \sin \theta \sin \phi , \quad (7.10b)$$

$$z = r \cos \theta . \quad (7.10c)$$

Substituting equations (7.10a) into the Euclidean metric (7.9) gives

$$ds^2 = dr^2 + r^2(d\theta^2 + \sin^2 \theta d\phi^2) . \quad (7.11)$$

Restricting to a surface $r = \text{constant}$ of constant radius then gives the metric of a 2-sphere of radius r

$$ds^2 = r^2(d\theta^2 + \sin^2 \theta d\phi^2) \quad (7.12)$$

as claimed.

The radius r in Schwarzschild coordinates is the **circumferential radius**, defined such that the proper

circumference of the 2-sphere measured by observers at rest in Schwarzschild coordinates is $2\pi r$. This is a coordinate-invariant definition of the meaning of r , which implies that r is a scalar.

7.4 Energy-momentum tensor

It is straightforward (especially if you use a computer algebraic manipulation program) to follow the cookbook summarized in §2.25 to check that the Einstein tensor that follows from the Schwarzschild metric (7.1) is zero. Einstein's equations then imply that the Schwarzschild geometry has zero energy-momentum tensor.

If the Schwarzschild geometry is empty, should not the spacetime be flat, the Minkowski spacetime? There are two answers to this question. Firstly, the Schwarzschild geometry describes the geometry of empty space around a static spherically symmetric mass, such as the Sun or Earth. The geometry inside the spherically symmetric mass is described by some other metric, which connects continuously and differentiably (but not necessarily doubly differentiably, if the spherical object has an abrupt surface) to the Schwarzschild metric.

The second answer is that the Schwarzschild geometry describes the geometry of a collapsed object, a black hole, which is singular at zero radius, $r = 0$, but is otherwise empty of energy-momentum.

Exercise 7.2 Derivation of the Schwarzschild metric. There are neater and more insightful ways to derive it, but the Schwarzschild metric can be derived by turning a mathematical crank without the need for deeper conceptual understanding. Start with the assumption that the metric of a static, spherically symmetric object can be written in polar coordinates $\{t, r, \theta, \phi\}$ as

$$ds^2 = -A(r) dt^2 + B(r) dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad (7.13)$$

where $A(r)$ and $B(r)$ are some to-be-determined functions of radius r . Write down the components of the metric $g_{\mu\nu}$, and deduce its inverse $g^{\mu\nu}$. Compute all the components of the coordinate connections $\Gamma_{\lambda\mu\nu}$, equation (2.63). Of the 40 distinct connections, 9 should be non-vanishing. Compute all the components of the Riemann tensor $R_{\kappa\lambda\mu\nu}$, equation (2.112). There should be 6 distinct non-zero components. Compute all the components of the Ricci tensor $R_{\kappa\mu}$, equation (2.121). There should be 4 distinct non-zero components. Now impose that the spacetime be empty, that is, the energy-momentum tensor is zero. Einstein's equations then demand that the Ricci tensor vanishes identically. Use the requirement that $g^{tt}R_{tt} - g^{rr}R_{rr} = 0$ to show that $AB = 1$. Then use $g^{tt}R_{tt} = 0$ to derive the functional form of A . Finally, use the Newtonian limit $-g_{tt} \approx 1 + 2\Phi$ with $\Phi = -GM/r$, valid at large radius r , to fix A .

7.5 Birkhoff's theorem

Birkhoff's theorem, whose proof is deferred to Chapter 20, Exercise 20.2, states that the geometry of empty space surrounding a spherically symmetric matter distribution is the Schwarzschild geometry. That

is, if the metric is of the form

$$ds^2 = A(t, r) dt^2 + B(t, r) dt dr + C(t, r) dr^2 + D(t, r) d\theta^2, \quad (7.14)$$

where the metric coefficients A , B , C , and D are allowed to be arbitrary functions of t and r , and if the energy momentum tensor vanishes, $T_{\mu\nu} = 0$, outside some value of the circumferential radius r' defined by $r'^2 = D$, then the geometry is necessarily Schwarzschild outside that radius.

This means that if a mass undergoes spherically symmetric pulsations, then those pulsations do not affect the geometry of the surrounding spacetime. This reflects the fact that there are no spherically symmetric gravitational waves.

7.6 Horizon

The **horizon** of the Schwarzschild geometry lies at the Schwarzschild radius $r = r_s$

$$r_s = \frac{2GM}{c^2}, \quad (7.15)$$

where units of c and G have been momentarily restored. Where does this come from? The Schwarzschild metric shows that the scalar spacetime distance squared ds^2 along an interval at rest in Schwarzschild coordinates, $dr = d\theta = d\phi = 0$, is timelike, lightlike, or spacelike depending on whether the radius is greater than, equal to, or less than the Schwarzschild radius r_s :

$$ds^2 = -\left(1 - \frac{r_s}{r}\right) dt^2 \quad \begin{cases} < 0 & \text{if } r > r_s, \\ = 0 & \text{if } r = r_s, \\ > 0 & \text{if } r < r_s. \end{cases} \quad (7.16)$$

Since the worldline of a massive observer must be timelike, it follows that a massive observer can remain at rest only outside the horizon, $r > r_s$. An object at rest at the horizon, $r = r_s$, follows a null geodesic, which is to say it is a possible worldline of a massless particle, a photon. Inside the horizon, $r < r_s$, neither massive nor massless objects can remain at rest. To remain at rest, a particle inside the horizon would have to go faster than light.

A full treatment of what is going on requires solving the geodesic equation in the Schwarzschild geometry, but the results may be anticipated already at this point. In effect, space is falling into the black hole. Outside the horizon, space is falling less than the speed of light; at the horizon space is falling at the speed of light; and inside the horizon, space is falling faster than light, carrying everything with it. This is why light cannot escape from a black hole: inside the horizon, space falls inward faster than light, carrying light inward even if that light is pointed radially outward. The statement that space is falling superluminally inside the horizon of a black hole is a coordinate-invariant statement: massive or massless particles are carried inward whatever their state of motion and whatever the coordinate system.

Whereas an interval of coordinate time t switches from timelike outside the horizon to spacelike inside the

horizon, an interval of coordinate radius r does the opposite: it switches from spacelike to timelike:

$$ds^2 = \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 \quad \begin{cases} > 0 & \text{if } r > r_s , \\ = \infty & \text{if } r = r_s , \\ < 0 & \text{if } r < r_s . \end{cases} \quad (7.17)$$

It appears then that the Schwarzschild time and radial coordinates swap roles inside the horizon. Inside the horizon, the radial coordinate becomes timelike, meaning that it becomes a possible worldline of a massive observer. That is, a trajectory at fixed t and decreasing r is a possible worldline. Again this reflects the fact that space is falling faster than light inside the horizon. A person inside the horizon is inevitably compelled, as their proper time goes by, to move to smaller radial coordinate r .

Concept question 7.3 **Going forwards or backwards in time inside the horizon.** Inside the horizon, can a person can go forwards or backwards in Schwarzschild time t ? What does that mean?

7.7 Proper time

The proper time experienced by an observer at rest in Schwarzschild coordinates, $dr = d\theta = d\phi = 0$, is

$$d\tau = \sqrt{-ds^2} = \left(1 - \frac{r_s}{r}\right)^{1/2} dt . \quad (7.18)$$

For an observer at rest at infinity, $r \rightarrow \infty$, the proper time is the same as the coordinate time,

$$d\tau \rightarrow dt \quad \text{as } r \rightarrow \infty . \quad (7.19)$$

Among other things, this implies that the Schwarzschild time coordinate t is a scalar: not only is it the unique coordinate with respect to which the metric is manifestly static, but it coincides with the proper time of observers at rest at infinity. This coordinate-invariant definition of Schwarzschild time t implies that it is a scalar.

At finite radii outside the horizon, $r > r_s$, the proper time $d\tau$ is less than the Schwarzschild time dt , so the clocks of observers at rest run slower at smaller than at larger radii.

At the horizon, $r = r_s$, the proper time $d\tau$ of an observer at rest goes to zero,

$$d\tau \rightarrow 0 \quad \text{as } r \rightarrow r_s . \quad (7.20)$$

This reflects the fact that an object at rest at the horizon is following a null geodesic, and as such experiences zero proper time.

7.8 Redshift

An observer at rest at infinity looking through a telescope at an emitter at rest at radius r sees the emitter redshifted by a factor

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} = \frac{d\tau_{\text{obs}}}{d\tau_{\text{em}}} = \left(1 - \frac{r_s}{r}\right)^{-1/2}. \quad (7.21)$$

This is an example of the universally valid statement that photons are good clocks: the redshift factor is given by the rate at which the emitter's clock appears to tick relative to the observer's own clock. Equation (7.21) is an example of the general formula (2.101) for the redshift between two comoving (= rest) observers in a stationary spacetime.

It should be emphasized that the redshift factor (7.21) is valid only for an observer and an emitter at rest in the Schwarzschild geometry. If the observer and emitter are not at rest, then additional special relativistic factors will fold into the redshift.

The redshift goes to infinity for an emitter at the horizon

$$1 + z \rightarrow \infty \quad \text{as} \quad r \rightarrow r_s. \quad (7.22)$$

Here the redshift tends to infinity regardless of the motion of the observer or emitter. An observer watching an emitter fall through the horizon will see the emitter appear to freeze at the horizon, becoming ever slower and more redshifted. Physically, photons emitted vertically upward at the horizon by an infaller remain at the horizon for ever, taking an infinite time to get out to the outside observer.

7.9 “Schwarzschild singularity”

The apparent singularity in the Schwarzschild metric at the horizon r_s is not a real singularity, because it can be removed by a change of coordinates, such as to Gullstrand-Painlevé coordinates, equation (7.27). Einstein, and other influential physicists such as Eddington, failed to appreciate this. Einstein thought that the “Schwarzschild singularity” at $r = r_s$ marked the physical boundary of the Schwarzschild spacetime. After all, an outside observer watching stuff fall in never sees anything beyond that boundary.

Schwarzschild's choice of coordinates was certainly a natural one. It was natural to search for static solutions, and his time coordinate t is the only one with respect to which the metric is manifestly static. The problem is that physically there can be no static observers inside the horizon: they must necessarily fall inward as time passes. The fact that Schwarzschild's coordinate system shows an apparent singularity at the horizon reflects the fact that the assumption of a static spacetime necessarily breaks down at the horizon, where space is falling at the speed of light.

Does stuff “actually” fall in, even though no outside observer ever sees it happen? The answer is yes: when a black hole forms, it does actually collapse, and when an observer falls through the horizon, they really do fall through the horizon. The reason that an outside observer sees everything freeze at the horizon is simply a light travel time effect: it takes an infinite time for light to lift off the horizon and make it to the outside world.

7.10 Weyl tensor

For Schwarzschild, the Einstein tensor vanishes identically (because the spacetime is by assumption empty of energy-momentum). The only part of the Riemann curvature tensor that does not vanish is the Weyl tensor. The non-vanishing Weyl tensor says that gravitational tidal forces are present, even though the spacetime is empty of energy-momentum. Non-vanishing gravitational tidal forces are the signature that spacetime is curved.

The covariant (all indices down) components $C_{\kappa\lambda\mu\nu}$ of the coordinate-frame Weyl tensor of the Schwarzschild geometry, computed from equation (3.1), appear at first sight to be a mess (go ahead, compute them). However, the mess is an artefact of looking at the tensor through the distorting lens of the coordinate basis vectors e_μ , which are not orthonormal. After tetrads, Chapter 11, it will be found that the 10 components of the Weyl tensor, the tidal part of the Riemann tensor, can be decomposed in any locally inertial frame into 5 complex components of spin 0, ± 1 , and ± 2 . In a locally inertial frame whose radial direction coincides with the radial direction of the Schwarzschild metric, all components of the Weyl tensor of the Schwarzschild geometry vanish except the real spin-0 component. Spin 0 means that the Weyl tensor is unchanged under a spatial rotation about the radial direction (and it is also unchanged by a Lorentz boost in the radial direction). This spin-0 component is a coordinate-invariant scalar, the Weyl scalar C . The fact that the Weyl tensor of the Schwarzschild geometry has only a single independent non-vanishing component is plausible from the fact that the non-zero components of the coordinate-frame Weyl tensor written with two indices up and two indices down are (no implicit summation over repeated indices)

$$-\frac{1}{2}C^{tr}_{tr} = -\frac{1}{2}C^{\theta\phi}_{\theta\phi} = C^{t\theta}_{t\theta} = C^{t\phi}_{t\phi} = C^{r\theta}_{r\theta} = C^{r\phi}_{r\phi} = C , \quad (7.23)$$

where C is the Weyl scalar,

$$C = -\frac{M}{r^3} . \quad (7.24)$$

The trick of writing the 4-index Weyl tensor with 2 indices up and 2 indices down, in order to reveal a simple pattern, works in a simple spacetime like Schwarzschild, but fails in more complicated spacetimes.

7.11 Singularity

The Weyl scalar, equation (7.24), goes to infinity at zero radius,

$$C \rightarrow \infty \quad \text{as} \quad r \rightarrow 0 . \quad (7.25)$$

The diverging Weyl tensor implies that the tidal force diverges at zero radius, signalling that there is a genuine **singularity** at zero radius in the Schwarzschild geometry.

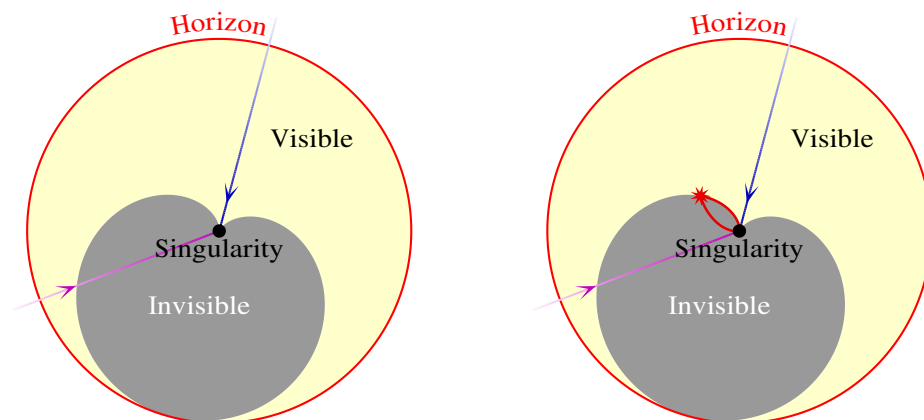


Figure 7.1 (Left) The light (yellow) shaded region shows the region visible to an infaller (blue) who falls radially to the singularity of a Schwarzschild black hole; the dark (grey) shaded region shows the region that remains invisible to the infaller. The invisible region has the shape of a cardioid, equation (7.64). If another infaller (purple) falls along a different radial direction, the two infallers not only fail to meet at the singularity, they lose causal contact with each other already some distance from the singularity. Since the two infallers fall to two causally disconnected places, the singularity cannot be a point. (Right) Same, showing the shortest causal path (red) joining the two infallers asymptotically near the singularity. The shortest causal path is a pair of light rays that start at the starred point, move in opposite azimuthal directions, and reach the infallers asymptotically near the singularity. The shortest causal path remains non-zero even though the spatial distance between the infallers tends to zero. Compare to Figure 23.2 for a Kerr black hole.

Concept question 7.4 Is the singularity of a Schwarzschild black hole a point? Is the singularity at the centre of the Schwarzschild geometry a point? **Answer.** No. Familiar experience in 3-dimensional space would suggest the answer is yes, but that conception is misleading. In the first place, general relativity fails at singularities: the locally inertial description of spacetime fails, and general relativity cannot continue worldlines of infallers beyond a singularity. Therefore singularities are not part of the spacetime described by general relativity. Presumably some other physical theory takes over at singularities, but what that theory is remains equivocal at the present time. In the second place, infallers who fall into a Schwarzschild black hole at different angular positions do not approach each other as they approach the singularity. Rather, the diverging tidal force near the singularity funnels each infaller along radially converging lines, effectively keeping the infallers isolated from each other. Moreover, the future lightcones of infallers who fall in at the same time t but at different angular positions cease to intersect once they are close enough to the singularity. Thus the infallers not only fail to touch each other, they cease even to be able to communicate with each other as they approach the singularity, as illustrated in Figure 7.1. The reader may object that the Schwarzschild metric shows that the proper angular distance between two observers separated by angle ϕ is $r d\phi$, which goes to zero at the singularity $r \rightarrow 0$. This objection fails because infallers approaching the singularity cease to be able to measure angular distances, since angularly separated points cease to be causally accessible to

the infaller. The region accessible to an infaller is cusp-like near the singularity. See Exercises 7.10 and 7.11 for a more quantitative treatment of this problem.

Concept question 7.5 Separation between infallers who fall in at different times. Consider two infallers who free-fall radially into the black hole at the same angular position, but at different times t . What is the proper spatial radial separation between the two observers at the instants they hit the singularity, at $r \rightarrow 0$? **Answer.** Infinity. At the same angular position, $d\theta = d\phi$, the proper radial separation is

$$dl = \sqrt{ds^2} = \sqrt{\frac{r_s}{r} - 1} dt \rightarrow \infty \quad \text{as } r \rightarrow 0. \quad (7.26)$$

7.12 Gullstrand-Painlevé metric

An alternative metric for the Schwarzschild geometry was discovered independently by Allvar Gullstrand and Paul Painlevé in 1921 (Gullstrand, 1922; Painlevé, 1921). (Gullstrand has priority because his paper, though published in 1922, was submitted in May 1921, whereas Painlevé's paper was a write-up of a presentation to L'Académie des Sciences in Paris in October 1921). After tetrads, it will become clear that the standard

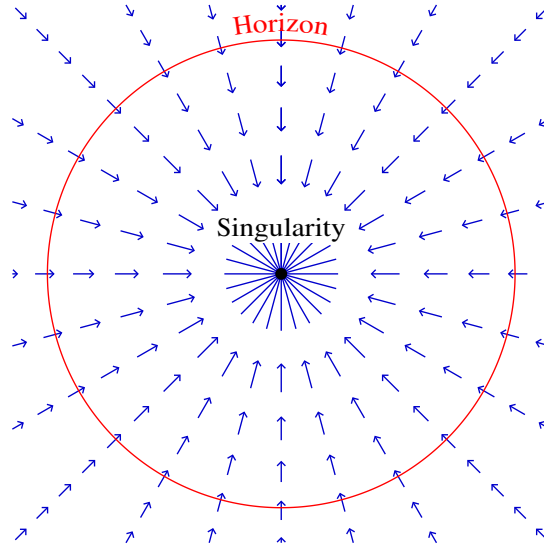


Figure 7.2 The Gullstrand-Painlevé metric for the Schwarzschild geometry encodes locally inertial frames (tetrads) that free-fall radially into the black hole at the Newtonian escape velocity β , equation (7.28). The infall velocity is less than the speed of light outside the horizon, equal to the speed of light at the horizon, and faster than light inside the horizon. The infall velocity tends to infinity at the central singularity.

way in which metrics are written encodes not only metric but also a tetrad. The Gullstrand-Painlevé line-element (7.27) encodes a tetrad that represents locally inertial frames free-falling radially into the black hole at the Newtonian escape velocity, Figure 7.2, although at the time no one, including Einstein, Gullstrand, and Painlevé, understood this. Unlike Schwarzschild coordinates, there is no singularity at the horizon in Gullstrand-Painlevé coordinates. It is striking that the mathematics was known long before physical understanding emerged.

The Gullstrand-Painlevé metric is

$$ds^2 = -dt_{\text{ff}}^2 + (dr - \beta dt_{\text{ff}})^2 + r^2 d\phi^2 . \quad (7.27)$$

Here β is the Newtonian escape velocity (with a minus sign because space is falling inward),

$$\beta = - \left(\frac{2GM}{r} \right)^{1/2} , \quad (7.28)$$

and t_{ff} is the proper time experienced by an object that free falls radially inward from zero velocity at infinity. The free fall time t_{ff} is related to the Schwarzschild time coordinate t by

$$dt_{\text{ff}} = dt - \frac{\beta}{1 - \beta^2} dr , \quad (7.29)$$

which integrates to

$$t_{\text{ff}} = t + r_s \left(2\sqrt{r/r_s} + \ln \left| \frac{\sqrt{r/r_s} - 1}{\sqrt{r/r_s} + 1} \right| \right) . \quad (7.30)$$

The time axis $\mathbf{e}_{t_{\text{ff}}}$ in Gullstrand-Painlevé coordinates is not orthogonal to the radial axis $\mathbf{e}_r$, but rather is tilted along the radial axis, $\mathbf{e}_{t_{\text{ff}}} \cdot \mathbf{e}_r = g_{t_{\text{ff}}r} = -\beta$.

The proper time of a person at rest in Gullstrand-Painlevé coordinates, $dr = d\theta = d\phi = 0$, is

$$d\tau = dt_{\text{ff}} \sqrt{1 - \beta^2} . \quad (7.31)$$

The horizon occurs where this proper time vanishes, which happens when the infall velocity β is the speed of light

$$|\beta| = 1 . \quad (7.32)$$

According to equation (7.28), this happens at $r = r_s$, which is the Schwarzschild radius, as it should be.

Exercise 7.6 Geodesics in the Schwarzschild geometry. The Schwarzschild metric is

$$ds^2 = -\Delta(r) dt^2 + \frac{1}{\Delta(r)} dr^2 + r^2 (d\theta^2 + \sin^2\theta d\phi^2) , \quad (7.33)$$

where $\Delta(r)$ is the horizon function

$$\Delta(r) = 1 - \frac{2M}{r} . \quad (7.34)$$

1. **Constants of motion.** Argue that, without loss of generality, the trajectory of a freely falling particle may be taken to lie in the equatorial plane, $\theta = \pi/2$. Argue that, for a massive particle, conservation of energy per unit rest mass E , angular momentum per unit rest mass L , and rest mass per unit rest mass implies that the 4-velocity $u^\mu \equiv dx^\mu/d\tau$ satisfies

$$u_t = -E , \quad (7.35a)$$

$$u_\phi = L , \quad (7.35b)$$

$$u_\mu u^\mu = -1 . \quad (7.35c)$$

2. **Effective potential.** Show that the radial component u^r of the 4-velocity satisfies

$$u^r = \pm (E^2 - U)^{1/2} , \quad (7.36)$$

where U is the effective potential

$$U = \left(1 + \frac{L^2}{r^2}\right) \Delta . \quad (7.37)$$

3. **Proper time in radial free-fall.** What is the proper time τ for an observer to free-fall from radius r to the singularity at zero radius, for the particular case of an observer who falls radially from rest at infinity. [Hint: What are the energy E and angular momentum L for an observer who falls radially starting from rest at infinity?]
4. **Proper time in radial free-fall—numbers.** Evaluate the proper time, in seconds, to fall from the horizon to the singularity in the case of a black hole with the mass $4 \times 10^6 M_\odot$ of the black hole at the centre of our Galaxy, the Milky Way.
5. **Circular orbits.** Circular orbits occur where the effective potential U is an extremum. Find the radii at which this occurs, as a function of angular momentum L . Solutions exist only if the absolute value $|L|$ of the angular momentum exceeds a certain critical value L_c . What is this critical value L_c ?
6. **Graph.** Graph the effective potential U for values of L (i) less than, (ii) equal to, (iii) greater than the critical value L_c . Describe physically, in words, what the possible orbital trajectories are for the various cases. [Hint: For cases (i) and (iii), values near the critical value L_c show the distinction most clearly.]
7. **Range of orbits.** Identify the ranges of radii over which circular orbits are: (i) stable, (ii) unstable, (iii) non-existent. [Hint: Stability depends on whether the extremum of the effective potential is a minimum or a maximum. Which is which? You will find it helps to consider U as a function of $1/r$ rather than r .]
8. **Angular momentum and energy in circular orbit.** Show that the angular momentum per unit mass for a circular orbit at radius r satisfies

$$|L| = \frac{r}{(r/M - 3)^{1/2}} , \quad (7.38)$$

and hence show also that the energy per unit mass in the circular orbit is

$$E = \frac{r - 2M}{[r(r - 3M)]^{1/2}} . \quad (7.39)$$

9. **Drop in orbit.** There is a certain circular orbit that has the same energy as a massive particle at rest at infinity. This is useful for starship captains to know, because it is possible to drop into this orbit using only a small amount of energy. What is the radius of the orbit? Is it stable or unstable?
10. **Orbital period.** Show that the orbital period t , as measured by an observer at rest at infinity, of a particle in circular orbit at radius r is given by Kepler's 3rd law (remarkably, Kepler's 3rd law remains true even in the fully general relativistic case, as long as t is taken to be the time measured at infinity),

$$\frac{GMt^2}{(2\pi)^2} = r^3 . \quad (7.40)$$

[Hint: Argue that the azimuthal angle ϕ evolves according to $d\phi/dt = u^\phi/u^t = L\Delta/(Er^2)$.]

Exercise 7.7 Null geodesics in the Schwarzschild geometry. The orbit equations (7.35) would appear to break down for photons, which have zero mass, hence infinite energy per unit mass E , and infinite angular momentum per unit mass L . Another way of looking at this is that photons follow null geodesics, $d\tau = 0$, so that τ , which does not change, is not a very useful time coordinate for expressing the equations of motion of photons. The difficulty is cured by introducing an affine parameter, equation (2.93), which functions as a good scalar coordinate along null geodesics.

1. **Constants of motion.** For a massless particle, the 4-velocity $v^\mu \equiv dx^\mu/d\lambda$, normalized to unit energy at infinity, satisfies

$$v_t = -1 , \quad (7.41a)$$

$$v_\phi = J , \quad (7.41b)$$

$$v_\mu v^\mu = 0 , \quad (7.41c)$$

where $J = L/E$ is the photon's angular momentum per unit energy.

2. **Effective potential.** Show that the radial component v^r of the photon 4-velocity satisfies

$$v^r = \pm (1 - V)^{1/2} , \quad (7.42)$$

where V is the effective potential

$$V = \frac{J^2}{r^2} \Delta . \quad (7.43)$$

3. **Photon sphere.** Circular orbits occur where the effective potential V is a minimum (stable orbit) or a maximum (unstable orbit). At what radius can photons orbit in circles? Is the orbit stable or unstable?
4. **Photon energy.** The photon energy $-v_t$, equation (7.41a), is normalized to one as measured by an observer at rest at infinity. Show that the energy of the photon measured by an observer on a trajectory with energy E per unit mass and angular momentum per unit mass L , relative to unit energy at infinity,

$$\omega_{\text{obs}} = u_\mu v^\mu = -\frac{E}{\Delta} \pm \frac{1}{\Delta} \sqrt{\left(E^2 - \left(1 + \frac{L^2}{r^2}\right) \Delta\right) \left(1 - \frac{J^2}{r^2} \Delta\right) + \frac{\mathbf{L} \cdot \mathbf{J}}{r^2}} , \quad (7.44)$$

where the $\pm$ sign is the sign of $u^r v^r$, which is positive or negative as the observer and photon are moving radially in the same or opposite directions.

Exercise 7.8 Geodesics in the Schwarzschild geometry in arbitrary dimensions. Standard general relativity breaks down in $N = 2$ spacetime dimensions, §11.19, and there are no black holes in $N = 2$ spacetime dimensions in the closest approximation to general relativity, Exercise 11.10 (there are however black holes in $N = 2$ spacetime dimensions in extensions of general relativity). The Schwarzschild metric in $N \geq 3$ spacetime dimensions is

$$ds^2 = -\Delta(r) dt^2 + \frac{1}{\Delta(r)} dr^2 + r^2 do^2, \quad (7.45)$$

where do^2 is the metric of a unit $N-2$ sphere, and $\Delta(r)$ is the horizon function

$$\Delta(r) = 1 - \frac{2M}{r^{N-3}}. \quad (7.46)$$

The Gullstrand-Painlevé form of the metric is equation (7.27) with infall velocity

$$\beta = -\left(\frac{2M}{r^{N-3}}\right)^{1/2}. \quad (7.47)$$

The Schwarzschild radius, where $\Delta = 0$, is

$$r_s = (2M)^{1/(N-3)}. \quad (7.48)$$

What happens when $N = 3$? What happens when $N \geq 5$? Argue that equations (7.35)–(7.37) hold, with Δ in the effective potential U , equation (7.37), being given by equation (7.46) (Tangherlini, 1963).

Solution. For $N = 3$, the horizon function 7.46 is constant $\Delta = 1 - 2M$. For $N = 3$, a coordinate transformation to coordinates $t' = t\sqrt{\Delta}$ and $r' = r/\sqrt{\Delta}$ brings the Schwarzschild line-element (7.45) to

$$ds^2 = -dt'^2 + dr'^2 + r'^2 \Delta do^2, \quad (7.49)$$

which is the metric of a cone, with angle $2\pi\sqrt{\Delta}$ around a circumference. The spacetime looks flat except for a conical vertex at $r' = 0$. A mass M bends geodesics around it, but there are no bound orbits.

The condition for a circular orbit is that the effective potential be an extremum, $dU/dr = 0$. The boundary between stable and unstable circular orbits occurs when the potential is a double extremum, $dU/dr = d^2U/dr^2 = 0$. The boundary between stable and unstable circular orbits occurs at

$$\frac{r_c}{r_s} = \left(\frac{N-1}{5-N}\right)^{1/(N-3)}, \quad \frac{L_c}{r_s} = \left(\frac{N-1}{5-N}\right)^{(5-N)/[2(N-3)]}, \quad (7.50)$$

which has real finite solutions only for $1 < N < 5$, and integral solutions only for $N = 2, 3, 4$. For $N = 2$, equations (7.50) do not apply. For $N = 3$, equations (7.50) give $r_c/r_s = e$ and $L_c/r_s = e$ (where e is the exponential); but these values are really valid not for $N = 3$, but rather for values of N infinitesimally close to but not equal to 3.

For $N \geq 5$, there are no stable circular orbits. For $N \geq 5$, the only circular orbits are unstable, which occur for $L > r_s$ if $N = 5$ or $L > 0$ if $N \geq 6$. Besides unstable circular orbits, there are unbound geodesics, and geodesics that fall into the black hole. The case $N = 4$ is the only dimension for which stable circular orbits exist.

Exercise 7.9 General relativistic precession of Mercury.

1. Conclude from Exercise 7.6 that the 4-velocity $u^\mu \equiv dx^\mu/d\tau$ of a massive particle on a geodesic in the equatorial plane of the Schwarzschild geometry satisfies

$$u^t = \frac{E}{\Delta}, \quad u^\phi = \frac{L}{r^2}, \quad u^r = \left[E^2 - \left(1 + \frac{L^2}{r^2} \Delta \right) \right]^{1/2}. \quad (7.51)$$

2. Letting $x \equiv 1/r$, show that

$$\phi = \int \frac{L dx}{[(E^2 - 1) + 2Mx - L^2x^2 + 2ML^2x^3]^{1/2}}. \quad (7.52)$$

[Hint: This is a straightforward application of equations (7.51). Do *not* try to solve this integral; leave it as given above.]

3. Suppose that the orbit varies between a known periapsis r_- and apoapsis r_+ . Define $x_- \equiv 1/r_-$ and $x_+ \equiv 1/r_+$ (note that $r_- < r_+$ so $x_- > x_+$). Argue that equation (7.52) must take the form

$$\phi = \int \frac{dx}{[(x - x_+)(x_- - x)(a - 2Mx)]^{1/2}}, \quad (7.53)$$

where

$$a \equiv 1 - 2M(x_- + x_+). \quad (7.54)$$

[Hint: This is not hard, but there are two things to do. First, you have to argue that, given the assumption that the orbit is a bounded stable orbit, there must be 3 real roots to the cubic, which must be ordered as $0 < x_+ < x_- < a/2M < \infty$. Second, you should compare the coefficients of x^3 and x^2 in the cubic in the integrands of (7.52) and (7.53).]

4. By the transformation

$$x = x_+ + (x_- - x_+)y \quad (7.55)$$

bring the integral (7.53) to the form

$$\phi = \int \frac{dy}{[y(1-y)(q - py)]^{1/2}}, \quad (7.56)$$

where

$$p = 2M(x_- - x_+), \quad q = 1 - 2M(x_- + 2x_+). \quad (7.57)$$

5. Argue that the angle ϕ integrated around a full period, from apoapsis at $y = 0$ to periapsis at $y = 1$ and back, is

$$\phi = \frac{4}{q^{1/2}} K(p/q), \quad (7.58)$$

where $K(m)$ is the complete elliptic integral of the first kind, one definition of which is

$$K(m) \equiv \frac{1}{2} \int_0^1 \frac{dy}{[y(1-y)(1-my)]^{1/2}}. \quad (7.59)$$

6. The Taylor series expansion of the complete elliptic integral is

$$K(m) = \frac{\pi}{2} \left(1 + \frac{m}{4} + \dots \right) . \quad (7.60)$$

Argue that to linear order in mass M , the angle around a full period is

$$\phi = 2\pi + 3\pi M(x_- + x_+) . \quad (7.61)$$

7. Calculate the predicted precession of the perihelion of the orbit of Mercury, expressing your answer in arcseconds per century. Google the perihelion and aphelion distances of Mercury, and its orbital period.

Exercise 7.10 A body cannot remain rigid as it approaches the Schwarzschild singularity. You have already found from Exercise 7.6 that the azimuthal angle ϕ at radius r of a particle of rest mass m on a geodesic with energy E and azimuthal angular momentum L in the equatorial plane of the Schwarzschild geometry satisfies

$$\phi = \int \frac{L dr}{\sqrt{(E^2 - m^2)r^4 - L^2 r^2 \Delta}} . \quad (7.62)$$

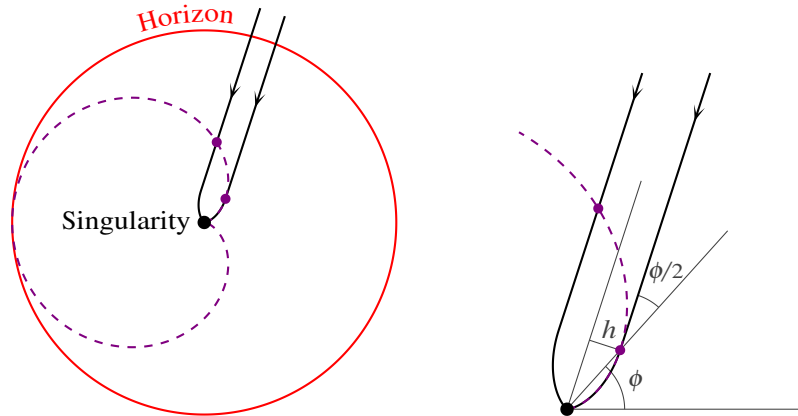


Figure 7.3 The arrowed lines, which are initially parallel, represent the worldtube of a body that remains as rigid as possible (having constant cross-sectional radius h) as it falls to the singularity at the centre of a Schwarzschild black hole. (The blow-up at right shows some details.) The dashed (purple) line shows geodesics with the maximum possible angular motion inside the horizon, namely null geodesics with infinite angular momentum per unit energy, $J = \infty$. Since the walls of the infalling body cannot exceed the speed of light, their horizontal motion near the singularity is bounded by that of $J = \infty$ null geodesics, as illustrated. The diagram gives the impression that the different (left and right) sides of the worldtube encounter each other at the singularity, but this is false. The left side of the tube can send a signal to the right side only as long as the two sides are connected by a $J = \infty$ null geodesic. The dashed line, marked with filled dots where the signal is emitted by the left side and observed by the right side, is the last such geodesic connecting the two sides: inside this dashed line the left side can no longer influence causally the right side.

1. Define $J \equiv L/E$ to be the angular momentum per unit energy. Argue that for photons, which are massless,

$$\phi = \int \frac{J dr}{\sqrt{r^4 - J^2 r^2 \Delta}} . \quad (7.63)$$

2. Argue that inside the horizon ($\Delta < 0$) the largest possible rate of change $d\phi/dr$ of the azimuthal angle ϕ with respect to radius r occurs for $J \rightarrow \infty$.
3. Show that a null geodesic with $J = \infty$ in a Schwarzschild black hole satisfies

$$r = r_s \sin^2(\phi/2) . \quad (7.64)$$

Equation (7.64) is the equation of a cardioid, illustrated by the dashed purple lines in Figure 7.3.

4. Parameterize the $J = \infty$ null geodesic satisfying equation (7.64) by $\{x, y\} \equiv \{r \cos \phi, r \sin \phi\}$. Show that

$$\frac{dy}{dx} = \tan(3\phi/2) . \quad (7.65)$$

Sketch the situation geometrically. Conclude that the radius h of a cylinder whose centre falls radially must satisfy $h \leq r \sin(\phi/2)$ in order that the walls of the cylinder not exceed the speed of light. Equivalently, conclude that a cylinder of radius h can remain rigid only down to a radius r satisfying

$$h \leq r^{3/2}/r_s^{1/2} . \quad (7.66)$$

5. Do the parts of a body that falls into a Schwarzschild black hole encounter each other at the singularity?
- Solution.** See Figure 7.3. The answer to part 5 is no, parts of a body that fall into a Schwarzschild black hole do not encounter each other at the singularity. Indeed, as illustrated in Figure 7.3, parts of a body cease to be in causal contact (cease to be able to influence each other) once they are close enough to the singularity. From the perspective of an infaller inside the horizon, the closest they ever see any point an angle ϕ away is at the edge of their past light cone, along the $J = \infty$ null geodesic.

Exercise 7.11 Causal distance between infallers near the singularity. The proper distance between two infallers who fall along different radial directions goes to zero at the singularity, but the causal distance between the two, the shortest causal path joining them, does not go to zero. The shortest causal path is the red line illustrated in the right panel of Figure 7.1, a pair of null geodesics each with the maximum possible angular momentum, $J = \infty$. A measure of distance along a null geodesic is the affine distance λ . Calculate the affine distance along the shortest causal path between infallers approaching the singularity.

Solution. Normalized to a frame at rest at infinity, the affine distance λ along a null geodesic is obtained by integrating $d\phi/d\lambda = r^2/J$, equation (7.41b), or equivalently $dr/d\lambda = v^r$ from equation (7.42), giving

$$\lambda = \frac{1}{J} \int r^2 d\phi = \int \frac{dr}{\sqrt{1 - J^2 \Delta/r^2}} . \quad (7.67)$$

Normalized to the frame of an observer, the affine distance λ_{obs} is

$$\lambda_{\text{obs}} = \omega_{\text{obs}} \lambda , \quad (7.68)$$

where ω_{obs} is the observed energy (7.44) of the photon relative to that at infinity. The observed affine distance

to an object coincides with proper distance to it measured by the observer in their immediate locally inertial vicinity. The shortest causal path joining infallers near the singularity is realised by a pair of photons emitted in opposite directions with maximum angular momentum, $J = \infty$, from a point half way (in angle) between the infallers, illustrated by the red line in the right panel of Figure 7.1. The causal path has two symmetrically equal parts, each following the path of a cardioid, equation (7.64). If the angular separation between the two infallers near the singularity is 2ϕ , then the observed affine distance along the shortest causal path is $2\lambda_{\text{obs}}$, twice the affine distance along each individual null segment,

$$\lambda_{\text{obs}} = \frac{\omega_{\text{obs}}}{J} \int_0^\phi r^2 d\phi = \frac{\omega_{\text{obs}} r_s^2}{J} \int_0^\phi \sin^4(\phi/2) d\phi = \frac{\omega_{\text{obs}} r_s^2}{J} \left(\frac{3}{8}\phi - \frac{1}{2}\sin\phi + \frac{1}{16}\sin 2\phi \right). \quad (7.69)$$

The ϕ -dependent factor in parentheses on the right hand side of equations (7.69) is $\approx \frac{1}{80}\phi^5$ at small separations ϕ . The observed energy ω_{obs} depends on the position and motion of the observer. Radially-falling observers ($L = 0$) near the singularity watching $J = \infty$ null geodesics see photon energy, from equation (7.44),

$$\omega_{\text{obs}} = \frac{J}{r_{\text{obs}}}, \quad (7.70)$$

so the factor on the right hand side of the expression (7.69) for the observed affine distance is

$$\frac{\omega_{\text{obs}} r_s^2}{J} = \frac{r_s^2}{r_{\text{obs}}}, \quad (7.71)$$

which diverges at the singularity, $r_{\text{obs}} \rightarrow 0$. The divergence is a symptom of the failure of general relativity, the cessation of the existence of locally inertial frames, at the singularity. Notwithstanding the divergence, the robust conclusion is that the causal distance between two infallers does not go to zero at the singular surface.

Exercise 7.12 Maximum transverse velocity of a light signal inside the horizon. Again consider two infallers who free-fall radially along radial paths at different angular positions. The maximum transverse velocity with which they can send signals to each other is, once again, along $J = \infty$ null geodesics. Show that this maximum transverse velocity is

$$\left. \frac{rd\phi}{dt_{\text{ff}}} \right|_{J=\infty} = \sqrt{1 - \frac{r}{r_s}}. \quad (7.72)$$

The maximum transverse velocity is always less than the speed of light, but tends to the speed of light at the singularity.

Solution. The relation between the radius r and angle ϕ along a $J = \infty$ null geodesic is given by equation (7.64). The relation between radius r and proper time t_{ff} for a radial free-faller follows from $dr/dt_{\text{ff}} = \beta$ in the Gullstrand-Painlevé metric (7.27).

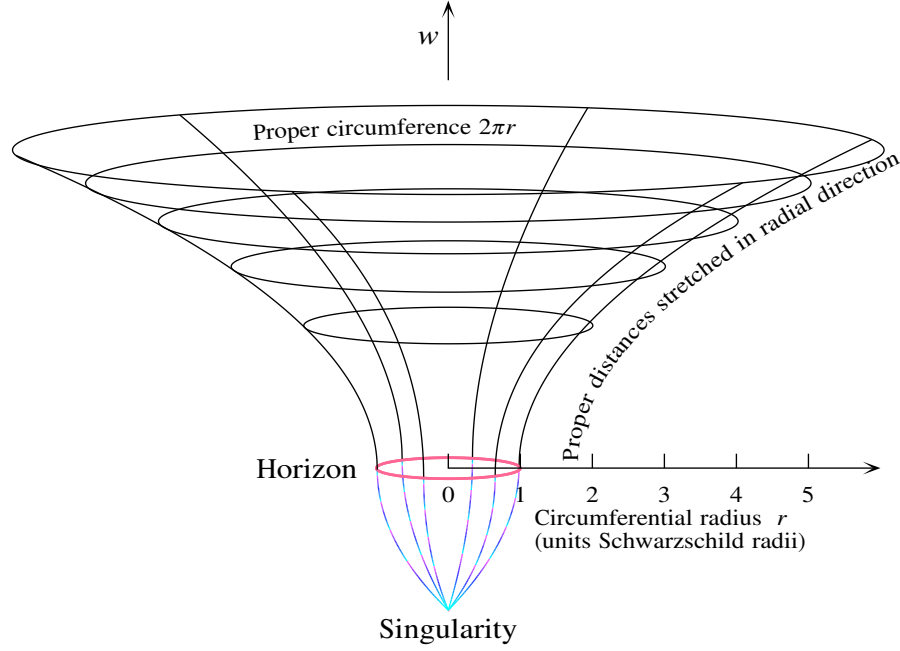


Figure 7.4 Embedding diagram of the Schwarzschild geometry. The 2-dimensional surface represents the 3-dimensional Schwarzschild geometry at a fixed instant of Schwarzschild time t . Each circle represents a sphere, of proper circumference $2\pi r$, as measured by observers at rest in the geometry. The proper radial distance measured by observers at rest is stretched in the radial direction, as shown in the diagram. The stretching is infinite at the horizon, so the spatial geometry there looks like a vertical cliff. Radial lines in the Schwarzschild geometry are spacelike outside the horizon, but timelike inside the horizon.

7.13 Embedding diagram

An **embedding diagram** is a visual aid to understanding geometry. It is a depiction of a lower dimensional geometry in a higher dimension. A classic example is the illustration of the geometry of a 2-sphere embedded in 3-dimensional space, Figure 2.2.

Figure 7.4 shows an embedding diagram of the spatial Schwarzschild geometry at a fixed instant of Schwarzschild time t . To the polar coordinates r, θ, ϕ of the 3D Schwarzschild spatial geometry, adjoin a fourth spatial coordinate w . The metric of 4D Euclidean space in the coordinates w, r, θ, ϕ , is

$$dl^2 = dw^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2. \quad (7.73)$$

The spatial Schwarzschild geometry is represented by a 3D surface embedded in the 4D Euclidean geometry, such that the proper distance dl in the radial direction satisfies equation (7.17), that is

$$dl^2 = \frac{dr^2}{1 - r_s/r} = dw^2 + dr^2. \quad (7.74)$$

Equation (7.74) rearranges to

$$dw = \frac{dr}{\sqrt{r/r_s - 1}} , \quad (7.75)$$

which integrates to

$$w = 2\sqrt{\frac{r}{r_s} - 1} . \quad (7.76)$$

The embedded Schwarzschild surface has the shape of a square root, infinitely steep at the horizon $r = r_s$, as illustrated by Figure 7.4.

Inside the horizon, proper radial distances change to being timelike, $dl^2 < 0$, equation (7.17). Here the Schwarzschild geometry at fixed Schwarzschild time t (which is a spacelike coordinate inside the horizon) can be embedded in a 4D Minkowski space in which the fourth coordinate w is timelike,

$$dl^2 = -dw^2 + dr^2 + r^2 d\phi^2 . \quad (7.77)$$

The embedded surface inside the horizon satisfies

$$w = -2\sqrt{1 - \frac{r}{r_s}} , \quad (7.78)$$

with a minus sign chosen so that the coordinate w is negative inside the horizon, whereas it is positive outside the horizon. The two embeddings (7.76) and (7.78) can be patched together at the horizon (though not doubly differentiably), as illustrated in Figure 7.4.

It should be emphasized that the embedding diagram of the Schwarzschild geometry at fixed Schwarzschild time t has a limited physical meaning. Fixing the time t means choosing a certain hypersurface through the geometry. Other choices of hypersurface will yield different embedding diagrams. For example, the Gullstrand-Painlevé metric (7.27) is spatially flat at fixed free-fall time t_{ff} , so in that case the embedding diagram would simply illustrate flat space, with no funny business at the horizon.

7.14 Schwarzschild spacetime diagram

In general relativity as in special relativity, a spacetime diagram is a plot of space versus time.

Figure 7.5 shows a spacetime diagram of the Schwarzschild geometry. In this diagram, Schwarzschild time t increases vertically upward, while circumferential radius r increases horizontally.

The more or less diagonal lines in Figure 7.5 are outgoing and infalling radial null geodesics. The radial null geodesics are not at 45° , as they would be in a special relativistic spacetime diagram. In Schwarzschild coordinates, light rays that fall radially ($d\theta = d\phi = 0$) inward or outward follow null geodesics

$$ds^2 = -\left(1 - \frac{r_s}{r}\right) dt^2 + \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 = 0 . \quad (7.79)$$

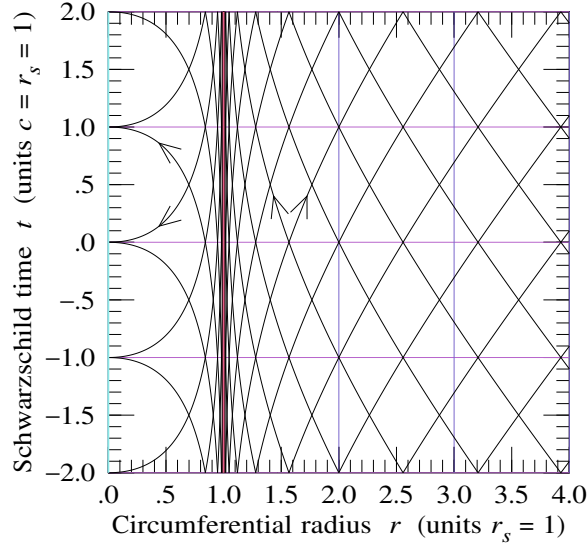


Figure 7.5 Spacetime diagram of the Schwarzschild geometry, in Schwarzschild coordinates. The horizontal axis is the circumferential radius r , while the vertical axis is Schwarzschild time t . The horizon (pink) is at one Schwarzschild radius, $r = r_s$. The singularity (cyan) is at zero radius, $r = 0$. The more or less diagonal lines (black) are outgoing and infalling null geodesics. The outgoing and infalling null geodesics appear not to cross the horizon, but this is an artefact of the Schwarzschild coordinate system.

Radial null geodesics thus follow

$$\frac{dr}{dt} = \pm \left(1 - \frac{r_s}{r}\right), \quad (7.80)$$

in which the $\pm$ sign is $+$ for outgoing, $-$ for infalling rays. Equation (7.80) shows that $dr/dt \rightarrow 0$ as $r \rightarrow r_s$, suggesting that null rays, whether infalling or outgoing, never cross the horizon. In the Schwarzschild spacetime diagram 7.5, null geodesics asymptote to the horizon, but never actually cross it. This feature of Schwarzschild coordinates was first noticed by Droste (1916), and contributed to the historical misconception that black holes stopped at their horizons. The failure of geodesics to cross the horizon is an artefact of Schwarzschild's choice of coordinates, which are adapted to observers at rest, whereas no locally inertial frame can remain at rest at the horizon.

7.15 Gullstrand-Painlevé spacetime diagram

Figure 7.6 shows a spacetime diagram of the Schwarzschild geometry in Gullstrand-Painlevé coordinates t_{ff} and r in place of Schwarzschild coordinates t and r . As the spacetime diagram shows, in Gullstrand-Painlevé

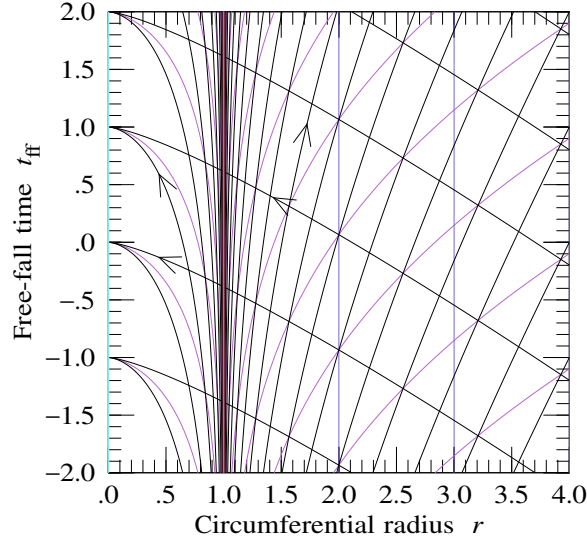


Figure 7.6 Gullstrand-Painlevé, or free-fall, spacetime diagram, in units $r_s = c = 1$. In this spacetime diagram the time coordinate is the Gullstrand-Painlevé time t_{ff} , which is the proper time of observers who free-fall radially from zero velocity at infinity. The radial coordinate r is the circumferential radius, and the horizon and singularity are at $r = r_s$ and $r = 0$, as in the Schwarzschild spacetime diagram, Figure 7.5. In contrast to the spacetime diagram in Schwarzschild coordinates, in Gullstrand-Painlevé coordinates infalling light rays do cross the horizon.

coordinates infalling light rays cross the horizon. Unfortunately, neither Gullstrand nor Painlevé, nor anyone else at that time, grasped the physical significance of their metric.

7.16 Eddington-Finkelstein spacetime diagram

In 1958, David Finkelstein (1958) carried out an elementary transformation of the time coordinate which seemed to show that infalling light rays could indeed pass through the horizon. It turned out that Eddington had already discovered the transformation in 1924 (Eddington, 1924), though at that time the physical implications were not grasped. Again, it is striking that the mathematics was in place long before physical understanding emerged.

In Schwarzschild coordinates, radially outgoing or infalling light rays follow equation (7.80). Equation (7.80) integrates to

$$t = \pm \left(r + r_s \ln \left| \frac{r}{r_s} - 1 \right| \right), \quad (7.81)$$

which shows that Schwarzschild time t approaches $\pm\infty$ logarithmically as null rays approach the horizon.

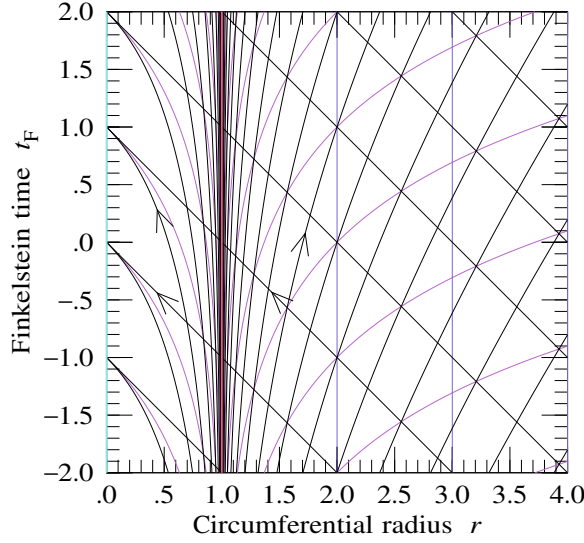


Figure 7.7 Finkelstein spacetime diagram, in units $r_s = c = 1$. Here the time coordinate is taken to be the Finkelstein time coordinate t_F , equation (7.82). The Finkelstein time coordinate t_F is constructed so that radially infalling light rays are at 45° .

Finkelstein defined his time coordinate t_F by

$$t_F \equiv t + r_s \ln \left| \frac{r}{r_s} - 1 \right|, \quad (7.82)$$

which has the property that infalling null rays follow

$$t_F + r = \text{constant}. \quad (7.83)$$

In other words, on a spacetime diagram in Finkelstein coordinates, Figure 7.7, radially infalling light rays move at 45° , the same as in a special relativistic spacetime diagram.

7.17 Kruskal-Szekeres spacetime diagram

After Finkelstein had transformed coordinates so that radially infalling light rays moved at 45° in a spacetime diagram, it was natural to look for coordinates in which outgoing as well as infalling light rays are at 45° . Kruskal and Szekeres independently provided such a transformation in 1960 (Kruskal, 1960; Szekeres, 1960).

Define the tortoise, or Regge-Wheeler (Regge and Wheeler, 1957), coordinate r^* by

$$r^* \equiv \int \frac{dr}{1 - r_s/r} = r + r_s \ln \left| \frac{r}{r_s} - 1 \right|. \quad (7.84)$$

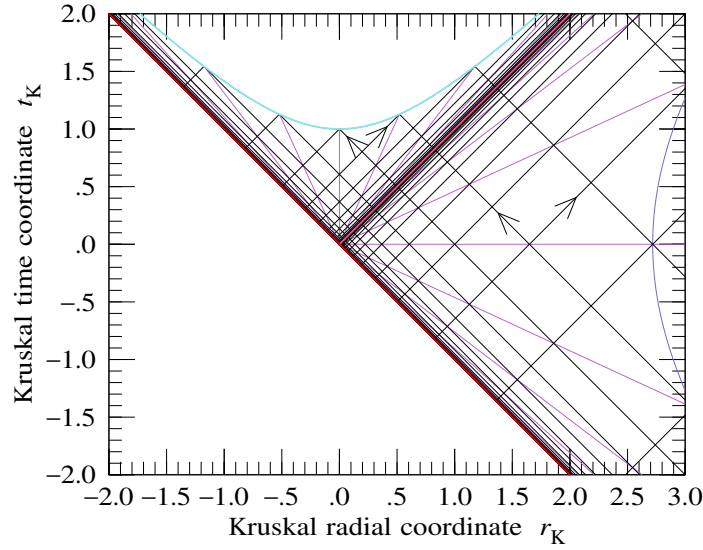


Figure 7.8 Kruskal-Szekeres spacetime diagram, in units $r_s = c = 1$. Kruskal-Szekeres coordinates are arranged such that not only infalling, but also outgoing null rays move at 45° on the spacetime diagram. The Kruskal-Szekeres spacetime diagram reveals the causal structure of the Schwarzschild geometry. The singularity (cyan) at $r = 0$, at the upper edge of the spacetime diagram, is revealed to be a spacelike surface. Besides the usual horizon (pink), there is an antihorizon (red), which was not apparent in Schwarzschild or Finkelstein coordinates. In the Kruskal-Szekeres spacetime diagram, lines of constant circumferential radius r (blue) are hyperboloids, while lines of constant Schwarzschild time t (violet) are straight lines passing through the origin, the same as in the spacetime wheel, Figure 1.14, or as in Rindler space. Contours of constant Schwarzschild time t (violet) and of constant circumferential radius r (blue) are spaced uniformly at intervals of 1 (in units $r_s = c = 1$), and similarly infalling and outgoing null rays (black) are spaced uniformly by 1.

Then radially infalling and outgoing null rays follow

$$\begin{aligned} r^* + t &= \text{constant} & \text{infalling} , \\ r^* - t &= \text{constant} & \text{outgoing} . \end{aligned} \tag{7.85}$$

In a spacetime diagram in coordinates t and r^* , infalling and outgoing light rays are indeed at 45° . Unfortunately the metric in these coordinates is still singular at the horizon $r = r_s$:

$$ds^2 = \left(1 - \frac{r_s}{r}\right) (-dt^2 + dr^{*2}) + r^2 d\phi^2 . \tag{7.86}$$

The singularity at the horizon can be eliminated by the following transformation into Kruskal-Szekeres

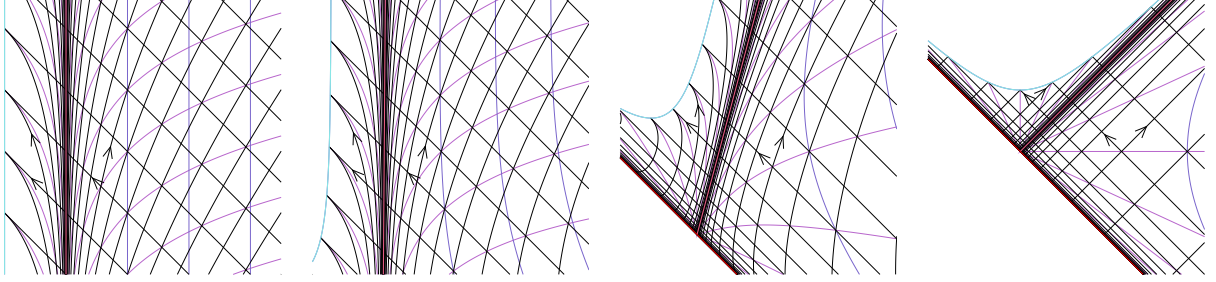


Figure 7.9 From left to right, the Finkelstein spacetime diagram, Figure 7.7, morphs into the Kruskal-Szekeres spacetime diagram, Figure 7.8. The morph illustrates how the antihorizon, or past horizon (red), emerges from the depths of $t = -\infty$. Like the horizon, the antihorizon is a null surface, thus appearing at 45° in the Kruskal-Szekeres spacetime diagram.

coordinates t_K and r_K :

$$\begin{aligned} r_K + t_K &= r_s \exp\left(\frac{r^* + t}{2r_s}\right), \\ r_K - t_K &= \pm r_s \exp\left(\frac{r^* - t}{2r_s}\right), \end{aligned} \quad (7.87)$$

where the $\pm$ sign in the last equation is $+$ outside the horizon, $-$ inside the horizon. The Kruskal-Szekeres metric is

$$ds^2 = \frac{4r_s}{r} e^{-r/r_s} (-dt_K^2 + dr_K^2) + r^2 d\phi^2, \quad (7.88)$$

which is non-singular at the horizon. The Schwarzschild radial coordinate r , which appears in the factors $(4r_s/r)e^{-r/r_s}$ and r^2 in the Kruskal metric, is to be understood as an implicit function of the Kruskal coordinates t_K and r_K .

7.18 Antihorizon

The Kruskal-Szekeres spacetime diagram reveals a new feature that was not apparent in Schwarzschild or Finkelstein coordinates. Dredged from the depths of $t = -\infty$ appears a null line $r_K + t_K = 0$, Figure 7.9. The null line is at radius $r = r_s$, but it does not correspond to the horizon that a person might fall into. The null line is called the **antihorizon**.

7.19 Analytically extended Schwarzschild geometry

The Schwarzschild geometry is analytic, and there is a unique analytic continuation of the geometry through the antihorizon. The extended geometry consists of two copies of the Schwarzschild geometry, glued along

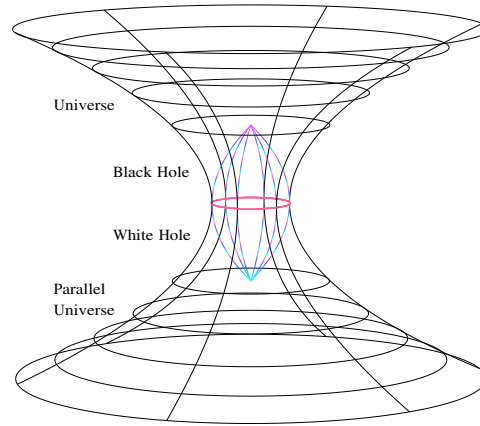


Figure 7.10 Embedding diagram of the analytically extended Schwarzschild geometry. The analytically extended geometry is constructed by gluing together two copies of the Schwarzschild geometry along the antihorizon. The extended geometry contains a Universe, a Parallel Universe, a Black Hole, and a White Hole.

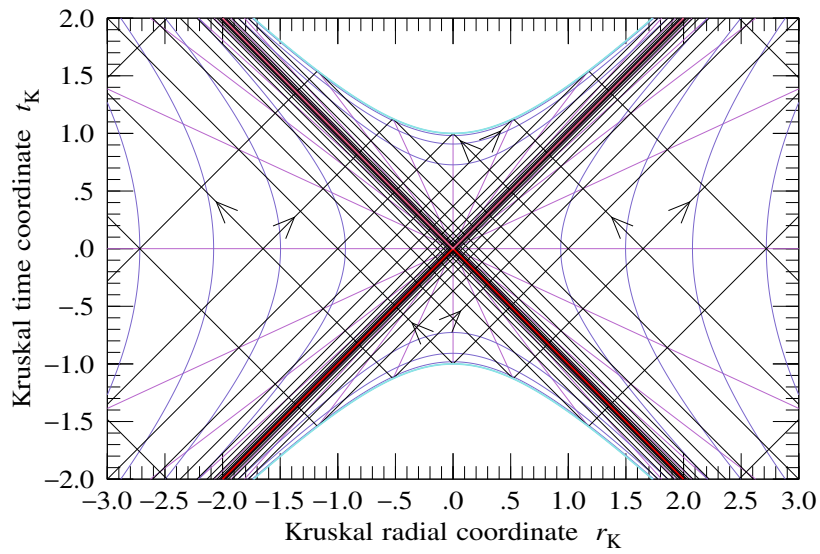


Figure 7.11 Analytically extended Kruskal-Szekeres spacetime diagram, in units $r_s = c = 1$. The analytically extended horizon and antihorizon (crossing pink/red lines at 45°) divide the spacetime into 4 regions, a Universe region at right, a Black Hole region bounded by the singularity at top, a Parallel Universe region at left, and a White Hole region bounded by a singularity at bottom. The White Hole is a time-reversed version of the Black Hole.

their antihorizons, as illustrated in the embedding diagram in Figure 7.10. The embedding diagram 7.10 gives the impression of a static wormhole, but this is an artefact of everything being frozen at the horizon in Schwarzschild coordinates.

Figure 7.11 shows the Kruskal spacetime diagram of the analytically extended Schwarzschild geometry. Whereas the original Schwarzschild geometry showed an asymptotically flat region and a black hole region separated by a horizon, the complete analytically extended Schwarzschild geometry shows two asymptotically flat regions, together with a black hole and a white hole. Relativists typically label the regions I, II, III, and IV, but I like to call them by name: “Universe,” “Black Hole,” “Parallel Universe,” and “White Hole.”

The **white hole** is a time-reversed version of the black hole. Whereas space falls inward faster than light inside the black hole, space falls outward faster than light inside the white hole. In the Gullstrand-Painlevé metric (7.27), the velocity $\beta = \pm(2M/r)^{1/2}$ is negative for the black hole, positive for the white hole.

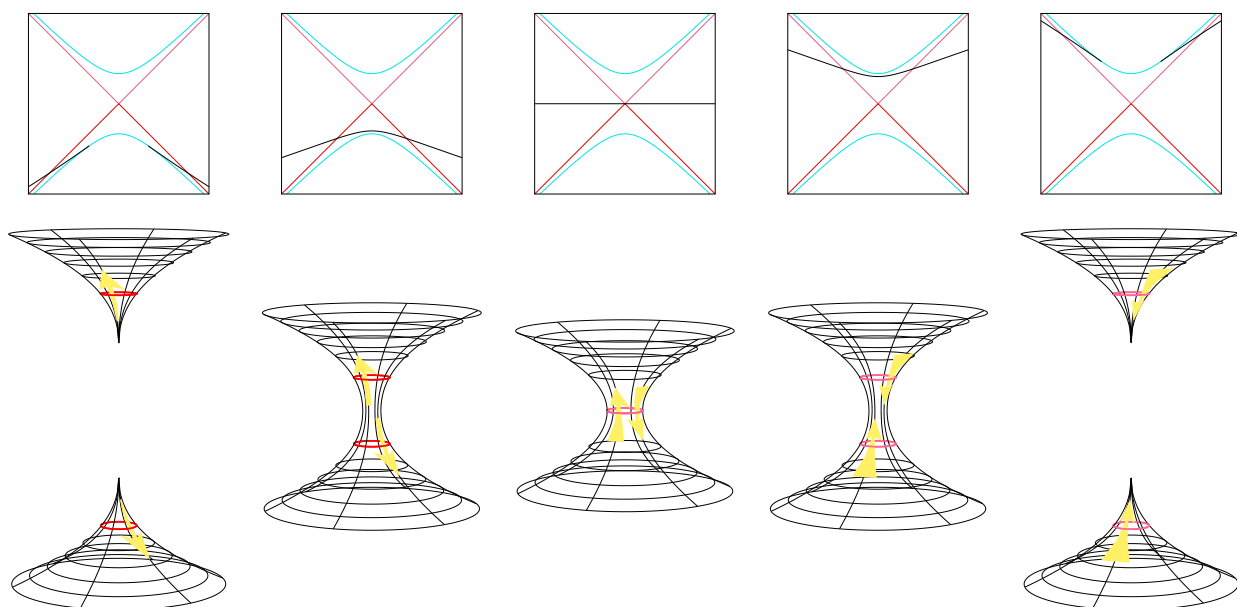


Figure 7.12 Sequence of embedding diagrams of spatial slices of the analytically extended Schwarzschild geometry, progressing in time from left to right. Two white holes merge, form an Einstein-Rosen bridge, then fall apart into two black holes. The wormhole formed by the Einstein-Rosen bridge is non-traversable. The (yellow) arrows indicate the direction in which an object can cross the horizon. At left, travellers in the two universes cannot fall into their respective white holes, because objects can cross the white hole horizons (red) only in the outward direction. The horizons cross in the middle diagram, without the arrows changing direction. After this point, travellers in the two universes can fall through their respective black hole horizons (pink) into the Einstein-Rosen bridge, and temporarily meet up with each other. Unfortunately, having fallen through the black hole horizons, they cannot exit, and are doomed to hit the singularity. The insets at top show the adopted spatial slicings on the Kruskal spacetime diagram. The adopted slicings are engineered to give the embedding diagrams an appealing look, and have no fundamental significance.

The Kruskal diagram shows that the universe and the parallel universe are connected, but only by spacelike lines. This spacelike connection is called the **Einstein-Rosen bridge**, and constitutes a wormhole connecting the two universes. Because the connection is spacelike, it is impossible for a traveller to pass through this wormhole. The wormhole is said to be **non-traversable**.

Figure 7.12 illustrates a sequence of embedding diagrams for spatial slices of the analytically extended Schwarzschild geometry. Although two travellers, one from the universe and one from the parallel universe, cannot travel to each other's universe, they can meet, but only inside the black hole. Inside the black hole, they can talk to each other, and they can see light from each other's universe. Sadly, the enlightenment is only temporary, because they are doomed soon to hit the central singularity.

It should be emphasized that the white hole and the wormhole in the Schwarzschild geometry are a mathematical construction with as far as anyone knows no relevance to reality. Nevertheless it is intriguing that such bizarre objects emerge already in the simplest general relativistic solution for a black hole.

7.20 Penrose diagrams

Roger Penrose, as so often, had a novel take on the business of spacetime diagrams (Penrose, 2011). Penrose conceived that the primary purpose of a spacetime diagram should be to portray the causal structure of the spacetime, and that the specific choice of coordinates was largely irrelevant. After all, general relativity allows arbitrary choices of coordinates.

In addition to requiring that light rays be at 45° , Penrose wanted to bring regions at infinity (in time or space) to a finite position on the spacetime diagram, so that the entire spacetime could be seen at once. Such diagrams are called **Penrose diagrams**, or **conformal diagrams**.

Penrose diagrams are bona-fide spacetime diagrams. Penrose time and space coordinates t_P and r_P can be defined by any conformal transformation of Kruskal-Szekeres coordinates

$$r_P \pm t_P = f(r_K \pm t_K) \quad (7.89)$$

for which $f(z)$ is finite as $z \rightarrow \pm\infty$. The transformation (7.89) brings spatial and temporal infinity to finite values of the coordinates, while keeping infalling and outgoing light rays at 45° in the spacetime diagram. It is common to draw a Penrose diagram with the singularity horizontal, which can be accomplished by choosing the function $f(z)$ to be odd, $f(-z) = -f(z)$. Figure 7.13 shows a spacetime diagram in Penrose coordinates with $f(z)$ set to (units $r_s = 1$)

$$f(z) = \frac{2}{\pi} \operatorname{atan} z . \quad (7.90)$$

Figure 7.14 illustrates a morph of the Kruskal-Szekeres spacetime diagram, Figure 7.8, into the Penrose spacetime diagram, Figure 7.13.

Figure 7.15 illustrates the Penrose diagram of the analytically extended Schwarzschild geometry.

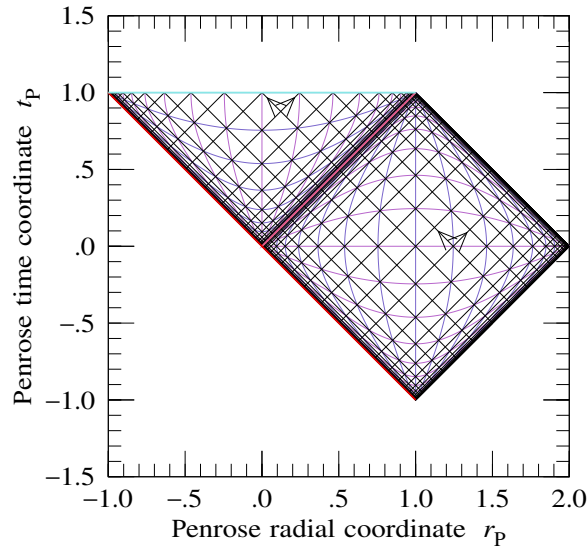


Figure 7.13 Penrose spacetime diagram, in units $r_s = c = 1$. The Penrose coordinates t_P and r_P here are defined by equations (7.89) and (7.90). Lines of constant Schwarzschild time t (violet), and infalling and outgoing null lines (black) are spaced uniformly at intervals of 1 (units $r_s = c = 1$), while lines of constant circumferential radius r (blue) are spaced uniformly in the tortoise coordinate r^* , equation (7.84), so that the intersections of t and r lines are also intersections of infalling and outgoing null lines.

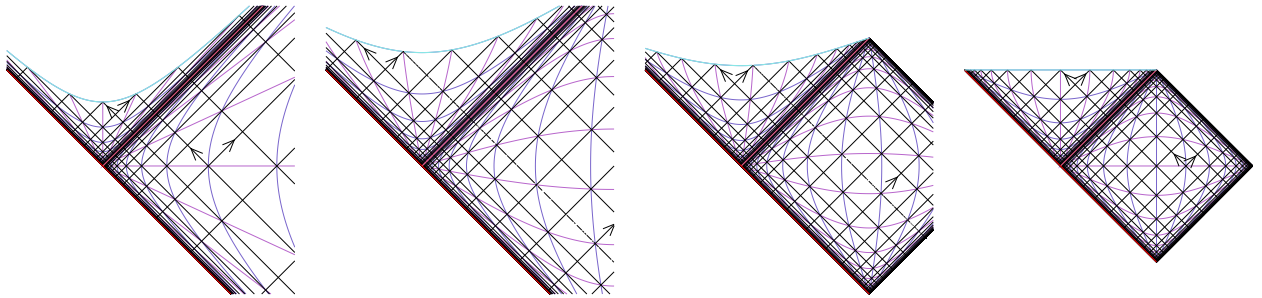


Figure 7.14 From left to right, the Kruskal-Szekeres spacetime diagram, Figure 7.8, morphs into the Penrose spacetime diagram, Figure 7.13.

7.21 Penrose diagrams as guides to spacetime

In the literature, Penrose diagrams are usually sketched, not calculated, the aim being to convey a conceptual understanding of the spacetime without obsessing over detail.

Figure 7.16 shows a Penrose diagram of the Schwarzschild geometry, with the Universe and Black Hole regions, and the various boundaries of the diagram, marked. The 45° edges of the Penrose diagram at infinite

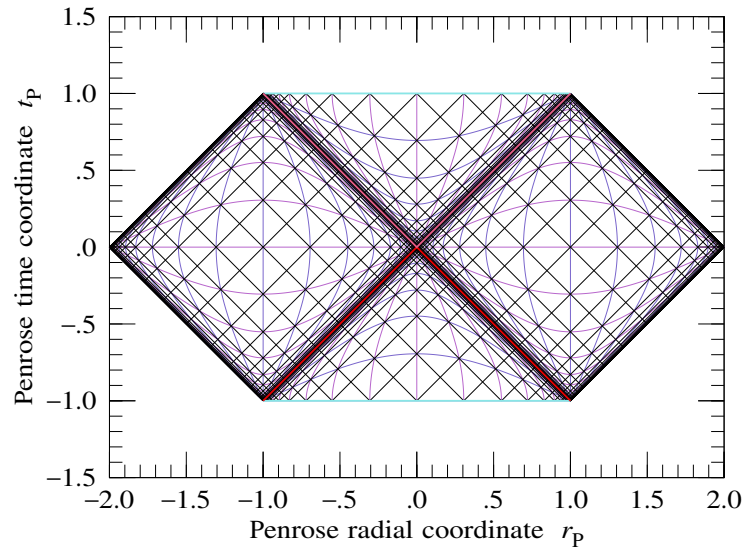


Figure 7.15 Penrose spacetime diagram of the analytically extended Schwarzschild geometry. This is the analytically extended version of Figure 7.13.

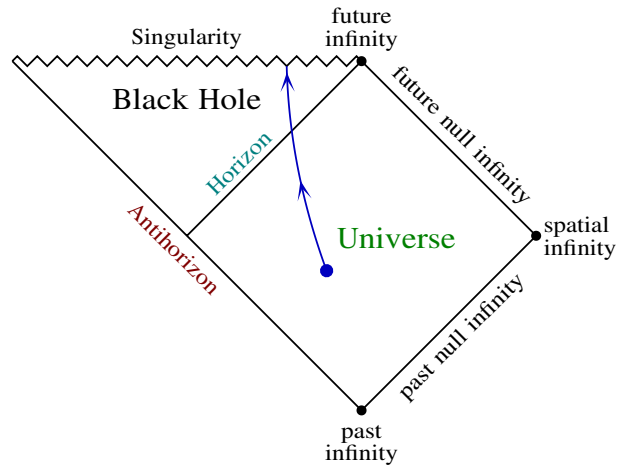


Figure 7.16 Penrose diagram of the Schwarzschild geometry, labelled with the Universe and Black Hole regions, and their various boundaries. The (blue) line at less than 45° from vertical is a possible trajectory of a person who falls through the horizon from the Universe into the Black Hole. Once inside the horizon, the infaller cannot avoid the Singularity.

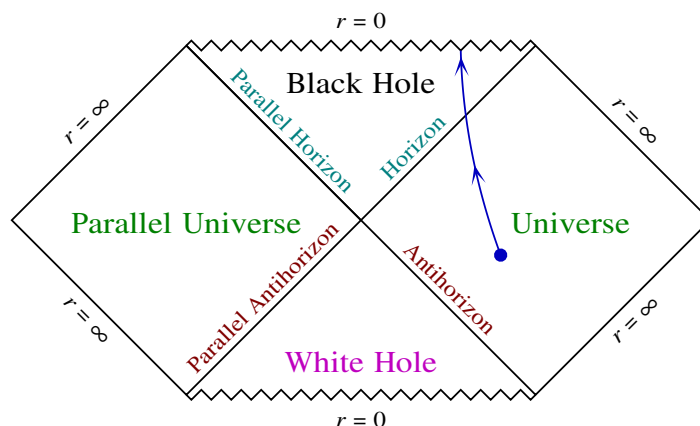


Figure 7.17 Penrose diagram of the analytically extended Schwarzschild geometry.

radius, $r = \infty$, are called **past** and **future null infinity**, often designated in the mathematical literature by $\mathcal{I}_+$ and $\mathcal{I}_-$ (commonly pronounced scri-plus and scri-minus, scri being short for script-I). The corners of the Penrose diagram in the infinite past or future are called **past** and **future infinity**, often designated i_- and i_+ , while the corner at infinite radius is called **spatial infinity**, often designated i_0 .

The Schwarzschild geometry, being asymptotically flat (Minkowski), has no boundary at infinity. Thus the boundary at infinity in the Penrose diagram is not part of the spacetime manifold. However, a worldline that extends into the indefinite past converges towards past infinity, while a worldline that extends into the indefinite future outside the black hole converges towards future infinity.

A Penrose diagram is an indispensable guide to finding your way around a complicated spacetime such as a black hole. However, a Penrose diagram can be deceiving, because the conformal mapping distorts the spacetime. Most of the physical spacetime in the Penrose diagram of the Schwarzschild geometry is compressed to the corners of the diagram, to past, future, and spatial infinity, and to the top left point at the intersection of the antihorizon with the singularity.

Figure 7.17 shows the Penrose diagram of the analytically extended Schwarzschild geometry, with the four regions, Universe, Black Hole, Parallel Universe, and White Hole marked. Again, relativists typically call these regions I, II, III, and IV, but I like to give them names. I've also given names to the various horizons. The names are unconventional, but reasonable.

Concept question 7.13 Penrose diagram of Minkowski space. Draw a Penrose diagram of Minkowski space.

7.22 Future and past horizons

Hawking and Ellis (1973) define the **future horizon** of the worldline of an observer to be the boundary of the past lightcone of the continuation of the worldline into the indefinite future. Likewise the **past horizon** of the worldline of an observer is the boundary of the future lightcone of the continuation of the worldline into the indefinite past. The definition of future and past horizons is observer-dependent.

The horizon of a Schwarzschild black hole is a future horizon for observers who remain at a finite distance outside the black hole for ever. The antihorizon of a Schwarzschild black hole is a past horizon for observers who remained a finite distance outside the black hole in the indefinite past.

The **causal diamond** of an observer is the part of spacetime bounded by the observer's past and future horizons. The causal diamond is the region of spacetime to which the observer can, at some point on their worldline, send a signal, and from which the observer can, at some other point on their worldline, receive a signal. For example, the Universe region of the Penrose diagram 7.16 is the causal diamond of an observer who starts at past infinity and ends at future infinity, without falling into the black hole.

7.23 Oppenheimer-Snyder collapse to a black hole

Realistic collapse of a star to a black hole is not expected to produce a white hole or parallel universe.

The simplest model of a collapsing star is a spherical ball of uniform density and zero pressure which free falls from zero velocity at infinity, a problem first solved by Oppenheimer and Snyder (1939). In this simple model, the interior of the star is described by a collapsing Friedmann-Lemaître-Robertson-Walker metric (see Chapter 10), while the exterior is described by the Schwarzschild solution. The assumption that the star collapses from zero velocity at infinity implies that the FLRW geometry is spatially flat, the simplest case. To continue the geometry between Schwarzschild and FLRW geometries, it is neatest to use the Gullstrand-Painlevé metric, with the Gullstrand-Painlevé infall velocity β at the edge of the star set equal to minus r times the Hubble parameter of the collapsing FLRW metric, $-rH \equiv -r d \ln a / dt$. Section 20.15 describes a systematic approach to solving the Oppenheimer-Snyder problem.

Figure 7.18 shows the star collapse as seen by an outside observer at rest at a radius of 10 Schwarzschild radii. The Figure is correctly ray-traced, taking into account the different travel times of light from the various parts of the star to the observer. The collapsing star appears to freeze at the horizon, taking on the appearance of a Schwarzschild black hole.

When Oppenheimer & Snyder first did their calculation, the result seemed paradoxical. An outsider saw the collapsing star freeze at its horizon and never get further, even to the end of time. Yet an observer who

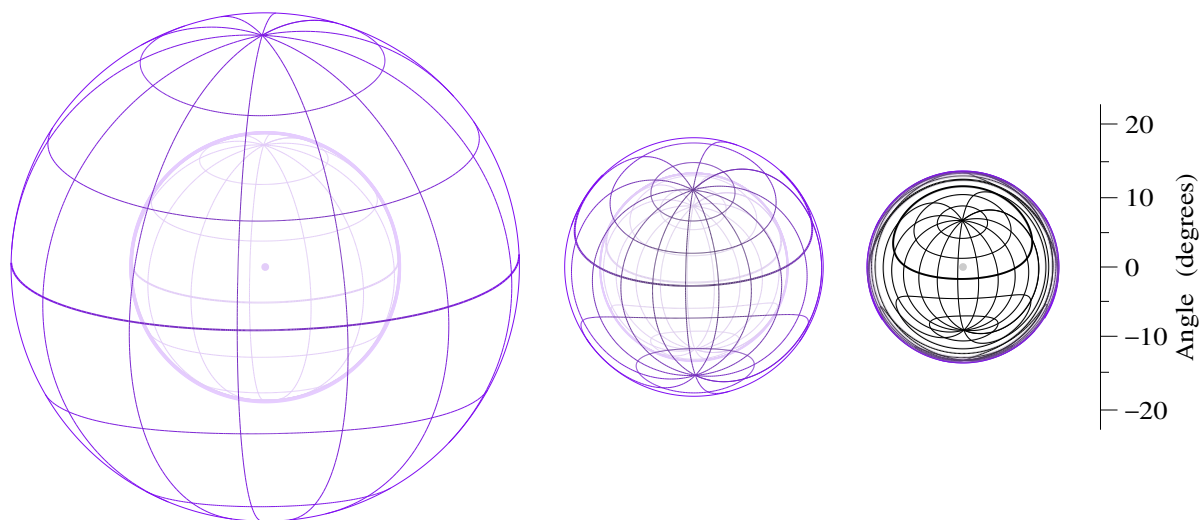


Figure 7.18 Three frames in the collapse of a uniform density, pressureless, spherical star from zero velocity at infinity (Oppenheimer and Snyder, 1939), as seen by an outside observer at rest at a radius of 10 Schwarzschild radii. The frames are spaced by 10 units of Schwarzschild time ($c = r_s = 1$). The star is made transparent, so you can see inside. Two layers are shown, one at the surface of the star, the other at half its radius. The centre of the star is shown as a dot. The frames are accurately ray-traced, and include the effect of the different light travel times from different parts of the star to the observer. As time goes by, from left to right, the collapsing star appears to freeze at the horizon, taking on the appearance of a Schwarzschild black hole. The different layers of the star appear to merge into one. The radius of the nearest point on the surface at the time of emission is 3.72, 1.50, and 1.01 Schwarzschild radii respectively.

collapsed with the star would find themselves falling uneventfully through the horizon to the central singularity in a finite proper time. How could these two perspectives be reconciled?

The solution is that the freezing at the horizon is an illusion. As pictured in Figure 7.2, space is falling at the speed of light at the horizon. Light emitted outward at the horizon just hangs there, barrelling at the speed of light through space that is falling at the speed of light. It takes an infinite time for light to lift off the horizon and make it to the outside world. The star really did collapse, but the infinite light travel time from the horizon gives the illusion that the star freezes at the horizon.

That radially outgoing light rays at the horizon remain on the horizon is apparent in the Penrose diagram, which shows the horizon as a null line, at 45° .

7.24 Apparent horizon

Since light can escape from the surface or interior of the collapsing star as long as it is even slightly larger than its Schwarzschild radius, it is possible to take the view that the horizon comes instantaneously into

being at the moment that the star collapses through its Schwarzschild radius. This definition of the horizon is called the apparent horizon. More generally, the **apparent horizon** is a null surface on which the congruence of light rays that form the surface are neither diverging nor converging. In spherically symmetric spacetimes, an apparent horizon is a place where radially moving null geodesics remain at rest in circumferential radius r ,

$$\frac{dr}{d\lambda} = 0 . \quad (7.91)$$

7.25 True horizon

An alternative definition of the horizon is to take it to be the boundary between outgoing null rays that fall into the black hole versus those that go to infinity. In any evolving situation, this definition of the horizon, which is called the **true horizon**, or **absolute horizon**, depends formally on what happens in the indefinite future, but in a slowly evolving system the absolute horizon can be located with some precision

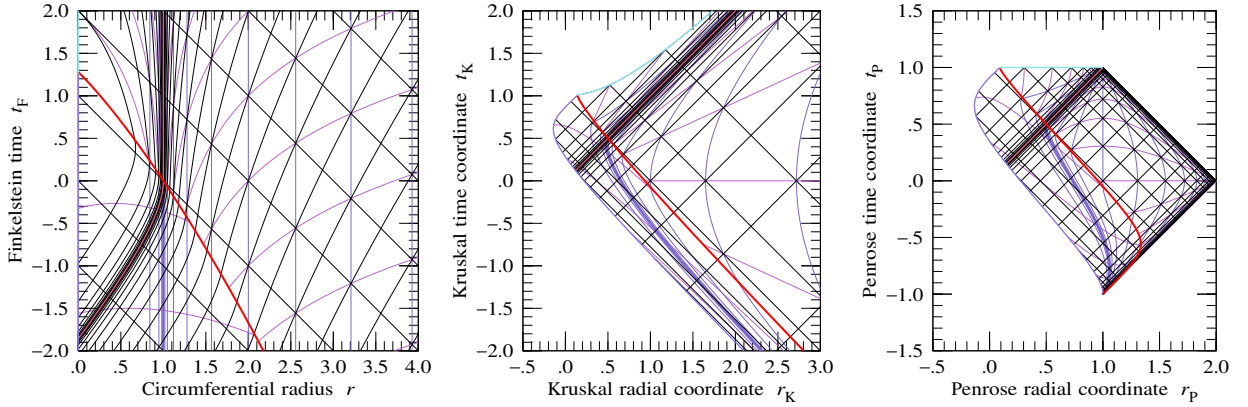


Figure 7.19 Finkelstein, Kruskal-Szekeres, and Penrose spacetime diagrams of the Oppenheimer-Snyder of a pressureless, spherical star. The thick (red) line is the surface of the collapsing star. The geometry outside the surface of the star is Schwarzschild, and the spacetime diagrams there look like those shown previously, Figures 7.7, 7.8, and 7.13. The geometry inside the surface of the star is that of a uniform density, pressureless Friedmann-Lemaître-Robertson-Walker universe. The lines of constant time (purple) are lines of constant Schwarzschild time outside the star's surface, and lines of constant FLRW time inside the star's surface. Lines of constant circumferential radius r (blue) are spaced uniformly in the tortoise coordinate r^* , equation (7.84), so before collapse appear bunched around the radius $r = r_s$ that after collapse becomes the horizon radius. The thick (pink) line at 45° in the Kruskal and Penrose diagrams is the true, or absolute, horizon, which divides the spacetime into a region where light rays are trapped, eventually falling to zero radius, and a region where light rays can escape to infinity. A singularity (cyan) forms when outgoing light rays can no longer escape from zero radius, which happens slightly before the surface of the collapsing star reaches zero radius.

without knowing the future. The true horizon is part of the future horizon of an observer who remains at a finite distance outside the black hole into the indefinite future.

Figure 7.19 shows Finkelstein, Kruskal, and Penrose spacetime diagrams of the Oppenheimer-Snyder collapse of a star to a Schwarzschild black hole. The diagrams show the freely-falling surface of the collapsing star, and the formation of the true horizon and of the singularity. The true horizon of the collapsing star forms before the star has collapsed, and grows to meet the apparent horizon as the star falls through its Schwarzschild radius. The central singularity forms slightly before the star has collapsed to zero radius. The formation of the singularity is marked by the fact that light rays emitted at zero radius cease to be able to move outward. In other words, the singularity forms when space starts to fall into it faster than light.

7.26 Penrose diagrams of Oppenheimer-Snyder collapse

Figure 7.20 shows a sequence of Penrose diagrams of Oppenheimer-Snyder collapse, progressing in time from left to right. The diagrams are drawn from the perspective of an observer before collapse on the left, to that of an observer after collapse on the right. The diagrams illustrate that, even though a Penrose diagram supposedly encompasses all of the spacetime, it crams most of the spacetime into a few boundary points, and the appearance of the diagram can vary dramatically depending on what part of the spacetime the diagram centres. In Figure 7.20, the Penrose diagram looks like Minkowski well before collapse, and like Schwarzschild well after collapse.

The Penrose diagrams in Figure 7.20 are drawn in the Penrose coordinates defined by equations (7.89) with the function $f(z)$ given by equation (7.90). Requiring the singularity to be horizontal, as is conventional, imposes that $f(z)$ be odd. Since other choices of $f(z)$ could be made, the shapes of the Penrose diagrams are not unique. However, other choices of smooth, monotonic, odd $f(z)$ give diagrams quite similar to those

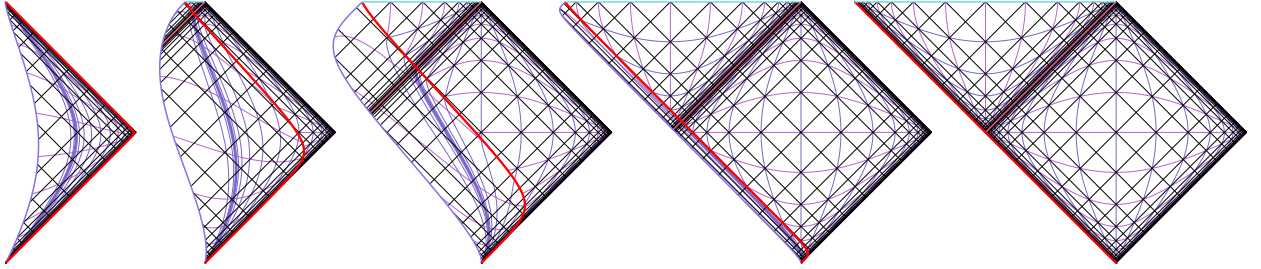


Figure 7.20 Sequence of Penrose diagrams illustrating the Oppenheimer-Snyder collapse of a pressureless, spherical star to a Schwarzschild black hole, progressing in time of collapse from left to right. On the left, the collapse is to the future of an observer at the centre of the diagram; on the right, the collapse is to the past of an observer at the centre of the diagram. The diagrams are at times -16 , -4 , 0 , 4 , and 16 Schwarzschild time units ($c = r_s = 1$) relative to the middle diagram. On the left the Penrose diagram resembles that of Minkowski space, while on the right the diagram resembles that of the Schwarzschild geometry. These Penrose diagrams are spacetime diagrams calculated in the Penrose coordinates defined by equations (7.89) and (7.90).

shown. In particular, as long as the singularity is chosen to be horizontal, it is impossible to arrange that the left edge of the diagram, defined by the centre of the collapsing star at $r = 0$, be vertical.

In the evolving Penrose diagram of Figure 7.20, spacetime appears to flow out of future infinity, the point at the top right of the diagram, down into past infinity, the point at the bottom of the diagram. Inside the horizon, as Schwarzschild time t goes by, spacetime appears to flow to the left, to the top left corner of the spacetime diagram. An infaller inside the horizon must of course follow a worldline at less than 45° from vertical. However, infallers who fall in at different times fall to different places on the spacelike singularity. From the perspective of an outside observer, infallers who fell in long ago are crammed to the top left corner of the Penrose diagram.

7.27 Illusory horizon

The simple Oppenheimer-Snyder model of stellar collapse shows that the antihorizon of the complete Schwarzschild geometry is replaced by the surface of the collapsing star, and that beyond the star's surface is not a parallel universe and a white hole, but merely the interior of the star, and the distant Universe glimpsed through the star's interior.

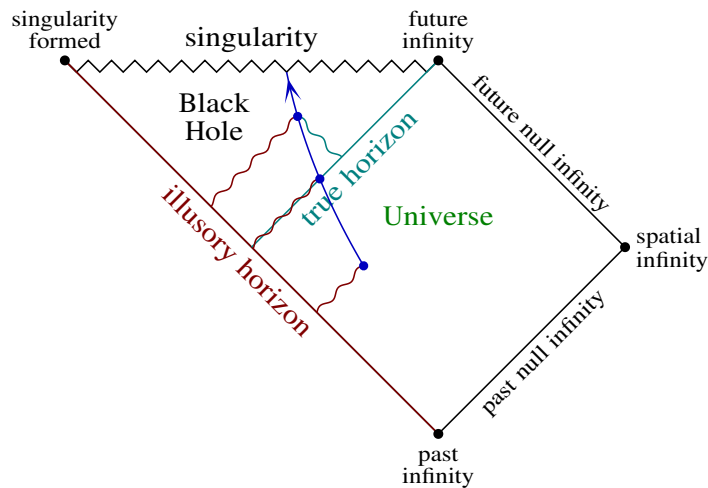


Figure 7.21 Penrose diagram of a collapsed spherical star at late times. The Penrose diagram looks essentially identical to the Penrose diagram 7.16 of the Schwarzschild geometry, except that the antihorizon is replaced by the illusory horizon. The wiggly lines show the paths of outgoing light rays from the illusory horizon, and ingoing light rays from the true horizon, as seen by an infaller who falls through the true horizon. An infaller looking directly towards the black hole sees the illusory horizon ahead of them, whether they are outside or inside the true horizon. The true horizon becomes visible to an infaller only after they have fallen through it. Once inside, the infaller sees the true horizon behind them, in the direction away from the black hole.

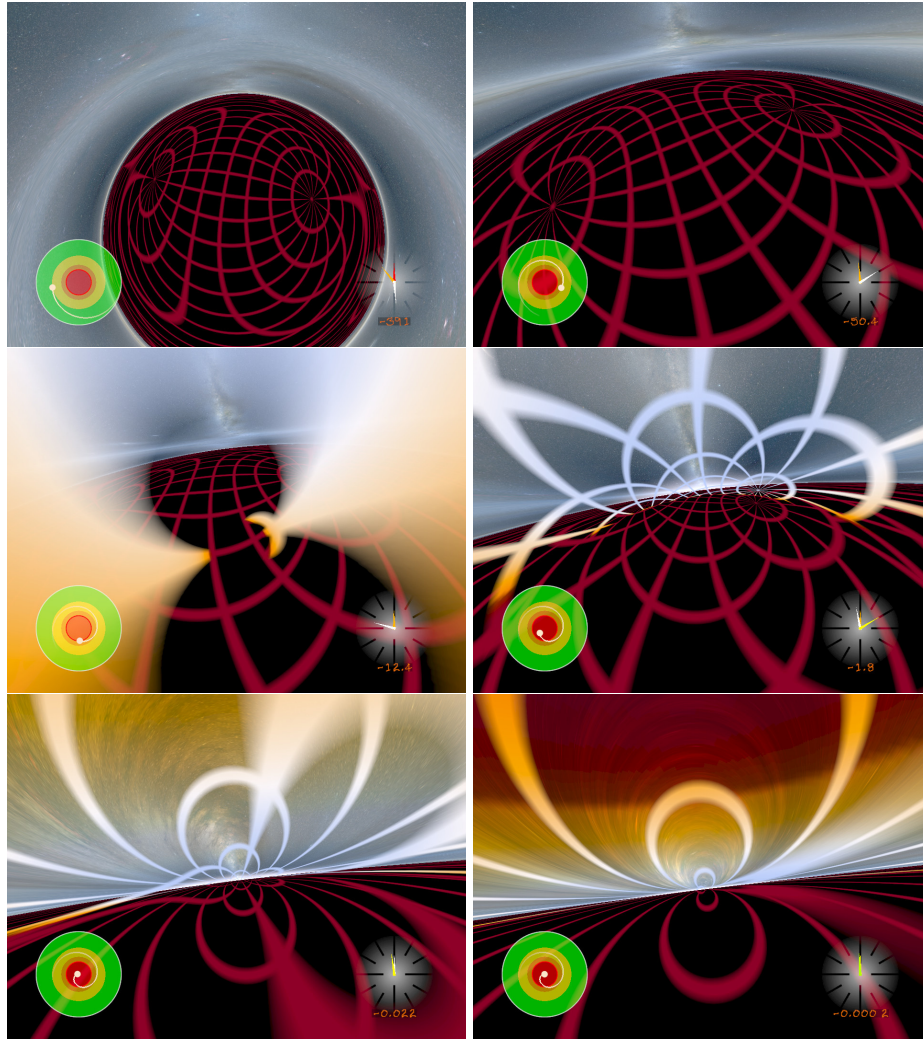


Figure 7.22 Six frames from a visualization of the view seen by an observer who free-falls into a Schwarzschild black hole. The infaller is on a geodesic with energy per unit mass $E = 1$, and angular momentum per unit mass $L = 1.96 r_s$. From left to right and top to bottom, the observer is at radii 3.008, 1.501, 0.987, 0.508, 0.102, and 0.0132 Schwarzschild radii. The illusory horizon is painted with a dark red grid, while the true horizon is painted with a grid coloured with an appropriately red- or blue-shifted blackbody colour. The schematic map at the lower left of each frame shows the trajectory (white line) of the observer through regions of stable circular orbits (green), unstable circular orbits (yellow), no circular orbits (orange), the horizon (red line), and inside the horizon (red). The clock at the lower right of each frame shows the proper time left to hit the singularity, in seconds, scaled to the mass $4 \times 10^6 M_\odot$ of the Milky Way's supermassive black hole (Ghez et al., 2005; Eisenhauer et al., 2005). The background is Axel Mellinger's Milky Way (Mellinger, 2009) (with permission).

As time goes by, the surface of the collapsing star becomes dimmer and more redshifted, taking on the appearance of the Schwarzschild antihorizon, Figure 7.18. The name **illusory horizon** for the exponentially dimming and redshifting surface was coined by Hamilton and Polhemus (2010). Figure 7.21 shows a Penrose diagram of a spherical collapsed star, with the true and illusory horizons marked. The Penrose diagram is just the limit of the sequence of the diagrams in Figure 7.20 from the perspective of an observer for whom the star collapsed long ago. The Penrose diagram 7.21 looks identical to the Penrose diagram of a Schwarzschild black hole, Figure 7.13, except that the antihorizon is replaced by the illusory horizon.

Unlike the antihorizon, the illusory horizon is not a future or past horizon, as defined by Hawking and Ellis (1973). As the Penrose diagrams 7.20 show, the illusory horizon is neither the boundary of the past lightcone of the future development of the worldline of any observer, nor the boundary of the future lightcone of the past development of the worldline of any observer.

An object similar to the illusory horizon, the **stretched horizon**, was introduced by Susskind, Thorlacius, and Uglum (1993). The stretched horizon was conceived as the place where, from the perspective of an outside observer, Hawking radiation comes from, and the place where, from the perspective of an outside observer, the interior quantum states of a black hole reside. The stretched horizon was argued to be located on a spacelike surface one Planck area above the true horizon. However, the restriction to an outside observer is too limiting, and the notion that the stretched horizon lives literally just above the true horizon has been a source of confusion in the theoretical physics literature. If you go down to the true horizon, you do not encounter the putative stretched horizon. The stretched horizon is an illusion, a mirage. Better call it the illusory horizon.

Figure 7.22 shows six frames from a visualization (Hamilton and Polhemus, 2010) of the appearance of a Schwarzschild black hole and its true and illusory horizons as perceived by an observer who free-falls through the true horizon. The illusory horizon, the exponentially redshifting image of the long-ago collapsed star, is painted with a dark red grid, as befits its dimmed, redshifted appearance. The true horizon is painted with blackbody colours blueshifted or redshifted according to the shift that the infalling observer would see on an emitter free-falling radially through the true horizon from zero velocity at infinity. When an infaller falls through the true horizon, they do not catch up with the illusory horizon, the image of the collapsed star, which remains ahead of them. The visualization gives the impression that the illusory horizon is a finite distance ahead of the infaller, and this impression is correct: the affine distance between the illusory horizon and an infaller at the true horizon is finite, not zero. Calculation of what an infaller sees involves working in the locally inertial frame (tetrad) of the infaller, so is deferred until after tetrads.

An infaller does not encounter the illusory horizon at the true horizon, but, as illustrated by the visualization 7.22, they do have the impression of encountering the illusory horizon at the singularity. The affine distance between the infaller and the illusory horizon tends to zero at the singularity.

7.28 Collapse of a shell of matter on to a black hole

The antihorizon of a Schwarzschild black hole is located at the horizon radius, one Schwarzschild radius. Where is the illusory horizon located? From the perspective of an observer watching a spherical black hole

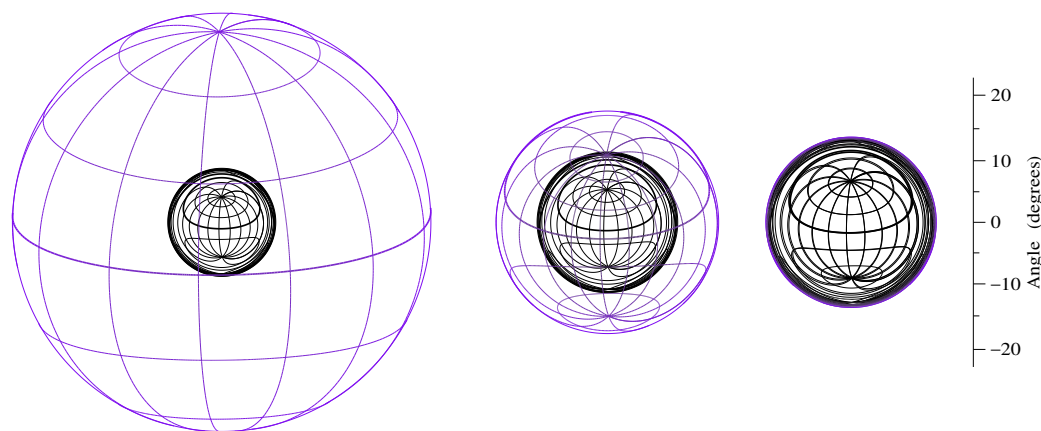


Figure 7.23 Three frames in the collapse of a thin spherical shell of matter on to a pre-existing Schwarzschild black hole, as seen by an outside observer at rest at a radius of 10 Schwarzschild radii (Schwarzschild radius of the final black hole). The frames are spaced by 10 units of Schwarzschild time ($c = r_s = 1$). The shell has the same mass as the original black hole, so the black hole doubles in mass from beginning to end. During the collapse, the horizon of the pre-existing black hole appears to expand outward, in due course reaching the size of the new black hole. The expansion of the image of the pre-existing black hole is caused by gravitational lensing by the shell.

that collapsed from a star long ago, the illusory horizon appears to be located at (exponentially close to) the antihorizon of the Schwarzschild black hole of the same mass.

What happens to the illusory horizon if the black hole accretes mass, and grows larger? Figure 7.23 shows three frames in the collapse of a thin spherical shell of pressureless matter on to a pre-existing black hole, Exercise 20.7. The shell collapses from zero velocity at infinity. As usual in this book, the frames are accurately ray-traced. The shell of matter here has the same mass as the pre-existing black hole, so the black hole doubles in mass as the shell collapses on to it. The visualization shows that the illusory horizon of the pre-existing black hole expands to meet the infalling shell of matter. The apparent expansion is caused by gravitational lensing of the pre-existing black hole by the shell. As time goes by, the shell appears to merge with the horizon of the pre-existing black hole. The merged shell and expanded horizon take on the appearance of the antihorizon of a Schwarzschild black hole of twice the original mass.

Figure 7.24 shows a Finkelstein spacetime diagram of the collapse of the shell of matter on to the black hole. The initial black hole has half the mass of the final black hole. The initial apparent horizon at $0.5r_s$, half the Schwarzschild radius of the final black hole, follows a null geodesic until the infalling shell hits it. The shell deflects the null geodesic, which falls to the central singularity. The true horizon follows a null geodesic that joins continuously with the apparent horizon of the final black hole.

Concept question 7.14 Penrose diagram of a thin spherical shell collapsing on to a Schwarzschild black hole. Sketch a Penrose diagram of a thin spherical shell collapsing on to a pre-existing Schwarzschild black hole.

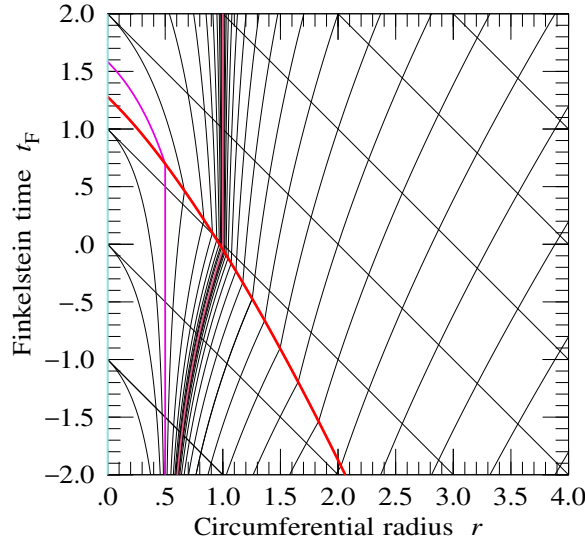


Figure 7.24 Finkelstein spacetime diagram of a thin spherical shell of matter collapsing on to a pre-existing Schwarzschild black hole, in units r_s of the Schwarzschild radius of the final black hole. The mass of the (red) shell equals that of the pre-existing black hole, so the black hole doubles in mass as a result of accreting the shell. Whereas the apparent horizon jumps discontinuously from $0.5r_s$ to $1r_s$ at the shell boundary, the true horizon increases continuously. The mathematics governing a thin spherical shell is addressed in Exercise 20.7.

schild black hole. Where are the apparent and true horizons? **Answer.** The Penrose diagram looks essentially the same as Figure 7.13 (differing in that lines of constant time and radius are different inside the shell). The apparent horizon before collapse follows an outgoing null (45°) line that hits the singularity inside the true horizon, consistent with the Finkelstein diagram 7.24.

7.29 The illusory horizon and black hole thermodynamics

As will be discussed later in this book, the illusory horizon plays a central role in the thermodynamics of black holes. The illusory horizon is the source of Hawking radiation, for observers both outside and inside the true horizon. If, as proposed by Susskind, Thorlacius, and Uglum (1993), there is a holographic mapping between the interior quantum states of a black hole and its horizon, then that holographic mapping must be to the illusory horizon, for observers both outside and inside the true horizon.

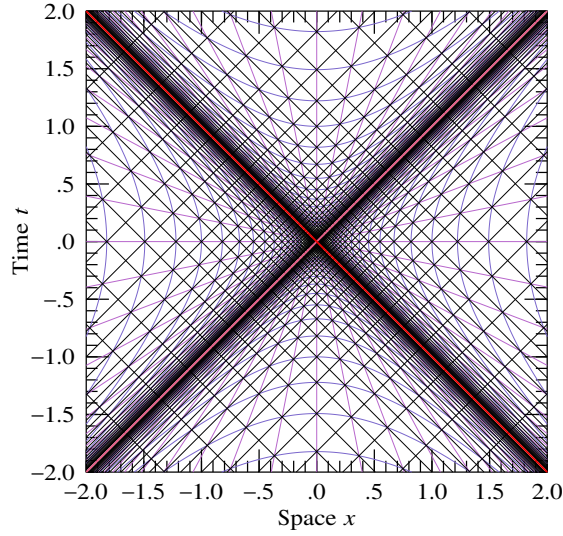


Figure 7.25 Rindler diagram, which is a Minkowski spacetime diagram showing lines of constant Rindler coordinates α and l , equations (7.92) and (7.94). The Rindler lines are uniformly spaced by 0.2 in α and $\ln l$. The spacetime diagram resembles that of the analytically extended Schwarzschild geometry in Kruskal coordinates, Figure 7.11. The null lines passing through the origin constitute future (line from lower left to upper right) and past (line from lower right to upper left) horizons for Rindler observers in the right quadrant.

7.30 Rindler space and Rindler horizons

Rindler space is Minkowski space expressed in the coordinates of, and as experienced by, a system of uniformly accelerating observers, called Rindler observers. A Rindler observer who accelerates uniformly in their own frame with proper acceleration $1/l$, passing through position $\{t, x\} = \{0, l\}$, follows a worldline in Minkowski space

$$\{t, x\} = l \{\sinh \alpha, \cosh \alpha\} \quad (7.92)$$

with fixed l and varying α . The Rindler observer's worldline follows a point on the rim of the rotating spacetime wheel, §1.8.2. The Rindler line-element is the Minkowski line-element expressed in Rindler coordinates $\{\alpha, l, y, z\}$, Exercise 2.11,

$$ds^2 = -l^2 d\alpha^2 + dl^2 + dy^2 + dz^2. \quad (7.93)$$

Despite the fact that Rindler spacetime is Minkowski spacetime in disguise, it nevertheless resembles Schwarzschild spacetime in that, from the perspective of Rindler observers, Rindler space contains horizons. Moreover Rindler observers are expected to see Hawking radiation, which in this context is called Unruh (1976) radiation, §43.1.

Figure 7.25 shows a Rindler diagram, a spacetime time diagram of Minkowski space, drawn in standard Minkowski coordinates t and x , showing lines of constant Rindler coordinates α and l . The Rindler spacelike

coordinate l is positive in the right quadrant, negative in the left quadrant. The Rindler coordinate vanishes, $l = 0$, at the boundaries of the right and left quadrants, which form the null lines at 45° passing through the origin in the Rindler diagram 7.25. The Rindler metric (7.93) has a coordinate singularity at $l = 0$. In the upper and lower quadrants, the Rindler coordinate l switches from being spacelike to timelike ($dl^2 < 0$). Rindler coordinates in the upper and lower quadrants are defined by

$$\{t, x\} = l \{\cosh \alpha, \sinh \alpha\} , \quad (7.94)$$

where the timelike coordinate l is positive in the upper quadrant, negative in the lower quadrant.

The null (45°) lines passing through the origin in Figure 7.25 are future and past horizons for Rindler observers in the right quadrant of the Rindler diagram. A Rindler observer following a worldline (7.92) in the right quadrant never gets to see the part of spacetime to the future of the null surface $x = t$, which therefore constitutes a future horizon for the Rindler observer. The same Rindler observer can never send a signal into the part of spacetime to the past of the null surface $x = -t$, which therefore constitutes a past horizon, an antihorizon, for the Rindler observer.

The Rindler diagram 7.25 resembles the Kruskal diagram 7.11 of the analytically extended Schwarzschild geometry, albeit without singularities. The Minkowski coordinates t and x are analogues of the Kruskal coordinates t_K and r_K , while the Rindler coordinates α and l are analogues of the Schwarzschild coordinates t and r . The Schwarzschild and Rindler time coordinates t and α are both Killing coordinates, §7.32. Lines of constant Schwarzschild and Rindler time t and α follow straight lines in the corresponding Kruskal and Rindler diagrams, Figures 7.11 and 7.25. The Schwarzschild and Rindler spatial coordinates r and l are spacelike in the right and left quadrants, timelike in the upper and lower quadrants.

7.30.1 Penrose diagram of Rindler space

Figure 7.26 is a Penrose diagram of Rindler space. This is just a Penrose diagram of Minkowski space showing lines of constant Rindler coordinates α and l . Penrose time and space coordinates t_P and x_P can be defined by any conformal transformation

$$t_P \pm x_P \equiv f(t \pm x) \quad (7.95)$$

for which $f(z)$ is finite at $z \rightarrow \pm\infty$. The Rindler lines acquire a symmetrical appearance on the Penrose diagram provided that the conformal function $f(z)$ is chosen to satisfy $f(z) + f(-z) = \text{constant}$. For the Penrose diagram in Figure 7.26, the conformal function $f(z)$ is

$$f(z) \equiv \text{sign}(z) + \frac{2}{\pi} \text{atan} \left(\frac{z - z^{-1}}{2} \right) . \quad (7.96)$$

The choice (7.96) is inspired by the form (10.186) of the coordinates that gives the Penrose diagram of de Sitter space a symmetrical appearance. The Penrose diagram 7.26 resembles that of the analytically extended Schwarzschild geometry, Figure 7.15, but without singularities.

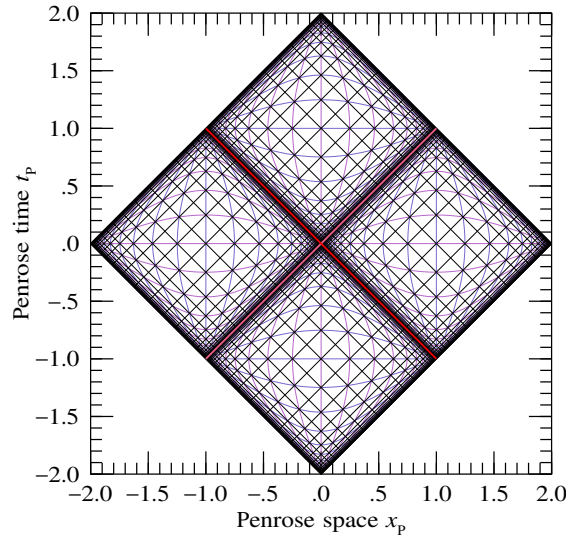


Figure 7.26 Penrose diagram of Rindler space. This is the Penrose diagram of Minkowski space corresponding to the Rindler diagram 7.25. The Penrose coordinates t_p and x_p are related to Minkowski coordinates t and x by equations (7.95). The Rindler lines are uniformly spaced by 0.4 in α and $\ln l$. The Penrose diagram resembles that of the analytically extended Schwarzschild geometry, Figure 7.15, but without singularities.

Concept question 7.15 Spherical Rindler space. The Rindler line-element (7.93) is plane-parallel, with all the Rindler observers accelerating in the x -direction. Would not a better analogue of a spherical black hole be the spherically symmetric Rindler line-element

$$ds^2 = -r_R^2 d\alpha^2 + dr_R^2 + r^2 d\phi^2, \quad (7.97)$$

where all Rindler observers accelerate in the radial direction with $\{t, r\} = r_R \{\sinh \alpha, \cosh \alpha\}$? **Answer.** The spherical Rindler line-element (7.97) is indeed a viable line-element. However, it does not provide a better analogue of a spherical black hole because the past and future horizons of a Rindler observer accelerating in, say, the x -direction are flat surfaces at $x \pm t = 0$, not spherical surfaces at $r \pm t = 0$.

7.31 Rindler observers who start at rest, then accelerate

Rindler space provides an analogue of the analytically extended Schwarzschild geometry. But a spherical black hole formed from the collapse of a star is not described by the analytically extended geometry. Rather, the analytic extension through the antihorizon is replaced by the interior of the collapsed star.

A Rindler analogue of a black hole that forms from the collapse of a star is obtained by considering a

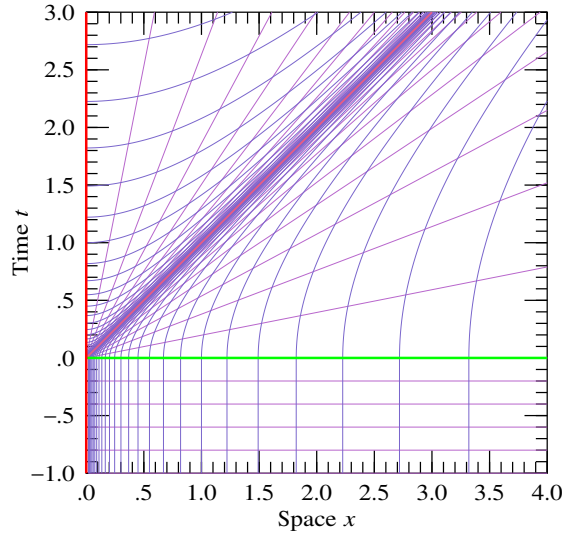


Figure 7.27 Minkowski spacetime diagram, showing worldlines of observers who start at rest, then begin accelerating uniformly, as Rindler observers, at $t = 0$. At $t \leq 0$, the lines are lines of constant Minkowski time and space t and x , while at $t \geq 0$, the lines are lines of constant Rindler time and space α and l , equations (7.92) and (7.94). The Rindler lines are uniformly spaced by 0.2 in α and $\ln l$. The null line starting at the origin $\{t, x\} = \{0, 0\}$ extending upward at 45° from vertical is a future horizon for the Rindler observers.

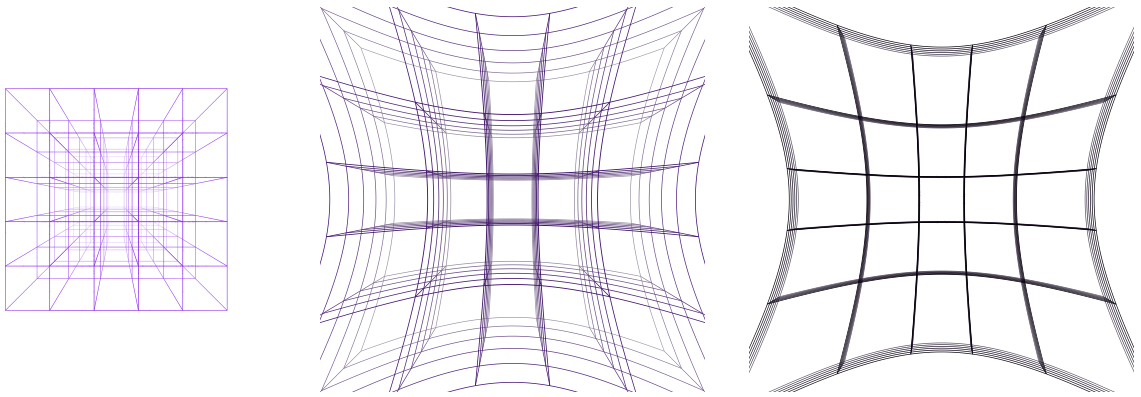


Figure 7.28 Three frames in the appearance of Minkowski space as seen by a uniformly accelerating observer, a Rindler observer. Minkowski space is represented by a unit box at rest, centred at the origin. The box is drawn as a $5 \times 5 \times 5$ lattice. The Rindler observer starts at rest at unit distance from the origin, and watches rearward while accelerating at unit acceleration away from the box. The field of view is 120° across the horizontal. The frames increase in time from left to right, and are at 0, 2, and 4 units of proper time after the Rindler observer begins accelerating. As time goes by, the lattice appears to freeze towards a two-dimensional surface, the illusory Rindler horizon.

system of Rindler observers who are initially at rest, and begin accelerating only at some time $t = 0$. The situation is illustrated in the spacetime diagram shown in Figure 7.27. This diagram is similar to the Rindler diagram 7.25, except that the Rindler observers start accelerating at $t = 0$ instead of having been accelerating into the indefinite past. Just as a black hole formed from the collapse of a star has a future horizon but no past horizon, so also the Rindler space of Rindler observers who start at rest contains a future horizon but no past horizon.

Despite having no past horizon, a Rindler observer who starts from rest sees an illusory horizon form, Figure 7.28, in much the same way that an observer watching a star collapse to a black hole sees an illusory horizon form, Figure 7.18. The illusory horizon is the exponentially dimming and redshifting image of Minkowski space around the Rindler observer. Figure 7.28 shows three frames in the appearance of a portion of Minkowski space as seen by a Rindler observer watching rearward. As time goes by, Minkowski space appears to compress and freeze toward a surface, the illusory horizon. The Rindler observer sees the illusory horizon dim and redshift exponentially. Exercise 7.16 quantifies the appearance of the Rindler illusory horizon, which forms a hyperbola around the Rindler observer, with the Rindler observer at its focus.

7.31.1 Penrose diagram of Rindler observers who start at rest, then accelerate

Figure 7.29 shows a sequence of Penrose diagrams drawn from the perspective of Rindler observers who start at rest and begin to accelerate at time $t = 0$, as in the spacetime diagram 7.27. These Penrose diagrams are

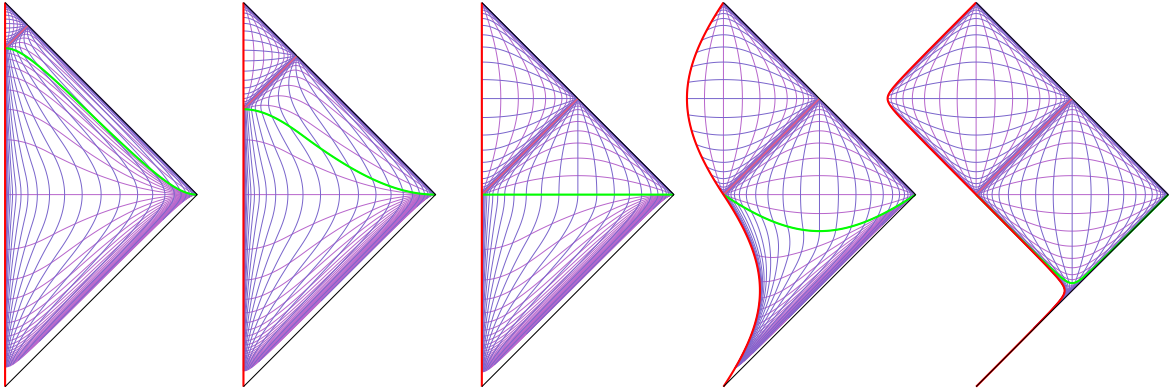


Figure 7.29 Sequence of Penrose diagrams of the Minkowski space shown in Figure 7.27, progressing in time from left to right. The left edge of each diagram is the surface at $x = 0$. The diagrams at left are in the frames of observers who are at rest relative to each other. In the middle diagram, the observers start to accelerate as Rindler observers. The diagrams at right are in the frames of the Rindler observers, which become progressively more Lorentz boosted compared to the rest frame. The diagrams are at times -8 , -2 , 0 , 2 , and 8 units of proper time of the observer who is initially at rest at unit distance ($x = 1$ in Figure 7.27) from the origin. The Rindler lines are uniformly spaced by 0.4 in α and $\ln l$. This sequence of Penrose diagrams resembles that of the Oppenheimer-Snyder collapse of a star shown in Figure 7.20.

calculated, not sketched, with Penrose coordinates given by equations (7.95). The left edge of each diagram is the surface at $x = 0$. This sequence resembles the sequence of Penrose diagrams of Oppenheimer-Snyder collapse of a star to a black hole, Figure 7.20, except that there is no singularity.

At left, before the observers start to accelerate, the Penrose diagram looks like that of Minkowski space. The Rindler portion of the spacetime (the part above the green line) is crammed along the top right edge of the Penrose diagram. At right, after the Rindler observers have started to accelerate, the Penrose diagram is tilted by the Lorentz boost of the Minkowski space. The Minkowski portion of the spacetime (the part below the green line) crams towards the bottom right edge of the diagram.

Aren't the Penrose diagrams in Figure 7.29 misleading because they omit the spacetime to the left of the diagrams, at $x < 0$? Since Rindler observers are confined to the right quadrant of Rindler space, they never get to see the region beyond their future horizon. Therefore there is no loss of generality to draw the Minkowski spacetime diagram 7.27 with reflection symmetry about $x = 0$. Applied to the Penrose diagrams 7.29, reflection symmetry means that light that passes that passes from $x < 0$ to $x > 0$ can be considered to “bounce” at 45° off the left edge of the diagram at $x = 0$. Whatever the case, as seen in Exercise 7.16, light emitted from $x < 0$ appears to a Rindler observer asymptotically to dim, redshift, and freeze at the observer's illusory horizon.

Exercise 7.16 Rindler illusory horizon. The purpose of this problem is to figure out the appearance to a Rindler observer of their illusory horizon. For simplicity, choose time units such that the Rindler observer accelerates with unit acceleration. The coordinates $\{x, y, z\}$ are spatial coordinates in Minkowski space. Starting from rest on the x -axis at position $x = 1$, the Rindler observer accelerates in the positive x -direction, reaching position $x = x_0$ in the rest frame. After a sufficiently long Rindler proper time α , the position $x_0 = \cosh \alpha$ is large.

1. **Shape.** Show that points $\{x, y, z\}$ that are close to the origin, in the sense of satisfying $|x| \ll x_0$ and $\sqrt{y^2 + z^2} \ll x_0$, appear to a Rindler observer to freeze towards a time-independent surface $\{l, y, z\}$, the illusory horizon, satisfying

$$l = \frac{1}{2}(y^2 + z^2 - 1) . \quad (7.98)$$

The Rindler observer sees their illusory horizon as a parabola with themselves at the focus, the origin.

2. **Redshift.** Show further that the Rindler observer sees points on the illusory horizon redshifting exponentially, at rate e^α .

Solution.

1. **Shape.** In the Minkowski rest frame, a spatial point $\{x, y, z\}$ relative to an observer at $\{x_0, 0, 0\}$ is at position $\{x - x_0, y, z\}$. If the observer is moving at velocity v in the x -direction, then according to the rules of 4-dimensional perspective, §1.13.2, the point appears in the observer's frame to lie at position $\{l, y, z\}$ with transverse coordinates y, z unchanged, and l given by

$$l = \gamma(x - x_0) + \gamma v \sqrt{(x - x_0)^2 + y^2 + z^2} , \quad (7.99)$$

where $\gamma = 1/\sqrt{1 - v^2}$ is the Lorentz gamma factor. Points near the origin, with $|x| < x_0$, are behind the

observer, satisfying $x - x_0 < 0$. Thus equation (7.99) factors to

$$l = \gamma(x - x_0) \left[1 - v \sqrt{1 + (y^2 + z^2)/(x - x_0)^2} \right] , \quad (7.100)$$

which rearranges to

$$\begin{aligned} l &= \frac{\gamma(x - x_0)[1 - v^2 - v^2(y^2 + z^2)/(x - x_0)^2]}{1 + v^2 \sqrt{1 + (y^2 + z^2)/(x - x_0)^2}} \\ &= \frac{1}{1 + v^2 \sqrt{1 + (y^2 + z^2)/(x - x_0)^2}} \left[\frac{x - x_0}{\gamma} - \frac{v^2(y^2 + z^2)\gamma}{x - x_0} \right] . \end{aligned} \quad (7.101)$$

For a Rindler observer, the position x_0 is just equal to the Lorentz gamma factor, $x_0 = \cosh \alpha = \gamma$. Under the conditions $x_0 \equiv \gamma \gg 1$, along with $x_0 \gg |x|$ and $x_0 \gg \sqrt{y^2 + z^2}$, equation (7.101) reduces to

$$l \approx -\frac{1}{2} + \frac{1}{2}(y^2 + z^2) , \quad (7.102)$$

yielding equation (7.98) as claimed.

2. **Redshift.** According to the rules of 4-dimensional perspective, §1.13.2, the redshift factor, the ratio $E_{\text{em}}/E_{\text{obs}}$ of emitted to observed photon energies from a point, equals the ratio of the emitted to observed distances to the point,

$$\frac{E_{\text{em}}}{E_{\text{obs}}} = \frac{\sqrt{(x - x_0)^2 + y^2 + z^2}}{\sqrt{l^2 + y^2 + z^2}} . \quad (7.103)$$

A point $\{l, y, z\}$ on the Rindler observer's illusory horizon appears fixed to the observer, l satisfying equation (7.102). The only quantity on the right hand side of equation (7.103) that varies with the Rindler observer's time α is x_0 . Under the conditions $x_0 \equiv \cosh \alpha \gg 1$, along with $x_0 \gg |x|$ and $x_0 \gg \sqrt{y^2 + z^2}$, the redshift factor satisfies

$$\frac{E_{\text{em}}}{E_{\text{obs}}} \lesssim x_0 \lesssim e^\alpha . \quad (7.104)$$

The redshift factor of a point on the Rindler observer's illusory horizon thus increases exponentially with Rindler time α .

Exercise 7.17 Area of the Rindler horizon. What is the area of a Rindler observer's horizon?

Solution. The area of the Rindler horizon is the area of the spatial yz plane orthogonal to the Rindler observer's boost plane tx . For a Rindler observer who starts accelerating at a finite time, the illusory horizon after α acceleration times is well-formed only over a region of size $\sqrt{y^2 + z^2} \lesssim e^\alpha$ about the origin. Thus the area of the illusory Rindler horizon is of order $\sim e^{2\alpha}$.

Exercise 7.18 Coordinates adapted to an observer accelerating in Minkowski space. Consider an observer in Minkowski space who is accelerating arbitrarily in the x -direction.

1. Let the rapidity of the observer in the Minkowski frame be α . The 4-velocity $u^m \equiv dx^m/d\tau$ of the observer in the Minkowski frame is then

$$u^m = \{\cosh \alpha, \sinh \alpha, 0, 0\} . \quad (7.105)$$

Argue that the proper acceleration $g(\tau)$ of the observer in their own frame is

$$g = \frac{d\alpha}{d\tau} . \quad (7.106)$$

2. Choose a spatial coordinate l such that at any instant τ of the observer's proper time, the coordinates $\{\tau, l\}$ comprise locally inertial coordinates in the frame of the observer, with $l = 0$ at the position of the observer. The Minkowski coordinates $\{t, x\}$ are related to the observer's locally inertial coordinates $\{\tau, l\}$ by

$$t(\tau, l) = t(\tau, 0) + l \sinh \alpha , \quad (7.107a)$$

$$x(\tau, l) = x(\tau, 0) + l \cosh \alpha . \quad (7.107b)$$

Show that the line-element in the coordinates $\{\tau, l, y, z\}$ is

$$ds^2 = -(1 + gl)^2 d\tau^2 + dl^2 + dy^2 + dz^2 . \quad (7.108)$$

7.32 Killing vectors

The Schwarzschild metric presents an opportunity to introduce the concept of **Killing vectors** (after Wilhelm Killing, not because the vectors kill things, though the latter is true), which are associated with symmetries of the spacetime. The flow through spacetime of the Killing vectors associate with a symmetry is called the **Killing vector field**. A coordinate that is constant along the flow lines of a Killing vector field is called a **Killing coordinate**.

7.32.1 Time translation symmetry

The time translation invariance of the Schwarzschild geometry is evident from the fact that the metric is independent of the Schwarzschild time coordinate t . Equivalently, the partial time derivative $\partial/\partial t$ of the Schwarzschild metric is zero. The associated Killing vector ξ^μ at each point of the spacetime is then defined by

$$\xi^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial t} , \quad (7.109)$$

so that in Schwarzschild coordinates $\{t, r, \theta, \phi\}$

$$\xi^\mu = \{1, 0, 0, 0\} . \quad (7.110)$$

In coordinate-independent notation, the Killing vector is

$$\boldsymbol{\xi} = e_\mu \xi^\mu = e_t . \quad (7.111)$$

The Schwarzschild time coordinate t is a Killing coordinate.

This may seem like overkill—couldn't one just say that the metric is independent of time t and be done

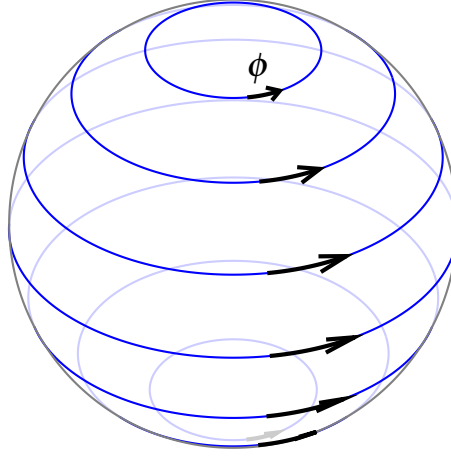


Figure 7.30 The Killing vector field associated with rotation of a 2-sphere about an axis.

with it? The answer is that symmetries are not always evident from the metric, as will be seen in the next section 7.32.2.

Because the Killing vector $\mathbf{e}_t$ is the unique timelike Killing vector of the Schwarzschild geometry, it has a definite meaning independent of the coordinate system. It follows that its scalar product with itself is a coordinate-independent scalar

$$\xi_\mu \xi^\mu = \mathbf{e}_t \cdot \mathbf{e}_t = g_{tt} = -\left(1 - \frac{2M}{r}\right). \quad (7.112)$$

In curved spacetimes, it is important to be able to identify scalars, which have a physical meaning independent of the choice of coordinates.

7.32.2 Spherical symmetry

The azimuthal rotational symmetry of the Schwarzschild metric is evident from the fact that the metric is independent of the azimuthal coordinate ϕ , implying that ϕ is a Killing coordinate. The associated Killing vector at each point of the spacetime is

$$\boldsymbol{\xi} = \mathbf{e}_\phi \quad (7.113)$$

with components $\{0, 0, 0, 1\}$ in Schwarzschild coordinates $\{t, r, \theta, \phi\}$. Figure 7.30 illustrates the Killing vector field corresponding to the azimuthal rotational symmetry.

The Schwarzschild metric is fully spherically symmetric, not just azimuthally symmetric. Since the 3D rotation group $O(3)$ is 3-dimensional, it is to be expected that there are three Killing vectors. You may recognize from quantum mechanics that $\partial/\partial\phi$ is (modulo factors of i and $\hbar$) the z -component of the angular momentum operator $\mathbf{L} = \{L_x, L_y, L_z\}$ in a coordinate system where the azimuthal axis is the z -axis. The 3

components of the angular momentum operator are given by:

$$iL_x = y \frac{\partial}{\partial z} - z \frac{\partial}{\partial y} = -\sin \phi \frac{\partial}{\partial \theta} - \cot \theta \cos \phi \frac{\partial}{\partial \phi} , \quad (7.114a)$$

$$iL_y = z \frac{\partial}{\partial x} - x \frac{\partial}{\partial z} = \cos \phi \frac{\partial}{\partial \theta} - \cot \theta \sin \phi \frac{\partial}{\partial \phi} , \quad (7.114b)$$

$$iL_z = x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} = \frac{\partial}{\partial \phi} . \quad (7.114c)$$

The 3 rotational Killing vectors are correspondingly:

$$\text{rotation about } x\text{-axis: } -\sin \phi \mathbf{e}_\theta - \cot \theta \cos \phi \mathbf{e}_\phi , \quad (7.115a)$$

$$\text{rotation about } y\text{-axis: } \cos \phi \mathbf{e}_\theta - \cot \theta \sin \phi \mathbf{e}_\phi , \quad (7.115b)$$

$$\text{rotation about } z\text{-axis: } \mathbf{e}_\phi . \quad (7.115c)$$

The 3 Killing vectors span the 2-dimensional surface of the unit sphere, and are therefore not linearly independent. Specifically, they satisfy

$$xL_x + yL_y + zL_z = 0 . \quad (7.116)$$

Note that although a linear combination of Killing vectors with constant coefficients is a Killing vector, a linear combination with non-constant coefficients is *not* necessarily a Killing vector.

You can check that the action of the x and y rotational Killing vectors on the metric does *not* kill the metric. For example, $iL_x g_{\phi\phi} = -2r^2 \sin \phi \sin \theta \cos \theta$ does not vanish. This example shows that a more powerful and general condition, described in the next section 7.32.3, is needed to establish whether a quantity is or is not a Killing vector.

Because spherical symmetry does not define a unique azimuthal axis $\mathbf{e}_\phi$, its scalar product with itself $\mathbf{e}_\phi \cdot \mathbf{e}_\phi = g_{\phi\phi} = -r^2 \sin^2 \theta$ is *not* a coordinate-invariant scalar. However, the sum of the scalar products of the 3 rotational Killing vectors is rotationally invariant, and is therefore a coordinate-invariant scalar

$$(-\sin \phi \mathbf{e}_\theta - \cot \theta \cos \phi \mathbf{e}_\phi)^2 + (\cos \phi \mathbf{e}_\theta - \cot \theta \sin \phi \mathbf{e}_\phi)^2 + \mathbf{e}_\phi^2 = g_{\theta\theta} + (\cot^2 \theta + 1)g_{\phi\phi} = -2r^2 . \quad (7.117)$$

This shows that the circumferential radius r is a scalar, as you would expect.

7.32.3 Killing equation

As seen in the previous section, a Killing vector does not always kill the metric in a given coordinate system. This is not really surprising given the arbitrariness of coordinates in general relativity. What is true is that a quantity is a Killing vector if and only if there exists a coordinate system (possibly in patches) such that the Killing vector kills the metric in that system.

Suppose that in some coordinate system the metric is independent of the coordinate ϕ . Then the covariant ϕ -momentum p_ϕ of a particle along a geodesic is a constant of motion, equation (4.50),

$$p_\phi = \text{constant} . \quad (7.118)$$

Equivalently

$$\xi^\nu p_\nu = \text{constant} , \quad (7.119)$$

where ξ^ν is the associated Killing vector, whose only non-zero component is $\xi^\phi = 1$ in this particular coordinate system. The converse is also true: if $\xi^\nu p_\nu = \text{constant}$ along all geodesics, then ξ^ν is a Killing vector. The constancy of $\xi^\nu p_\nu$ along all geodesics is equivalent to the condition that its affine derivative vanish along all geodesics

$$\frac{d\xi^\nu p_\nu}{d\lambda} = 0 . \quad (7.120)$$

But this is equivalent to

$$0 = p^\mu \mathring{D}_\mu (\xi^\nu p_\nu) = p^\mu p^\nu \mathring{D}_\mu \xi_\nu = \frac{1}{2} p^\mu p^\nu (\mathring{D}_\mu \xi_\nu + \mathring{D}_\nu \xi_\mu) , \quad (7.121)$$

the $\circ$ atop $\mathring{D}_\mu$ serving as a reminder that this is the torsion-free covariant derivative, §2.12. The second equality of equations (7.121) follows from the geodesic equation, $p^\mu \mathring{D}_\mu p_\nu = 0$, and the last equality is true because of the symmetry of $p^\mu p^\nu$ in $\mu \leftrightarrow \nu$. A necessary and sufficient condition for equation (7.121) to be true for all geodesics is that

$$\boxed{\mathring{D}_{(\mu} \xi_{\nu)} = 0} , \quad (7.122)$$

which is **Killing's equation**. This equation is the desired necessary and sufficient condition for ξ^ν to be a Killing vector. It is a generally covariant equation, valid in any coordinate system. Equation (7.122) can also be written as the statement that the Lie derivative of the metric, equation (7.160), along the Killing direction ξ^ν vanishes,

$$\mathcal{L}_\xi g_{\mu\nu} = 0 . \quad (7.123)$$

7.32.4 Conformal Killing vector

Sometimes a spacetime has a weaker **conformal symmetry** in which, instead of the metric being independent of a coordinate (in some system of coordinates), the metric depends on a coordinate ϕ only through an overall scaling, $g_{\mu\nu} \propto e^{2\phi}$, equation (4.53). In that case the covariant momentum p_ϕ is constant only along null geodesics, equation (4.56),

$$p_\phi = \text{constant} \quad \text{along null geodesics} . \quad (7.124)$$

The associated **conformal Killing vector** ξ^ν , satisfying equation (7.119), is the vector whose only non-zero component is $\xi^\phi = 1$ in a coordinate system where ϕ is one of the coordinates. Equation (7.121) is modified to

$$0 = p^\mu p^\nu (\mathring{D}_{(\mu} \xi_{\nu)} - \frac{1}{4} g_{\mu\nu} \mathring{D}_\kappa \xi^\kappa) , \quad (7.125)$$

which holds because $p^\mu p^\nu g_{\mu\nu} = 0$ for null geodesics. A necessary and sufficient for equation (7.125) to hold for all null geodesics is the **conformal Killing equation**

$$\mathring{D}_{(\mu} \xi_{\nu)} - \frac{1}{4} g_{\mu\nu} \mathring{D}_\kappa \xi^\kappa = 0 , \quad (7.126)$$

the left hand side of which is the trace-free part of $\mathring{D}_{(\mu}\xi_{\nu)}$. The factor of $\frac{1}{4}$ in equations (7.125) and (7.126) is for 4 spacetime dimensions (where $g^{\mu\nu}g_{\mu\nu} = \frac{1}{4}$); the factor should be replaced by $1/N$ in N spacetime dimensions.

7.33 Killing tensors

Some symmetries are expressed by **Killing tensors** $\xi^{\mu\nu}$ rather than Killing vectors. Whereas for a Killing vector, $\xi^\nu p_\nu$ is a constant of motion along geodesics, equation (7.119), for a Killing tensor

$$\xi^{\mu\nu} p_\mu p_\nu = \text{constant} . \quad (7.127)$$

A Killing tensor $\xi^{\mu\nu}$ is symmetric without loss of generality. The metric $g_{\mu\nu}$ is itself a Killing tensor in any spacetime, since

$$g^{\mu\nu} p_\mu p_\nu = -m^2 = \text{constant} . \quad (7.128)$$

The condition of the constancy of $\xi^{\mu\nu} p_\mu p_\nu$ along geodesics is equivalent to the condition that its affine derivative vanishes along all geodesics, analogously to equation (7.120). A necessary and sufficient condition for this to be true is Killing's equation

$$\mathring{D}_{(\kappa}\xi_{\mu\nu)} = 0 , \quad (7.129)$$

where the parentheses denote symmetrization over all indices.

A **conformal Killing tensor** is one that satisfies equation (7.127) only along null geodesics. The corresponding Killing equation is the trace-free part of equation (7.129).

7.34 Lie derivative

It was remarked above that Killing's equation (7.122) can be recast as the statement that the Lie derivative of the metric along the Killing vector vanishes, equation (7.123). This section presents an exposition of the Lie derivative.

The **Lie derivative** of a coordinate tensor, whose mathematical form is derived in §§7.34.2–7.34.6, is physically minus the rate of change of the coordinate tensor with respect to a prescribed change in the coordinates, equation (7.130). The change in coordinates should be understood as leaving the spacetime itself and physical quantities within it unchanged.

Let the coordinates x^μ be changed by an infinitesimal amount ϵ with a prescribed shape $\xi^\mu(x)$ as a function of spacetime,

$$x^\mu \rightarrow x'^\mu = x^\mu + \epsilon \xi^\mu . \quad (7.130)$$

The Lie derivative of a coordinate tensor $A_{\mu\nu\dots}^{\kappa\lambda\dots}$ is defined such that the change in the coordinate tensor under the coordinate transformation (7.130) is given by ϵ times minus its Lie derivative, denoted $\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots}$,

$$A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) \rightarrow A'_{\mu\nu\dots}^{\kappa\lambda\dots}(x) = A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) - \epsilon \mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} . \quad (7.131)$$

Equivalently,

$$\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} = \lim_{\epsilon \rightarrow 0} \frac{A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) - A'_{\mu\nu\dots}{}^{\kappa\lambda\dots}(x)}{\epsilon} . \quad (7.132)$$

The reason for the minus sign in the definition (7.131) of the Lie derivative is that, as will be seen below, equation (7.157), the principal term in the expansion of the Lie derivative of a tensor $A_{\dots}$ in terms of ordinary derivatives is just its directed derivative along the direction ξ^κ ,

$$\mathcal{L}_\xi A_{\dots} = \xi^\kappa \frac{\partial A_{\dots}}{\partial x^\kappa} + \dots . \quad (7.133)$$

As its name suggests, the Lie derivative acts like a derivative: it is linear, and it satisfies the Leibniz rule. The Lie derivative is also a covariant derivative: the Lie derivative of a coordinate tensor is a coordinate tensor. Whereas the usual covariant derivative of a tensor is a tensor of rank one higher, the Lie derivative of a tensor is a tensor of the same rank. The Lie derivative can be expressed entirely in terms of coordinate derivatives without any connection coefficients, or equivalently in terms of torsion-free covariant derivatives.

Concept question 7.19 What use is a Lie derivative? Answer. The general rule to remember is that the change in any object under an infinitesimal coordinate transformation is, by construction, (minus) its Lie derivative. A prominent application of the Lie derivative is in general relativistic perturbation theory, Chapter 26, where it is essential to distinguish between genuine physical perturbations of the spacetime geometry and perturbations associated with transformations of the coordinates. Another important application of the Lie derivative is to derive the general relativistic law of conservation of energy-momentum, §16.11.2. The conservation law is a consequence of symmetry of the general relativistic action under coordinate transformations. Finally (the nominal motivation for introducing Lie derivatives here), if a spacetime possesses some special symmetry under a coordinate transformation, then that symmetry may be expressed as the vanishing of the Lie derivative of the metric with respect to the symmetry, equation (7.123).

7.34.1 The difference between the covariant derivative and the Lie derivative

The usual covariant derivative of a tensor A (dropping indices for brevity) follows from the difference between the tensor $A(x')$ evaluated at a shifted position x' , and the tensor $A(x)$ evaluated at the original position x parallel-transported to the shifted position x' ,

$$DA \propto A(x') - A(x)_{\text{parallel-transported}} . \quad (7.134)$$

Now if the shift between x' and x is the result of an infinitesimal coordinate transformation, $x' = x + \epsilon \xi$, then there is another object $A'(x')$ available, which is the tensor $A(x)$ transformed into the new (primed) coordinate frame. The Lie derivative is the difference between the tensor $A(x')$ evaluated at a shifted position x' , and the tensor $A(x)$ evaluated at the original position x , transformed into the new frame, and parallel-transported to the shifted position x' ,

$$\mathcal{L}_\xi A \propto A(x') - A'(x')_{\text{parallel-transported}} . \quad (7.135)$$

Concept question 7.19 discusses the physical justification for this mathematical artifice.

7.34.2 Lie derivative of a coordinate scalar

Under a coordinate transformation (7.130), a coordinate-frame scalar $\Phi(x)$ remains unchanged

$$\Phi(x) \rightarrow \Phi'(x') = \Phi(x) . \quad (7.136)$$

Here the scalar $\Phi'(x')$ is evaluated at position x' , which is the same as the original physical position x since all that has changed is the coordinates, not the physical position. However, the Lie derivative gives the change in a tensor evaluated at fixed coordinate position x , not at fixed physical position. The value of Φ' at x is related to that at x' by

$$\Phi'(x) = \Phi'(x' - \epsilon\xi) = \Phi'(x') - \epsilon\xi^\kappa \frac{\partial\Phi'}{\partial x^\kappa} . \quad (7.137)$$

Since ϵ is a small quantity, and Φ' differs from Φ by a small quantity, the last term $\epsilon\xi^\kappa \partial\Phi'/\partial x^\kappa$ in equation (7.141) can be replaced by $\epsilon\xi^\kappa \partial\Phi/\partial x^\kappa$ to linear order in ϵ . Putting equations (7.136) and (7.137) together shows that the coordinate scalar Φ changes under a coordinate transformation (7.130) as

$$\Phi(x) \rightarrow \Phi'(x) = \Phi(x) - \epsilon\mathcal{L}_\xi\Phi , \quad (7.138)$$

where $\mathcal{L}_\xi\Phi$ is the Lie derivative of the scalar Φ ,

$$\boxed{\mathcal{L}_\xi\Phi = \xi^\kappa \frac{\partial\Phi}{\partial x^\kappa}} \quad \text{a coordinate scalar} . \quad (7.139)$$

7.34.3 Lie derivative of a contravariant coordinate vector

A similar argument applies to coordinate vectors. Under an infinitesimal coordinate transformation (7.130), a contravariant coordinate 4-vector $A^\mu(x)$ transforms in the usual way as

$$A^\mu(x) \rightarrow A'^\mu(x') = A^\mu(x) \frac{\partial x'^\mu}{\partial x^\kappa} = A^\mu(x) + \epsilon A^\kappa(x) \frac{\partial \xi^\mu}{\partial x^\kappa} . \quad (7.140)$$

As in the scalar case, the vector $A'^\mu(x')$ is evaluated at position x' , which is the same as the original physical position since all that has changed is the coordinates, not the physical position. Again, the Lie derivative gives the change in the vector evaluated at coordinate position x , not x' . The value of A'^μ at x is related to that at x' by

$$A'^\mu(x) = A'^\mu(x' - \epsilon\xi) = A'^\mu(x') - \epsilon\xi^\kappa \frac{\partial A'^\mu}{\partial x^\kappa} . \quad (7.141)$$

The last term $\epsilon\xi^\kappa \partial A'^\mu/\partial x^\kappa$ in equation (7.141) can be replaced by $\epsilon\xi^\kappa \partial A^\mu/\partial x^\kappa$ to linear order in the infinitesimal parameter ϵ . Putting equations (7.140) and (7.141) together shows that the coordinate 4-vector A^μ changes under a coordinate transformation (7.130) as

$$A^\mu(x) \rightarrow A'^\mu(x) = A^\mu(x) - \epsilon\mathcal{L}_\xi A^\mu , \quad (7.142)$$

where $\mathcal{L}_\xi A^\mu$ is the Lie derivative of the contravariant vector A^μ ,

$$\boxed{\mathcal{L}_\xi A^\mu = \xi^\kappa \frac{\partial A^\mu}{\partial x^\kappa} - A^\kappa \frac{\partial \xi^\mu}{\partial x^\kappa}} \quad \text{a coordinate vector .} \quad (7.143)$$

The ordinary partial derivatives in equation (7.143) can be replaced by torsion-free covariant derivatives (the $\circ$ atop $\mathring{D}_\kappa$ is a reminder that it is the torsion-free covariant derivative)

$$\mathcal{L}_\xi A^\mu = \xi^\kappa \mathring{D}_\kappa A^\mu - A^\kappa \mathring{D}_\kappa \xi^\mu \quad \text{a coordinate vector .} \quad (7.144)$$

The replacement by a torsion-free covariant derivative holds because the contribution $\mathring{\Gamma}_{\nu\kappa}^\mu (\xi^\kappa A^\nu - A^\kappa \xi^\nu)$ from the torsion-free coordinate connection vanishes, because the torsion-free connection is symmetric in its last two indices, equation (2.56). Equation (7.144) holds, and the Lie derivative is a tensor, regardless of whether torsion is present. An equivalent expression for the Lie derivative of a coordinate vector A^μ in terms of torsion-full covariant derivatives D_κ is

$$\mathcal{L}_\xi A^\mu = \xi^\kappa D_\kappa A^\mu - A^\kappa D_\kappa \xi^\mu + A^\kappa \xi^\lambda S_{\kappa\lambda}^\mu \quad \text{a coordinate vector ,} \quad (7.145)$$

where $S_{\kappa\lambda}^\mu$ is the torsion. The torsion term in equation (7.145) is just such as to cancel the torsion part of the torsion-full covariant derivatives.

Exercise 7.20 Equivalence of expressions for the Lie derivative. Confirm that equations (7.143), (7.144), and (7.145) are all equivalent.

7.34.4 Lie bracket

If A^μ and B^μ are two contravariant coordinate vectors, then the Lie derivative with respect to A^μ of B^μ is minus the Lie derivative with respect to B^μ of A^μ ,

$$\mathcal{L}_A B^\mu = A^\kappa \frac{\partial B^\mu}{\partial x^\kappa} - B^\kappa \frac{\partial A^\mu}{\partial x^\kappa} = -\mathcal{L}_B A^\mu \quad \text{a coordinate vector .} \quad (7.146)$$

This antisymmetric property motivates defining the antisymmetric **Lie bracket** of two vectors $\mathbf{A} \equiv e_\mu A^\mu$ and $\mathbf{B} \equiv e_\mu B^\mu$ to be

$$\boxed{[\mathbf{A}, \mathbf{B}] \equiv \mathcal{L}_\mathbf{A} \mathbf{B} = e_\mu \mathcal{L}_\mathbf{A} B^\mu = -[\mathbf{B}, \mathbf{A}]} \quad (7.147)$$

The Lie bracket elevates the space of vectors on the manifold to a Lie algebra.

Exercise 7.21 Commutator of Lie derivatives.

1. Show that if $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are vectors, then the commutator of Lie derivatives of $\mathbf{C}$ is

$$[\mathcal{L}_\mathbf{A}, \mathcal{L}_\mathbf{B}] \mathbf{C} = [[\mathbf{A}, \mathbf{B}], \mathbf{C}] \quad (7.148)$$

2. Show that the commutator of Lie derivatives is the Lie derivative of the commutator,

$$[\mathcal{L}_\mathbf{A}, \mathcal{L}_\mathbf{B}] = \mathcal{L}_{[\mathbf{A}, \mathbf{B}]} \quad (7.149)$$

Solution.

1. This is an application of the Jacobi identity

$$[\mathbf{A}, [\mathbf{B}, \mathbf{C}]] + [\mathbf{B}, [\mathbf{C}, \mathbf{A}]] + [\mathbf{C}, [\mathbf{A}, \mathbf{B}]] = 0 . \quad (7.150)$$

The commutator of Lie derivatives of $\mathbf{C}$ is

$$[\mathcal{L}_{\mathbf{A}}, \mathcal{L}_{\mathbf{B}}]\mathbf{C} = \mathcal{L}_{\mathbf{A}}(\mathcal{L}_{\mathbf{B}}\mathbf{C}) - \mathcal{L}_{\mathbf{B}}(\mathcal{L}_{\mathbf{A}}\mathbf{C}) = [\mathbf{A}, [\mathbf{B}, \mathbf{C}]] - [\mathbf{B}, [\mathbf{A}, \mathbf{C}]] = [[\mathbf{A}, \mathbf{B}], \mathbf{C}] . \quad (7.151)$$

2. It is straightforward to check that equation (7.149) holds when acting on scalars. Equation (7.149) also holds when acting on vectors, since the rightmost side of equation (7.151) is, from equation (7.147), $\mathcal{L}_{[\mathbf{A}, \mathbf{B}]} \mathbf{C}$, so that for vectors $\mathbf{C}$,

$$[\mathcal{L}_{\mathbf{A}}, \mathcal{L}_{\mathbf{B}}]\mathbf{C} = \mathcal{L}_{[\mathbf{A}, \mathbf{B}]} \mathbf{C} . \quad (7.152)$$

Since $\mathcal{L}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{B}}$ satisfy the Leibniz rule, so also does their commutator $[\mathcal{L}_{\mathbf{A}}, \mathcal{L}_{\mathbf{B}}]$. It then follows that equation (7.149) holds when acting on arbitrary products. Thus equation (7.149) holds acting on a general tensor.

7.34.5 Lie derivative of a covariant coordinate vector

Under a coordinate transformation (7.130), a covariant coordinate 4-vector $A_{\mu}(x)$ transforms in the usual way as

$$A_{\mu}(x) \rightarrow A'_{\mu}(x') = A_{\kappa}(x) \frac{\partial x^{\kappa}}{\partial x'^{\mu}} = A_{\mu}(x) - \epsilon A_{\kappa}(x) \frac{\partial \xi^{\kappa}}{\partial x^{\mu}} . \quad (7.153)$$

Again, the vector $A'_{\mu}(x')$ is evaluated at position x' , which is the same as the original physical position x since all that has changed is the coordinates, not the physical position. And again, the Lie derivative gives the change in the vector evaluated at coordinate position x , not the physical position x' . The value of A'_{μ} at x is related to that at x' by

$$A'_{\mu}(x) = A'_{\mu}(x' - \epsilon \xi) = A'_{\mu}(x') - \epsilon \xi^{\kappa} \frac{\partial A'_{\mu}}{\partial x^{\kappa}} , \quad (7.154)$$

and again, the last term $\epsilon \xi^{\kappa} \partial A'_{\mu} / \partial x^{\kappa}$ in equation (7.154) can be replaced by $\epsilon \xi^{\kappa} \partial A_{\mu} / \partial x^{\kappa}$ to linear order in ϵ . Putting equations (7.140) and (7.141) together shows that the covariant coordinate 4-vector A_{μ} changes under a coordinate transformation (7.130) as

$$A_{\mu}(x) \rightarrow A'_{\mu}(x) = A_{\mu}(x) - \epsilon \mathcal{L}_{\xi} A_{\mu} , \quad (7.155)$$

where $\mathcal{L}_{\xi} A_{\mu}$ is the Lie derivative of the covariant vector A_{μ} ,

$$\boxed{\mathcal{L}_{\xi} A_{\mu} = \xi^{\kappa} \frac{\partial A_{\mu}}{\partial x^{\kappa}} + A_{\kappa} \frac{\partial \xi^{\kappa}}{\partial x^{\mu}}} \quad \text{a coordinate vector} . \quad (7.156)$$

As in the Lie derivative of a contravariant vector, equation (7.144), the coordinate derivatives in the Lie derivative (7.156) of a covariant vector can be replaced by torsion-free covariant derivatives.

7.34.6 Lie derivative of a coordinate tensor

In general, the Lie derivative of a coordinate tensor $A_{\mu\nu\dots}^{\kappa\lambda\dots}$ is defined by

$$\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} \equiv \xi^\pi \frac{\partial A_{\mu\nu\dots}^{\kappa\lambda\dots}}{\partial x^\pi} + A_{\pi\nu\dots}^{\kappa\lambda\dots} \frac{\partial \xi^\pi}{\partial x^\mu} + A_{\mu\pi\dots}^{\kappa\lambda\dots} \frac{\partial \xi^\pi}{\partial x^\nu} \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} \frac{\partial \xi^\kappa}{\partial x^\pi} - A_{\mu\nu\dots}^{\kappa\pi\dots} \frac{\partial \xi^\lambda}{\partial x^\pi} \quad \text{a coordinate tensor ,} \quad (7.157)$$

with an overall ∂A term, and a $+\partial\xi$ term for each covariant index and a $-\partial\xi$ term for each contravariant index. As in the Lie derivative of a vector, equation (7.144), the coordinate derivatives in the Lie derivative (7.157) of a tensor can be replaced by torsion-free covariant derivatives,

$$\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} = \xi^\pi \mathring{D}_\pi A_{\mu\nu\dots}^{\kappa\lambda\dots} + A_{\pi\nu\dots}^{\kappa\lambda\dots} \mathring{D}_\mu \xi^\pi + A_{\mu\pi\dots}^{\kappa\lambda\dots} \mathring{D}_\nu \xi^\pi \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} \mathring{D}_\pi \xi^\kappa - A_{\mu\nu\dots}^{\kappa\pi\dots} \mathring{D}_\pi \xi^\lambda \quad \text{a coordinate tensor .} \quad (7.158)$$

Equivalently, in terms of torsion-full covariant derivatives,

$$\begin{aligned} \mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} = & \xi^\pi D_\pi A_{\mu\nu\dots}^{\kappa\lambda\dots} + A_{\pi\nu\dots}^{\kappa\lambda\dots} D_\mu \xi^\pi + A_{\mu\pi\dots}^{\kappa\lambda\dots} D_\nu \xi^\pi \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} D_\pi \xi^\kappa - A_{\mu\nu\dots}^{\kappa\pi\dots} D_\pi \xi^\lambda \dots \\ & + (A_{\pi\nu\dots}^{\kappa\lambda\dots} S_{\mu\rho}^\pi + A_{\mu\pi\dots}^{\kappa\lambda\dots} S_{\nu\rho}^\pi \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} S_{\pi\rho}^\kappa - A_{\mu\nu\dots}^{\kappa\pi\dots} S_{\pi\rho}^\lambda) \xi^\rho \quad \text{a coordinate tensor .} \end{aligned} \quad (7.159)$$

Exercise 7.22 Lie derivative of the metric. What is the Lie derivative of the metric tensor $g_{\mu\nu}$ along the direction ξ^κ ?

Solution. The Lie derivative of the metric $g_{\mu\nu}$ along ξ^κ is

$$\begin{aligned} \mathcal{L}_\xi g_{\mu\nu} &= \xi^\kappa \frac{\partial g_{\mu\nu}}{\partial x^\kappa} + g_{\kappa\nu} \frac{\partial \xi^\kappa}{\partial x^\mu} + g_{\mu\kappa} \frac{\partial \xi^\kappa}{\partial x^\nu} \\ &= \frac{\partial \xi_\nu}{\partial x^\mu} + \frac{\partial \xi_\mu}{\partial x^\nu} - 2\mathring{\Gamma}_{\kappa\mu\nu} \xi^\kappa \\ &= \mathring{D}_\mu \xi_\nu + \mathring{D}_\nu \xi_\mu , \end{aligned} \quad (7.160)$$

where $\mathring{\Gamma}_{\kappa\mu\nu}$ is the torsion-free coordinate-frame connection, equation (2.63), and $\mathring{D}_\mu$ is the torsion-free covariant derivative.

Exercise 7.23 Lie derivative of the inverse metric. Show that the Lie derivative of a Kronecker delta is zero,

$$\mathcal{L}_\xi \delta_\mu^\kappa = 0 . \quad (7.161)$$

Show that the Lie derivative of the inverse metric tensor $g^{\kappa\lambda}$ is

$$\mathcal{L}_\xi g^{\kappa\lambda} = -g^{\kappa\mu} g^{\lambda\nu} \mathcal{L}_\xi g_{\mu\nu} = -(\mathring{D}^\kappa \xi^\lambda + \mathring{D}^\lambda \xi^\kappa) . \quad (7.162)$$

Exercise 7.24 Lie derivative of the metric determinant. Show that the Lie derivative of the metric determinant is

$$\mathcal{L}_\xi \ln |g| = g^{\mu\nu} \mathcal{L}_\xi g_{\mu\nu} = \xi^\mu \frac{\partial \ln |g|}{\partial x^\mu} + 2 \frac{\partial \xi^\mu}{\partial x^\mu} . \quad (7.163)$$

Solution. The first equality of equation (7.163) follows because a Lie derivative is a variation (with respect to a coordinate transformation), and the variation of the determinant of any matrix is given by equation (2.77). The second equality of equation (7.163) follows from the first line of equations (7.160).

Reissner-Nordström Black Hole

The Reissner-Nordström geometry, discovered independently by Hans Reissner (1916), Hermann Weyl (1917), and Gunnar Nordström (1918), describes the unique spherically symmetric static solution for a black hole with mass and electric charge in asymptotically flat spacetime.

As with the Schwarzschild geometry, the mathematics of the Reissner-Nordström geometry was understood long before conceptual understanding emerged. The meaning of the Reissner-Nordström geometry was eventually clarified by Graves and Brill (1960).

8.1 Reissner-Nordström metric

The **Reissner-Nordström metric** for a black hole of mass M and electric charge Q is, in geometric units $c = G = 1$,

$$\boxed{ds^2 = -\Delta dt^2 + \Delta^{-1} dr^2 + r^2 d\phi^2}, \quad (8.1)$$

where $\Delta(r)$ is the horizon function,

$$\Delta \equiv 1 - \frac{2M}{r} + \frac{Q^2}{r^2}. \quad (8.2)$$

The Reissner-Nordström metric (8.1) looks like the Schwarzschild metric (7.1) with the replacement

$$M \rightarrow M(r) = M - \frac{Q^2}{2r}. \quad (8.3)$$

The quantity $M(r)$ in equation (8.3) has a coordinate-independent interpretation as the mass $M(r)$ interior to radius r , which here is the mass M at infinity, minus the mass in the electric field $E = Q/r^2$ outside r ,

$$\int_r^\infty \frac{E^2}{8\pi} 4\pi r^2 dr = \int_r^\infty \frac{Q^2}{8\pi r^4} 4\pi r^2 dr = \frac{Q^2}{2r}. \quad (8.4)$$

The units of Q here are gaussian; in Heaviside units the electric field is $E = Q/(4\pi r^2)$, the energy density is $E^2/2$, and the charge term in the horizon function would be $Q^2/(8\pi r^2)$. Equations (8.4) seem like a

Newtonian calculation of the energy in the electric field, but it turns out to be valid also in general relativity, essentially because the radial electric field E is unchanged by a Lorentz boost along the radial direction.

Real astronomical black holes probably have very little electric charge, because the Universe as a whole appears almost electrically neutral (and Maxwell's equations in fact demand that the Universe in its entirety should be exactly electrically neutral), and a charged black hole would quickly neutralize itself. It would probably not neutralize itself completely, but have some small residual positive charge, because protons (positive charge) are more massive than electrons (negative charge), so it is slightly easier for protons than electrons to overcome a Coulomb barrier.

Nevertheless, the Reissner-Nordström solution is of more than passing interest because its internal geometry resembles that of the Kerr solution for a rotating black hole.

Concept question 8.1 Units of charge of a charged black hole. What is the charge Q in standard (either gaussian or SI) units?

8.2 Energy-momentum tensor

The Einstein tensor of the Reissner-Nordström metric (8.1) is diagonal, with elements given by

$$G^\nu_\mu = \begin{pmatrix} G^t_t & 0 & 0 & 0 \\ 0 & G^r_r & 0 & 0 \\ 0 & 0 & G^\theta_\theta & 0 \\ 0 & 0 & 0 & G^\phi_\phi \end{pmatrix} = 8\pi \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p_r & 0 & 0 \\ 0 & 0 & p_\perp & 0 \\ 0 & 0 & 0 & p_\perp \end{pmatrix} = \frac{Q^2}{r^4} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (8.5)$$

The trick of writing one index up and the other down on the Einstein tensor G^ν_μ partially cancels the distorting effect of the metric, yielding the proper energy density ρ , the proper radial pressure p_r , and transverse pressure $p_\perp$, up to factors of ± 1 . A more systematic way to extract proper quantities is to work in the tetrad formalism, Chapter 11.

The energy-momentum tensor is that of a radial electric field

$$E = \frac{Q}{r^2}. \quad (8.6)$$

Notice that the radial pressure p_r is negative, while the transverse pressure $p_\perp$ is positive. It is no coincidence that the sum of the energy density and pressures is twice the energy density, $\rho + p_r + 2p_\perp = 2\rho$.

The negative pressure, or tension, of the radial electric field produces a gravitational repulsion that dominates at small radii, and that is responsible for much of the strange phenomenology of the Reissner-Nordström geometry. The gravitational repulsion mimics the centrifugal repulsion inside a rotating black hole, for which reason the Reissner-Nordström geometry is often used a surrogate for the rotating Kerr-Newman geometry.

At this point, the statements that the energy-momentum tensor is that of a radial electric field, and that the radial tension produces a gravitational repulsion that dominates at small radii, are true but unjustified assertions.

8.3 Weyl tensor

As with the Schwarzschild geometry (indeed, any spherically symmetric geometry), only 1 of the 10 independent spin components of the Weyl tensor is non-vanishing, the real spin-0 component, the Weyl scalar C . The Weyl scalar for the Reissner-Nordström geometry is

$$C = -\frac{M}{r^3} + \frac{Q^2}{r^4} . \quad (8.7)$$

The Weyl scalar goes to infinity at zero radius,

$$C \rightarrow \infty \quad \text{as } r \rightarrow 0 , \quad (8.8)$$

signalling the presence of a genuine singularity at zero radius, where the curvature, the tidal force, diverges.

8.4 Horizons

The Reissner-Nordström geometry has not one but two horizons. The horizons occur where an object at rest in the geometry, $dr = d\theta = d\phi = 0$, follows a null geodesic, $ds^2 = 0$, which occurs where the horizon function Δ , equation (8.2), vanishes,

$$\Delta = 0 . \quad (8.9)$$

This is a quadratic equation in r , and it has two solutions, an **outer horizon** r_+ and an **inner horizon** r_-

$$r_{\pm} = M \pm \sqrt{M^2 - Q^2} . \quad (8.10)$$

It is straightforward to check that the Reissner-Nordström time coordinate t is timelike outside the outer horizon, $r > r_+$, spacelike between the horizons $r_- < r < r_+$, and again timelike inside the inner horizon $r < r_-$. Conversely, the radial coordinate r is spacelike outside the outer horizon, $r > r_+$, timelike between the horizons $r_- < r < r_+$, and spacelike inside the inner horizon $r < r_-$.

The physical meaning of this strange behaviour is akin to that of the Schwarzschild geometry. As in the Schwarzschild geometry, outside the outer horizon space is falling at less than the speed of light; at the outer horizon space hits the speed of light; and inside the outer horizon space is falling faster than light. But a new ingredient appears. The gravitational repulsion caused by the negative pressure of the electric field slows down the flow of space, so that it slows back down to the speed of light at the inner horizon. Inside the inner horizon space is falling at less than the speed of light.

8.5 Gullstrand-Painlevé metric

Deeper insight into the Reissner-Nordström geometry comes from examining its Gullstrand-Painlevé metric. The Gullstrand-Painlevé metric for the Reissner-Nordström geometry has the same form as that for the

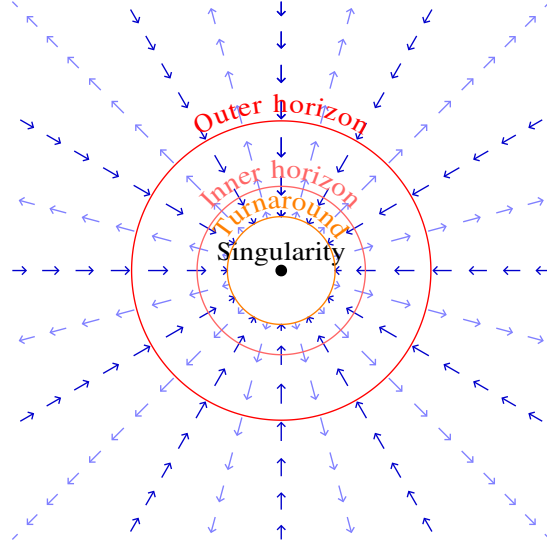


Figure 8.1 Depiction of the Gullstrand-Painlevé metric for the Reissner-Nordström geometry, for a black hole of charge $Q = 0.96M$. The Gullstrand-Painlevé line-element defines locally inertial frames attached to observers who free-fall radially from zero velocity at infinity. Frames fall at less than the speed of light outside the outer horizon, hit the speed of light at the outer horizon, and fall faster than light in the black hole region inside the outer horizon. The gravitational attraction from the mass of the black hole is counteracted by a gravitational repulsion produced by the tension (negative radial pressure) of the electric field. The repulsion grows stronger at smaller radii, slowing the inflow. The inflow slows back down to the speed of light at the inner horizon, comes to a halt at the turnaround radius, turns around, and accelerates outward. Now moving outward, the flow hits the speed of light at the inner horizon, and passes outward through the inner horizon into a new region of spacetime, a white hole, where frames are moving outward faster than light. The repulsion from the tension of the electric field weakens at larger radii, slowing the outflow. The outflow drops back down to the speed of light at the outer horizon of the white hole, and exits the outer horizon into a new piece of spacetime.

Schwarzschild geometry,

$$ds^2 = -dt_{\text{ff}}^2 + (dr - \beta dt_{\text{ff}})^2 + r^2 d\phi^2. \quad (8.11)$$

The velocity β is again the escape velocity, but this is now

$$\beta = \mp \sqrt{\frac{2M(r)}{r}}, \quad (8.12)$$

where $M(r) = M - Q^2/2r$ is the interior mass already given as equation (8.3). Horizons occur where the magnitude of the velocity β equals the speed of light

$$|\beta| = 1, \quad (8.13)$$

which happens at the outer and inner horizons $r = r_+$ and $r = r_-$, equation (8.10).

The Gullstrand-Painlevé metric once again paints the picture of space falling into the black hole. Outside the outer horizon r_+ space falls at less than the speed of light, at the horizon space falls at the speed of light, and inside the horizon space falls faster than light. But the gravitational repulsion produced by the tension of the radial electric field starts to slow down the inflow of space, so that the infall velocity reaches a maximum at $r = Q^2/M$. The infall slows back down to the speed of light at the inner horizon r_- . Inside the inner horizon, the flow of space slows all the way to zero velocity, $\beta = 0$, at the turnaround radius

$$r_0 = \frac{Q^2}{2M} . \quad (8.14)$$

Space then turns around, the velocity β becoming positive, and accelerates back up to the speed of light. Space is now accelerating outward, to larger radii r . The outfall velocity reaches the speed of light at the inner horizon r_- , but now the motion is outward, not inward. Passing back out through the inner horizon, space is falling outward faster than light. This is not the black hole, but an altogether new piece of spacetime, a white hole. The white hole looks like a time-reversed black hole. As space falls outward, the gravitational repulsion produced by the tension of the radial electric field declines, and the outflow slows. The outflow slows back to the speed of light at the outer horizon r_+ of the white hole. Outside the outer horizon of the white hole is a new universe, where once again space is flowing at less than the speed of light.

What happens inward of the turnaround radius r_0 , equation (8.14)? Inside this radius the interior mass $M(r)$, equation (8.3), is negative, and the velocity β is imaginary. The interior mass $M(r)$ diverges to negative infinity towards the central singularity at $r \rightarrow 0$. The singularity is timelike, and infinitely gravitationally repulsive, unlike the central singularity of the Schwarzschild geometry. Is it physically realistic to have a singularity that has infinite negative mass and is infinitely gravitationally repulsive? Undoubtedly not.

8.6 Radial null geodesics

In Reissner-Nordström coordinates, light rays that fall radially ($d\theta = d\phi = 0$) follow

$$\frac{dr}{dt} = \pm \Delta . \quad (8.15)$$

Equation (8.15) shows that $dr/dt \rightarrow 0$ as $r \rightarrow r_{\pm}$, suggesting that null rays can never cross a horizon. As in the Schwarzschild geometry, this is an artefact of the choice of coordinate system. As in the Schwarzschild geometry, the Reissner-Nordström metric (8.1) appears singular at the horizons, where $\Delta = 0$, but this is a coordinate singularity, not a true singularity, as is evident from the fact that the Riemann curvature tensor remains finite at the horizons.

Figure 8.2 shows a spacetime diagram of the Reissner-Nordström geometry in Reissner-Nordström coordinates. The spacetime diagram illustrates the apparent freezing of infalling and outgoing null geodesics at both outer and inner horizons.

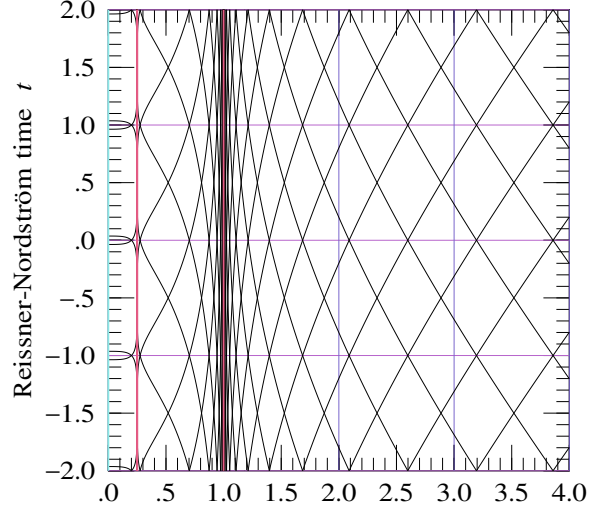


Figure 8.2 Spacetime diagram of the Reissner-Nordström geometry, in Reissner-Nordström coordinates, for a black hole of charge $Q = 0.8M$, plotted in units of the outer horizon radius r_+ of the black hole. The geometry has two horizons (pink), an outer horizon, and an inner horizon at $r_- = 0.25r_+$. The more or less diagonal lines (black) are outgoing and infalling null geodesics. The outgoing and infalling null geodesics appear not to cross the horizon, but this is an artefact of the Reissner-Nordström coordinate system.

8.7 Finkelstein coordinates

Finkelstein and Kruskal-Szekeres coordinates can be constructed for the Reissner-Nordström geometry just as in the Schwarzschild geometry.

Introduce the tortoise coordinate r^* defined by

$$r^* \equiv \int \frac{dr}{\Delta} = r + \frac{1}{2\kappa_+} \ln \left| 1 - \frac{r}{r_+} \right| + \frac{1}{2\kappa_-} \ln \left| 1 - \frac{r}{r_-} \right| , \quad (8.16)$$

where $\kappa_{\pm}$ are the surface gravities at the two horizons

$$\kappa_{\pm} = \pm \frac{r_+ - r_-}{2r_{\pm}^2} . \quad (8.17)$$

Radially infalling and outgoing null geodesics follow

$$\begin{aligned} r^* + t &= \text{constant} && \text{infalling} , \\ r^* - t &= \text{constant} && \text{outgoing} . \end{aligned} \quad (8.18)$$

Finkelstein time t_F is defined by

$$t_F + r = t + r^* , \quad (8.19)$$

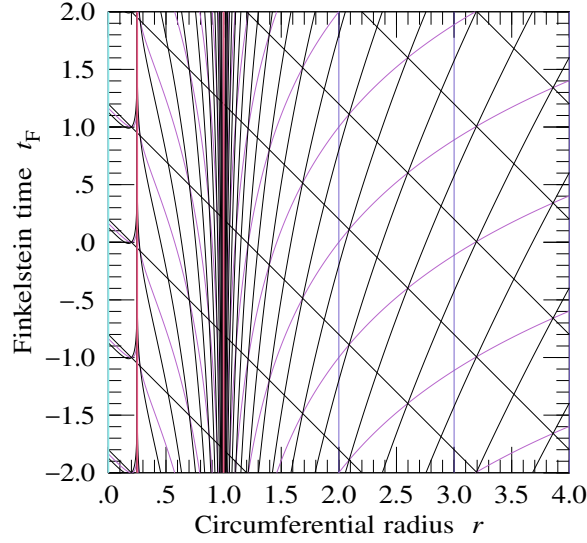


Figure 8.3 Finkelstein spacetime diagram of the Reissner-Nordström geometry, for a black hole of charge $Q = 0.8M$, plotted in units of the outer horizon radius r_+ of the black hole. The Finkelstein time coordinate t_F is constructed so that radially infalling light rays are at 45° .

which is constructed so that infalling null rays follow $t_F + r = 0$. Figure 8.3 shows the Finkelstein spacetime diagram of the Reissner-Nordström geometry.

8.8 Kruskal-Szekeres coordinates

With respect to the coordinates t and r^* , the Reissner-Nordström line-element is

$$ds^2 = \Delta (-dt^2 + dr^{*2}) + r^2 d\phi^2 . \quad (8.20)$$

This metric is still ill-behaved at the horizons, where $\Delta = 0$ and where the tortoise coordinate r^* diverges logarithmically, with $r^* \rightarrow -\infty$ as $r \rightarrow r_+$ and $r^* \rightarrow +\infty$ as $r \rightarrow r_-$. The misbehaviour at the two horizons can be removed by transforming to Kruskal coordinates r_K and t_K defined by

$$r_K + t_K \equiv f(r^* + t) , \quad (8.21a)$$

$$r_K - t_K \equiv sf(r^* - t) + 2nk , \quad (8.21b)$$

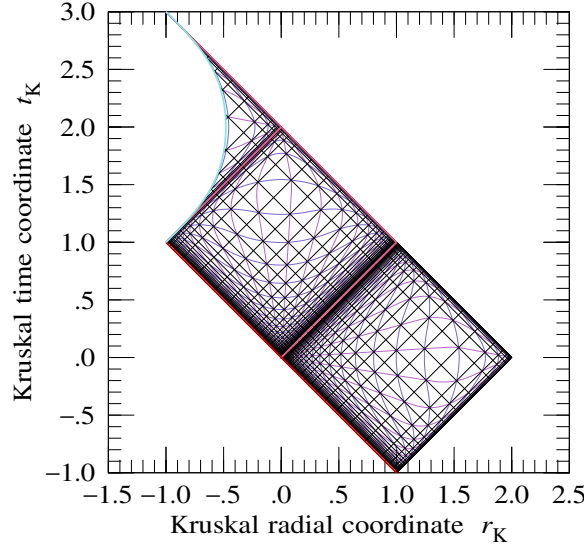


Figure 8.4 Kruskal spacetime diagram of the Reissner-Nordström geometry, plotted in units of k , equation (8.23), for a black hole of charge $Q = 0.96M$. The Kruskal coordinates t_K and r_K are defined by equations (8.21), and are constructed so that radially infalling and outgoing light rays are at 45° . Lines of constant Reissner-Nordström time t (violet), and infalling and outgoing null lines (black) are spaced uniformly at intervals of 1 (units $r_+ = 1$), while lines of constant circumferential radius r (blue) are spaced uniformly in the tortoise coordinate r^* , equation (8.16), so that the intersections of t and r lines are also intersections of infalling and outgoing null lines.

where the function $f(z)$ is

$$f(z) \equiv \begin{cases} \frac{e^{\kappa_+ z}}{\kappa_+} & z \leq 0, \\ \frac{e^{\kappa_- z}}{\kappa_-} + k & z \geq 0, \end{cases} \quad (8.22)$$

which varies from $f(z) \rightarrow 0$ as $z \rightarrow -\infty$, to $f(z) \rightarrow k$ as $z \rightarrow +\infty$, and is continuous and differentiable at the junction $z = 0$. The constant k is

$$k \equiv \frac{1}{\kappa_+} - \frac{1}{\kappa_-} = \frac{2(r_+^2 + r_-^2)}{r_+ - r_-}. \quad (8.23)$$

The constants s and n in equation (8.21b) are a sign and an integer that fix the sign and offset of the Kruskal coordinates in each quadrant of the Kruskal diagram. Figure 8.4 shows the resulting Kruskal spacetime diagram, containing three quadrants, a region outside the outer horizon, a region between the two horizons, and a region inside the inner horizon. The integers $\{s, n\}$ in the three quadrants are $\{1, 0\}$ in the region outside the outer horizon, $\{-1, 0\}$ in the region between the two horizons, and $\{1, -1\}$ in the region inside the inner horizon.

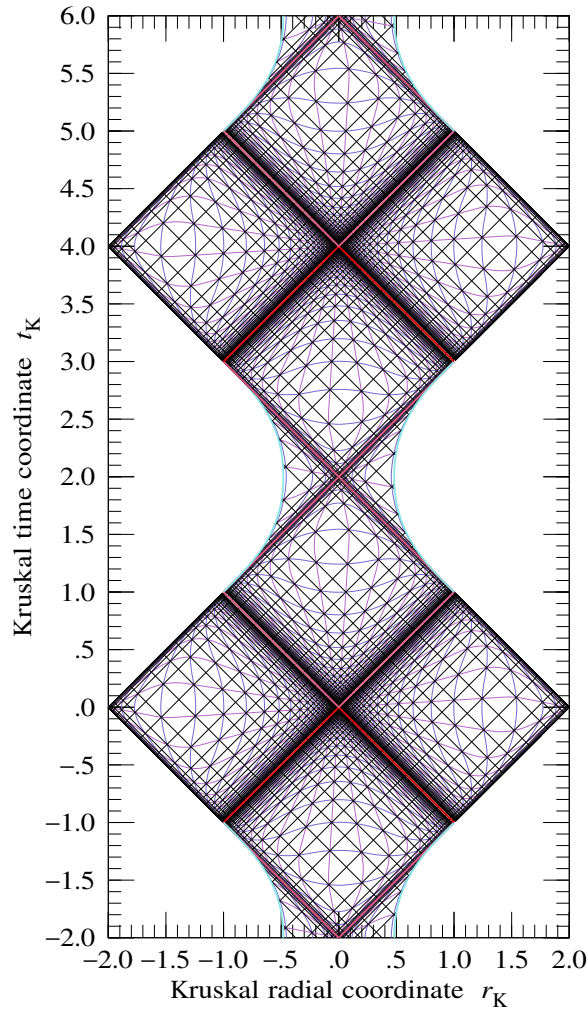


Figure 8.5 Kruskal spacetime diagram of the analytically extended Reissner-Nordström geometry, plotted in units of k , equation (8.23), for a black hole of charge $Q = 0.96M$.

The transformation (8.21) to Kruskal coordinates brings infinite time t and radius r to finite values, as in a Penrose diagram. This is associated with the fact that the tortoise coordinate r^* is $+\infty$ at both $r = \infty$ and $r = r_-$, so any transformation of $r^* \pm t$ that maps the inner horizon r_- to a finite coordinate also maps infinite radius to a finite coordinate. It would be possible to allow r_K to be infinite at infinite r , as in Schwarzschild, by choosing different Kruskal coordinate transformations for the regions near the inner horizon and near

infinity, but it is advantageous to enforce the same transformation, since the Kruskal coordinate system can then be extended analytically across both inner and outer horizons.

The Kruskal diagram 8.4 shows that the singularity of the Reissner-Nordström geometry is timelike, not spacelike. This is associated with the fact that the singularity is gravitationally repulsive, not attractive.

The Penrose diagram of the Reissner-Nordström geometry is commonly drawn with the singularity vertical. The singularity in the Kruskal diagram 8.4 is not vertical. It is possible to construct Kruskal-like coordinates such that the singularity is vertical in the resulting spacetime diagram, for example by setting $\kappa_- = -\kappa_+$ in the Kruskal transformation formulae (8.22) and (8.23). However, the metric coefficients in t_K and r_K are then zero, not finite, at the inner horizon. If the metric coefficients are required to be finite at both outer and inner horizons, then it is impossible to construct a Kruskal coordinate transformation that makes the singularity vertical.

8.9 Analytically extended Reissner-Nordström geometry

Like the Schwarzschild geometry, the Reissner-Nordström geometry can be analytically extended. Figure 8.5 shows the Kruskal spacetime diagram of the analytically extended geometry. The extension is considerably more complicated than that for Schwarzschild, as discussed in the next section.

8.10 Penrose diagram

Figure 8.6 shows a Penrose diagram of the analytic continuation of the Reissner-Nordström geometry. This is essentially a schematic version of the Kruskal diagram 8.5, with the various parts of the geometry labelled. The analytic continuation consists of an infinite ladder of universes and parallel universes connected to each other by black hole $\rightarrow$ wormhole $\rightarrow$ white hole tunnels. I call the various pieces of spacetime “Universe,” “Parallel Universe,” “Black Hole,” “Wormhole,” “Parallel Wormhole,” and “White Hole.” These pieces repeat in an infinite ladder. The various horizons in the Penrose diagram are labelled with descriptive names. Relativists tend to use more abstract terminology.

The Wormhole and Parallel Wormhole contain separate central singularities, the “Singularity” and the “Parallel Singularity,” which are oppositely charged. If the black hole is positively charged as measured by observers in the Universe, then it is negatively charged as measured by observers in the Parallel Universe, and the Wormhole contains a positive charge singularity while the Parallel Wormhole contains a negative charge singularity.

Where does the electric charge of the Reissner-Nordström geometry “actually” reside? This comes down to the question of how observers detect the presence of charge. Observers detect charge by the electric field that it produces. Equip all (radially moving) observers with a gyroscope that they orient consistently in the same radial direction, which can be taken to be towards the black hole as measured by observers in the Universe. Observers in the Parallel Universe find that their gyroscope is pointed away from the black hole. Inside the black hole, observers from either Universe agree that the gyroscope is pointed towards the

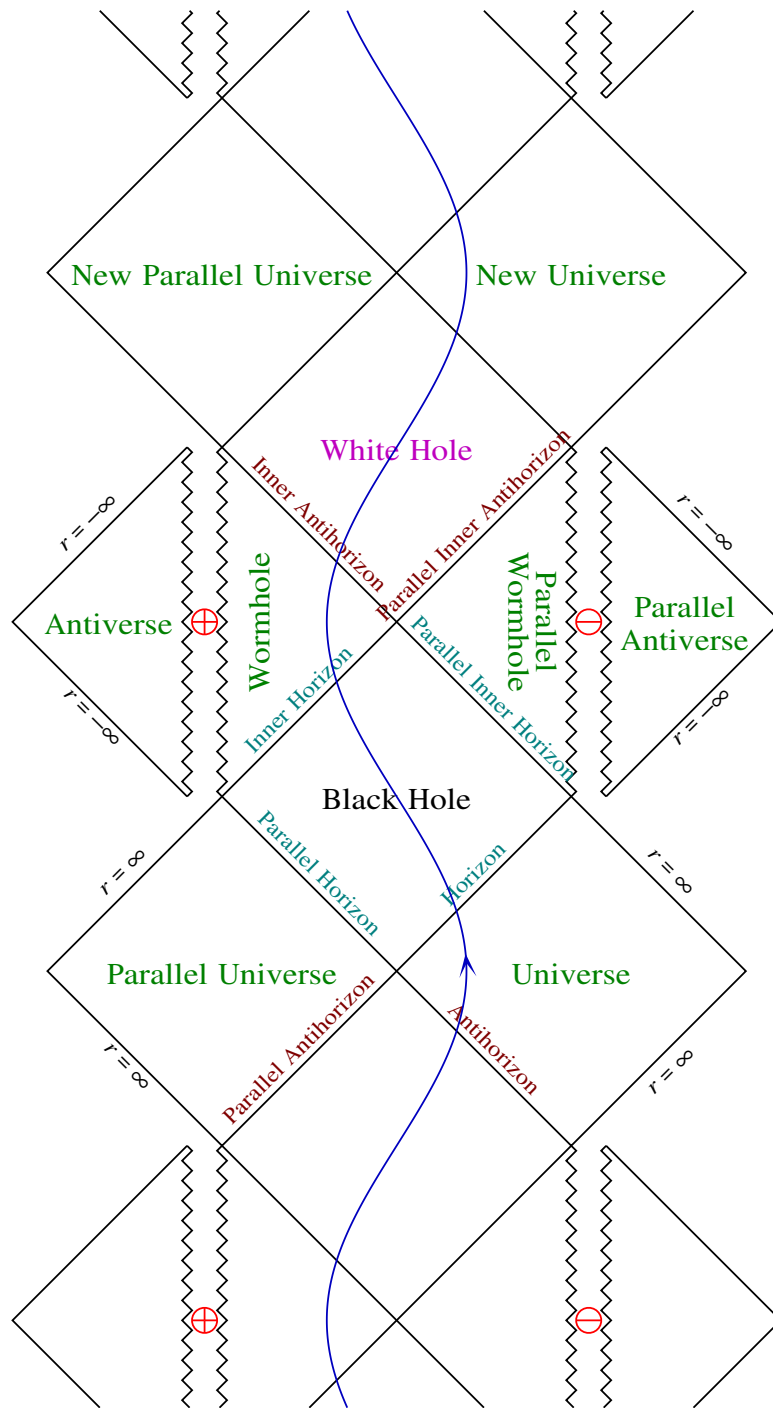


Figure 8.6 Penrose diagram of the analytically extended Reissner-Nordström geometry.

Wormhole, and away from the Parallel Wormhole. All observers agree that the electric field is pointed in the same radial direction. Observers who end up inside the Wormhole measure an electric field that appears to emanate from the Singularity, and which they therefore attribute to charge in the Singularity. Observers who end up inside the Parallel Wormhole measure an electric field that appears to emanate in the opposite direction from the Parallel Singularity, and which they therefore attribute to charge of opposite sign in the Parallel Singularity. Strange, but all consistent.

8.11 Antiverse: Reissner-Nordström geometry with negative mass

It is also possible to consider the Reissner-Nordström geometry for negative values of the radius r . I call the extension to negative r the “Antiverse.” There is also a “Parallel Antiverse.”

Changing the sign of r in the Reissner-Nordström metric (8.1) is equivalent to changing the sign of the mass M . Thus the Reissner-Nordström metric with negative r describes a charged black hole of negative mass

$$M < 0 . \quad (8.24)$$

The negative mass black hole is gravitationally repulsive at all radii, and it has no horizons.

8.12 Outgoing, ingoing

The black hole in the Reissner-Nordström geometry has not one but two inner horizons. The inner horizon plays a central role in the inflationary instability described in §8.13 below.

The inner horizons can be called outgoing and ingoing. Persons freely falling in the Black Hole region are all moving inward in coordinate radius r , but they may be moving either forward or backward in Reissner-Nordström coordinate time t . In the Black Hole region, the conserved energy along a geodesic is positive if the time coordinate t is decreasing, negative if the time coordinate t is increasing¹. Persons with positive energy are **ingoing**, while persons with negative energy are **outgoing**. Both outgoing and ingoing persons fall inward, to smaller radii, but outgoing persons think that the inward direction is towards the Parallel Wormhole, while ingoing persons think that the inward direction is in the opposite direction, towards the Wormhole. Outgoing persons fall through the outgoing inner horizon, while ingoing persons fall through the ingoing inner horizon.

Coordinate time t moves forwards in the Universe and Wormhole regions, and geodesics have positive energy in these regions. Conversely, coordinate time t moves backwards in the Parallel Universe and Parallel Wormhole regions, and geodesics have negative energy in these regions. Of course, all observers, wherever

¹ The fact that positive energy geodesics go backwards in Reissner-Nordström coordinate time t in the Black Hole region is counter-intuitive, but it does make sense. An outgoing infaller who fell through the horizon earlier can meet an ingoing infaller who falls in later. Thus outgoers, who have negative energy, progress forward in time t , while ingoers, who have positive energy, progress backward in time t .

they may be, always perceive their own proper time to be moving forward in the usual fashion, at the rate of one second per second.

8.13 The inflationary instability

Roger Penrose (1968) first pointed out that a person passing through the outgoing inner horizon (also called the Cauchy horizon) of the Reissner-Nordström geometry would see the outside Universe infinitely blueshifted, and he suggested that this would destabilize the geometry. Perturbation theory calculations, starting with Simpson & Penrose (1973) and culminating with Chandrasekhar and Hartle (1982), confirmed that waves become infinitely blueshifted as they approach the outgoing inner horizon, and that their energy

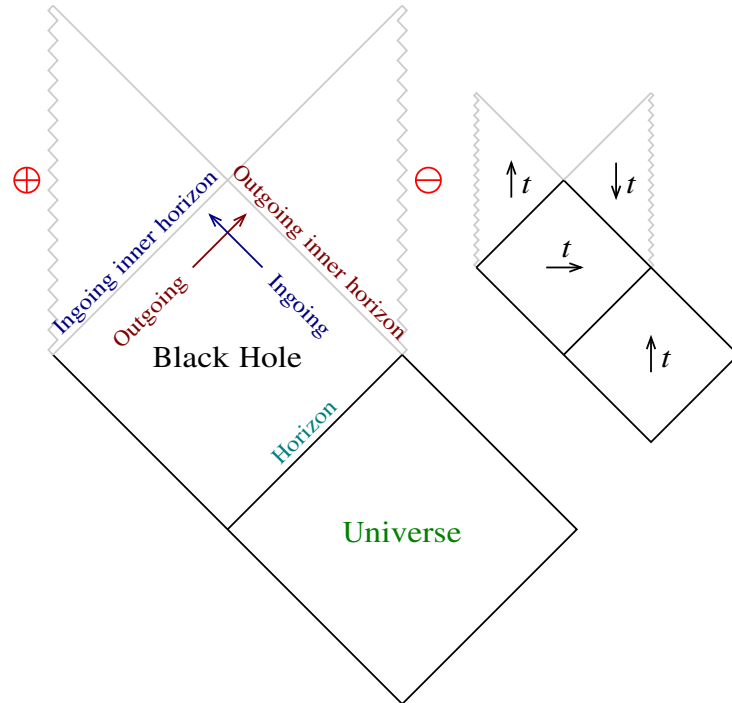


Figure 8.7 Penrose diagram illustrating why the Reissner-Nordström geometry is subject to the inflationary instability. Outgoing and ingoing streams just outside the inner horizon must pass through separate outgoing and ingoing inner horizons into causally separated pieces of spacetime where the timelike time coordinate t goes in opposite directions. To accomplish this, the outgoing and ingoing streams must exceed the speed of light through each other, which physically they cannot do. The inflationary instability is driven by the pressure of the relativistic counter-streaming between outgoing and ingoing streams. The inset shows the direction of coordinate time t in the various regions. Proper time of course always increases upward in a Penrose diagram.

density diverges. The perturbation theory calculations were widely construed as indicating that the Reissner-Nordström geometry was “unstable,” although the precise nature of this instability remained obscure.

It was not until a seminal paper by Poisson & Israel (1990) that the nonlinear nature of the instability at the inner horizon was clarified. Poisson & Israel showed that the Reissner-Nordström geometry is subject to an exponentially growing instability which they dubbed **mass inflation**. The term refers to the fact that the interior mass $M(r)$ grows exponentially during mass inflation. The interior mass $M(r)$ has the property of being a gauge-invariant, scalar quantity, so it has a physical meaning independent of the coordinate system.

What causes mass inflation? Actually it has nothing to do with mass: the inflating mass is just a symptom of the underlying cause. What causes mass inflation is relativistic counter-streaming between outgoing and ingoing streams. Since the name mass inflation can be misleading, I prefer to call it the **inflationary instability**. As the Penrose diagram of the Reissner-Nordström geometry shows, outgoing and ingoing streams must drop through separate outgoing and ingoing inner horizons into separate pieces of spacetime, the Wormhole and the Parallel Wormhole. The regions of spacetime must be separate because coordinate time t is timelike in both regions, but going in opposite directions in the two regions, forward in the Wormhole, backward in the Parallel Wormhole, as illustrated in Figure 8.7. In other words, outgoing and ingoing streams cannot co-exist in the same subluminal region of spacetime because they would have to be moving in opposite directions in time, which cannot be.

In the Reissner-Nordström geometry, outgoing and ingoing streams resolve their differences by exceeding the speed of light relative to each other, and passing into causally separated regions. As the outgoing and ingoing streams drop through their respective inner horizons, they each see the other stream infinitely blueshifted.

In reality however, this cannot occur: outgoing and ingoing streams cannot exceed the speed of light relative to each other. Instead, as the outgoing and ingoing streams move ever faster through each other in their effort to drop through the inner horizon, their counter-streaming generates a radial pressure. The pressure, which is positive, exerts an inward gravitational force. As the counter-streaming approaches the speed of light, the gravitational force produced by the counter-streaming pressure eventually exceeds the gravitational force produced by the background Reissner-Nordström geometry. At this point, the inflationary instability begins.

The gravitational force produced by the counter-streaming is inwards, but, in the strange way that general relativity operates, the inward direction is in opposite directions for the ingoing streams, towards the black hole for the ingoing stream, and away from the black hole for the outgoing stream. Consequently the counter-streaming pressure simply accelerates the outgoing and ingoing streams ever faster through each other. The result is an exponential feedback instability. The increasing pressure accelerates the streams faster through each other, which increases the pressure, which increases the acceleration.

The interior mass is not the only thing that increases exponentially during mass inflation. The proper density and pressure, and the Weyl scalar (all gauge-invariant scalars) exponentiate together.

Exercise 8.2 Blueshift of a photon crossing the inner horizon of a Reissner-Nordström black hole. Show that, in the Reissner-Nordström geometry, the blueshift of a photon with energy $v_t = \mp 1$ and angular momentum per unit energy $\mathbf{v}_\perp = \mathbf{J}$ observed by observer on a geodesic with energy per unit mass $u_t = -E$ and angular momentum per unit mass $\mathbf{u}_\perp = \mathbf{L}$ is (the minus sign in $-u_\mu v^\mu$ makes the blueshift positive)

$$-u_\mu v^\mu = \frac{\text{something}}{\Delta} . \quad (8.25)$$

Argue that the blueshift diverges at the horizon for outgoing observers observing ingoing photons, and for ingoing observers observing outgoing photons.

Solution. The solution for geodesics is similar to that in the Schwarzschild geometry, Exercise 7.6. The radial velocities u^r and v^r are both necessarily negative just above the inner horizon. The blueshift of a photon is

$$\begin{aligned} -u_\mu v^\mu &= -(g^{tt}u_t v_t + g_{rr}u^r v^r + g^{\perp\perp}u_\perp v_\perp) \\ &= \frac{\mp E + \sqrt{[E^2 - (1 + L^2/r^2)\Delta][1 - (J^2/r^2)\Delta]}}{-\Delta} - \frac{LJ}{r^2} . \end{aligned} \quad (8.26)$$

Note that Δ is negative between the outer and inner horizons. The $\mp$ sign of $\mp E$ is negative if u_t and v_t have the same sign, positive if u_t and v_t have opposite signs. The latter case holds for outgoing observers observing ingoing photons, or for ingoing observers observing outgoing photons, in which case the blueshift near the inner horizon, where $\Delta \rightarrow -0$, diverges as

$$-u_\mu v^\mu \rightarrow \left| \frac{2u_t v_t}{\Delta} \right| \quad \text{as } \Delta \rightarrow -0 \quad \text{if } u_t v_t < 0 . \quad (8.27)$$

8.14 The X point

The point in the Reissner-Nordström geometry where the outgoing and ingoing inner horizons intersect, the X point, is a special one. This is the point through which geodesics of zero energy, $E = 0$, must pass. Persons with zero energy who reach the X point see both outgoing and ingoing streams, coming from opposite directions, infinitely blueshifted.

8.15 Extremal Reissner-Nordström geometry

So far the discussion of the Reissner-Nordström geometry has centred on the case $Q < M$ (or more generally, $|Q| < |M|$) where there are separate outer and inner horizons. In the special case that the charge and mass are equal,

$$Q = M , \quad (8.28)$$

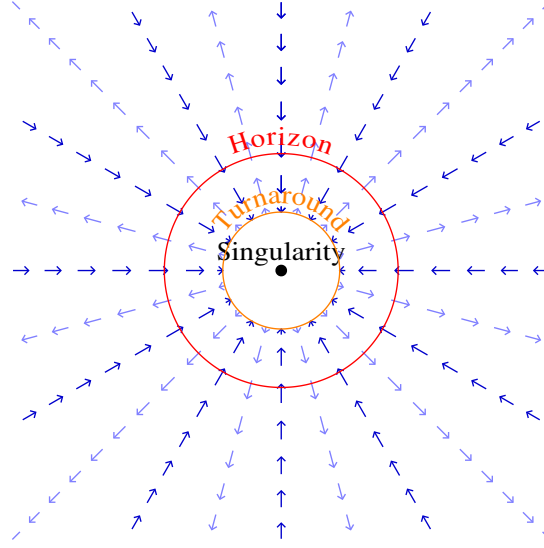


Figure 8.8 Depiction of the Gullstrand-Painlevé metric for the extremal Reissner-Nordström geometry, with $Q = M$. In the extremal geometry, the inner and outer horizons are at the same radius, so there is only one horizon.

the inner and outer horizons merge into one, $r_+ = r_-$, equation (8.10). This special case describes the **extremal Reissner-Nordström geometry**.

The extremal Reissner-Nordström geometry is of particular interest in quantum gravity because its Hawking temperature is zero, and in string theory because extremal black holes have a higher degree of symmetry, making them more tractable for theoretical investigation.

Figure 8.8 shows the Gullstrand-Painlevé model of an extremal Reissner-Nordström black hole. It looks like that of a non-extremal Reissner-Nordström black hole except that the two horizons merge into one. The infall velocity β into an extremal black hole reaches its maximum, the speed of light, at the horizon.

The Penrose diagram of the extremal Reissner-Nordström geometry, Figure 8.9, differs from that of the standard Reissner-Nordström geometry in having no Black Hole, White Hole, or Parallel regions. The fact that extremal black hole differs topologically from a non-extremal black hole suggests that it would be physically impossible by any causal mechanism to change a black hole from non-extremal to extremal.

8.16 Super-extremal Reissner-Nordström geometry

The Reissner-Nordström geometry with charge greater than mass,

$$Q > M, \quad (8.29)$$

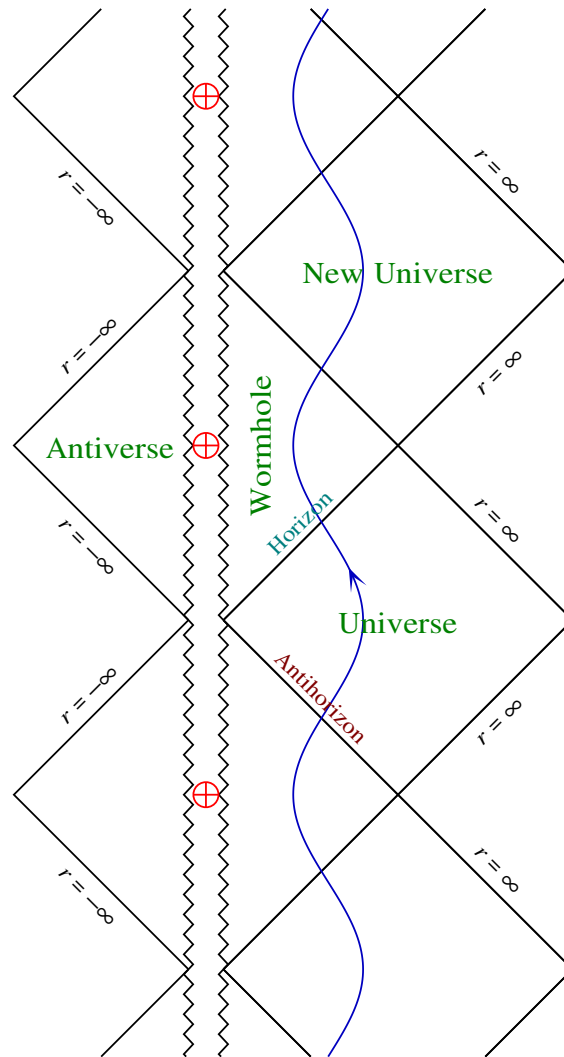


Figure 8.9 Penrose diagram of the extremal Reissner-Nordström geometry.

has no horizons. The geometry is called **super-extremal**. The change in geometry from an extremal black hole, with horizon at finite radius $r_+ = r_- = M$, to a super-extremal black hole without horizons is discontinuous. This suggests that there is no way to pack a black hole with more charge than its mass. Indeed, if you try to force additional charge into an extremal black hole, then the work needed to do so increases its mass so that the charge Q does not exceed its mass M .

Real fundamental particles nevertheless have charge far exceeding their mass. For example, the charge-to-

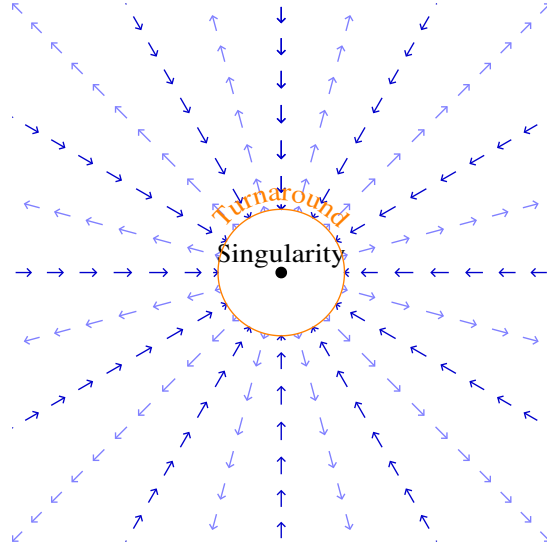


Figure 8.10 Depiction of the Gullstrand-Painlevé metric for a super-extremal Reissner-Nordström geometry, with $Q = 1.04M$. The super-extremal geometry has no horizons.

mass ratio of a proton is

$$\frac{e}{m_p} \approx 10^{18} \quad (8.30)$$

where e is the square root of the fine-structure constant $\alpha \equiv e^2/\hbar c \approx 1/137$, and $m_p \approx 10^{-19}$ is the mass of the proton in Planck units $c = G = \hbar = 1$. However, the Schwarzschild radius of such a fundamental particle is far tinier than its Compton wavelength $\sim \hbar/m$ (or its classical radius $e^2/m = \alpha\hbar/m$), so quantum mechanics, not general relativity, governs the structure of these fundamental particles.

8.17 Reissner-Nordström geometry with imaginary charge

It is possible formally to consider the Reissner-Nordström geometry with imaginary charge Q

$$Q^2 < 0. \quad (8.31)$$

This is completely unphysical. If charge were imaginary, then electromagnetic energy would be negative.

However, the Reissner-Nordström metric with $Q^2 < 0$ is well-defined, and it is possible to calculate geodesics in that geometry. What makes the geometry interesting is that the singularity, instead of being gravitationally repulsive, becomes gravitationally attractive. Thus particles, instead of bouncing off the singularity, are attracted to it, and it turns out to be possible to continue geodesics through the singularity. Mathematically, the geometry can be considered as the Kerr-Newman geometry in the limit of zero spin. In

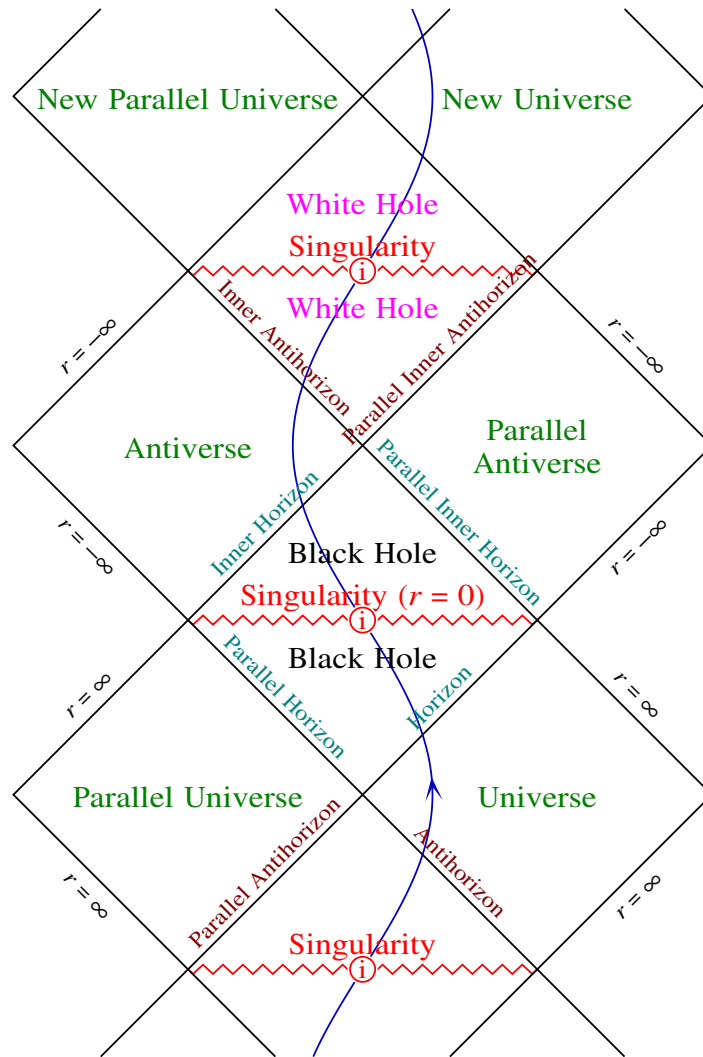


Figure 8.11 Penrose diagram of the Reissner-Nordström geometry with imaginary charge Q . If charge were imaginary, then electromagnetic energy would be negative, which is completely unphysical. But the metric is well-defined, and the spacetime is fun.

the Kerr-Newman geometry, geodesics can pass from positive to negative radius r , and the passage through the singularity of the Reissner-Nordström geometry can be regarded as this process in the limit of zero spin.

Suffice to say that it is intriguing to see what it looks like to pass through the singularity of a charged

black hole of imaginary charge, however unrealistic. The Penrose diagram is even more eventful than that for the usual Reissner-Nordström geometry.

Kerr-Newman Black Hole

The geometry of a stationary, rotating, uncharged black hole in asymptotically flat empty space was discovered unexpectedly by Roy Kerr in 1963 (Kerr, 1963). Kerr's own account of the history of the discovery is at Kerr (2009). You can read in that paper that the discovery was not mere chance: Kerr used sophisticated mathematical methods to make it. The extension to a rotating electrically charged black hole was made shortly thereafter by Ted Newman (Newman et al., 1965). Newman told me (private communication 2009) that, after seeing Kerr's work, he quickly realised that the extension to a charged black hole was straightforward. He set the problem to the graduate students in his relativity class, who became coauthors of Newman et al. (1965).

The importance of the Kerr-Newman geometry stems in part from the no-hair theorem, which states that this geometry is the unique end state of spacetime outside the horizon of an undisturbed black hole in asymptotically flat space.

9.1 Boyer-Lindquist metric

The Boyer-Lindquist metric of the Kerr-Newman geometry is

$$ds^2 = -\frac{R^2\Delta}{\rho^2} (dt - a \sin^2\theta d\phi)^2 + \frac{\rho^2}{R^2\Delta} dr^2 + \rho^2 d\theta^2 + \frac{R^4 \sin^2\theta}{\rho^2} \left(d\phi - \frac{a}{R^2} dt\right)^2, \quad (9.1)$$

where R and ρ are defined by

$$R \equiv \sqrt{r^2 + a^2}, \quad \rho \equiv \sqrt{r^2 + a^2 \cos^2\theta}, \quad (9.2)$$

and Δ is the horizon function defined by

$$\Delta \equiv 1 - \frac{2Mr}{R^2} + \frac{Q^2}{R^2}. \quad (9.3)$$

If $M = Q = 0$, so that $\Delta = 1$, the Boyer-Lindquist metric (9.1) goes over to the metric of Minkowski space expressed in oblate spheroidal coordinates, §9.2.

At large radius r , the Boyer-Lindquist metric is

$$ds^2 \rightarrow - \left(1 - \frac{2M}{r}\right) dt^2 - \frac{4aM \sin^2 \theta}{r} dt d\phi + \left(1 + \frac{2M}{r}\right) dr^2 + r^2 (d\theta^2 + \sin^2 \theta d\phi^2) . \quad (9.4)$$

By comparison, the weak-field metric in Newtonian gauge, equation (27.62), around an object of mass M and angular momentum L takes the form

$$ds^2 = - (1 + 2\Psi) dt^2 - 2Wr \sin \theta dt d\phi + (1 - 2\Phi)(dr^2 + r^2 d\theta^2) , \quad (9.5)$$

where, from equations (27.80) and (27.87), the scalar Ψ , Φ and vector W potentials are

$$\Psi = \Phi = -\frac{M}{r} , \quad W = -\frac{2L \sin \theta}{r^2} . \quad (9.6)$$

The asymptotic Boyer-Lindquist metric (9.4) is not quite in the Newtonian form (9.5), but a transformation of the radial coordinate brings it to Newtonian form, Exercise 7.1. Comparison of the two metrics establishes that M is the mass of the black hole and $a = L/M$ is its angular momentum per unit mass. For positive a , the black hole rotates right-handedly about its polar axis $\theta = 0$.

The Boyer-Lindquist line-element (9.1) defines not only a metric but also a tetrad. The Boyer-Lindquist coordinates and tetrad are carefully chosen to exhibit the symmetries of the geometry. In the locally inertial frame defined by the Boyer-Lindquist tetrad, the energy-momentum tensor (which is non-vanishing for charged Kerr-Newman) and the Weyl tensor are both diagonal. These assertions become apparent only in the tetrad frame, §19.3, and are obscure in the coordinate frame.

9.2 Oblate spheroidal coordinates

Boyer-Lindquist coordinates r, θ, ϕ are **oblate spheroidal** coordinates (not polar coordinates). Corresponding Cartesian coordinates are

$$\begin{aligned} x &= R \sin \theta \cos \phi , \\ y &= R \sin \theta \sin \phi , \\ z &= r \cos \theta . \end{aligned} \quad (9.7)$$

Surfaces of constant r are confocal oblate spheroids, satisfying

$$\frac{x^2 + y^2}{R^2} + \frac{z^2}{r^2} = 1 . \quad (9.8)$$

Equation (9.8) implies that the spheroidal coordinate r is given in terms of x, y, z by the quadratic equation

$$r^4 - r^2(x^2 + y^2 + z^2 - a^2) - a^2 z^2 = 0 . \quad (9.9)$$

Figure 9.1 illustrates the spatial geometry of a Kerr black hole, and of a Kerr-Newman black hole, in Boyer-Lindquist coordinates.

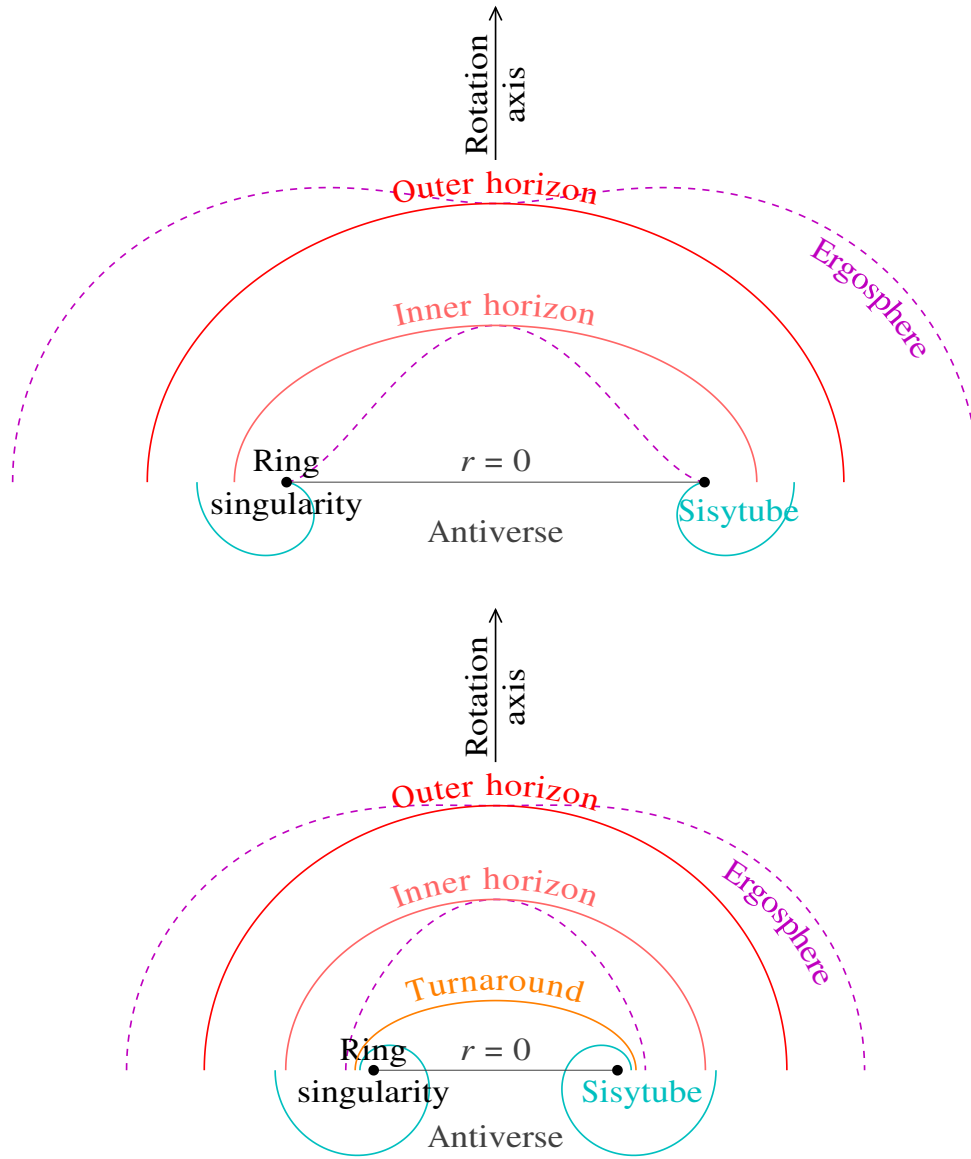


Figure 9.1 Spatial geometry of (upper) a Kerr black hole with spin parameter $a = 0.96M$, and (lower) a Kerr-Newman black hole with charge $Q = 0.8M$ and spin parameter $a = 0.56M$. The upper half of each diagram shows $r \geq 0$, while the lower half shows $r \leq 0$, the Antiverse. The outer and inner horizons are confocal oblate spheroids whose focus is the ring singularity. For the Kerr geometry, the turnaround radius is at $r = 0$. The Sisytube is a torus enclosing the ring singularity, that contains closed timelike curves.

9.3 Time and rotation symmetries

The Boyer-Lindquist metric coefficients are independent of the time coordinate t and of the azimuthal angle ϕ . This shows that the Kerr-Newman geometry has time translation symmetry, and rotational symmetry about its azimuthal axis. The time and rotation symmetries means that the tangent vectors $\mathbf{e}_t$ and $\mathbf{e}_\phi$ in Boyer-Lindquist coordinates are Killing vectors. It follows that their scalar products

$$\begin{aligned} \mathbf{e}_t \cdot \mathbf{e}_t = g_{tt} &= -\frac{1}{\rho^2} (R^2 \Delta - a^2 \sin^2 \theta) , \\ \mathbf{e}_t \cdot \mathbf{e}_\phi = g_{t\phi} &= -\frac{a R^2 \sin^2 \theta}{\rho^2} (1 - \Delta) , \\ \mathbf{e}_\phi \cdot \mathbf{e}_\phi = g_{\phi\phi} &= \frac{R^2 \sin^2 \theta}{\rho^2} (R^2 - a^2 \sin^2 \theta \Delta) , \end{aligned} \quad (9.10)$$

are all gauge-invariant scalar quantities. As will be seen below, $g_{tt} = 0$ defines the boundary of ergospheres, $g_{t\phi} = 0$ defines the turnaround radius, and $g_{\phi\phi} = 0$ defines the boundary of the sisytube, the toroidal region containing closed timelike curves.

The Boyer-Lindquist time t and azimuthal angle ϕ are arranged further to satisfy the condition that $\mathbf{e}_t$ and $\mathbf{e}_\phi$ are each orthogonal to both $\mathbf{e}_r$ and $\mathbf{e}_\theta$.

9.4 Ring singularity

The Kerr-Newman geometry contains a **ring singularity** where the Weyl tensor (9.26) diverges, $\rho = 0$, or equivalently at

$$\boxed{r = 0 \text{ and } \theta = \pi/2} . \quad (9.11)$$

The ring singularity is at the focus of the confocal ellipsoids of the Boyer-Lindquist metric. Physically, the singularity is kept open by the centrifugal force.

Figure 9.2 illustrates contours of constant ρ in a Kerr black hole.

9.5 Horizons

The horizon of a Kerr-Newman black hole rotates, as observed by a distant observer, so it is incorrect to try to solve for the location of the horizon by assuming that the horizon is at rest. The worldline of a photon that sits on the horizon, battling against the inflow of space, remains at fixed radius r and polar angle θ , but it moves in time t and azimuthal angle ϕ . The photon's 4-velocity is $v^\mu = \{v^t, 0, 0, v^\phi\}$, and the condition that it is on a null geodesic is

$$0 = v_\mu v^\mu = g_{\mu\nu} v^\mu v^\nu = g_{tt} (v^t)^2 + 2 g_{t\phi} v^t v^\phi + g_{\phi\phi} (v^\phi)^2 . \quad (9.12)$$

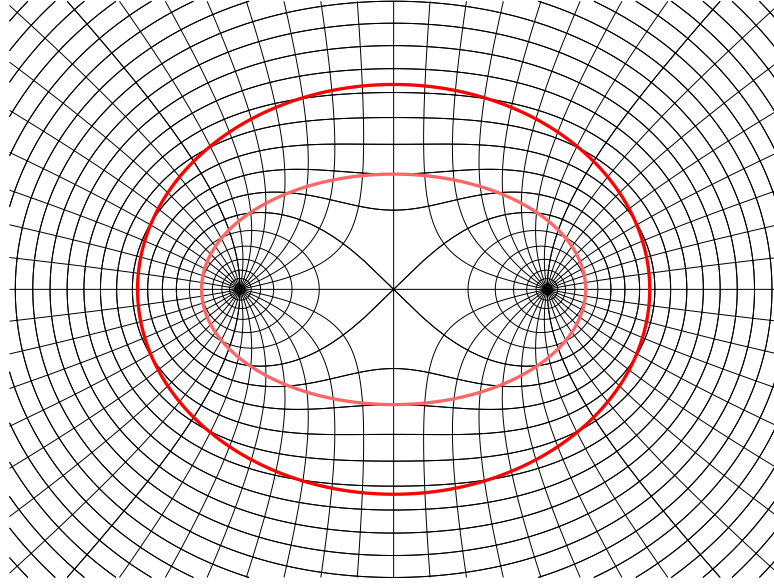


Figure 9.2 *Not* a mouse's eye view of a snake coming down its mousehole, uhoh. Contours of constant ρ and their covariant normals $\partial\rho/\partial x^\mu$ in a spatial cross-section of a Kerr black hole of spin parameter $a = 0.96M$, in Boyer-Lindquist coordinates. The thicker contours are the outer and inner horizons, which are confocal spheroids with the ring singularity at their focus. The ring singularity is at $\rho = 0$, the snake's eyes.

This equation has solutions provided that the determinant of the 2×2 matrix of metric coefficients in t and ϕ is less than or equal to zero (why?). The determinant is

$$g_{tt}g_{\phi\phi} - g_{t\phi}^2 = -R^2 \sin^2\theta \Delta, \quad (9.13)$$

where Δ is the horizon function defined above, equation (9.3). Thus if $\Delta \geq 0$, then there exist null geodesics such that a photon can be instantaneously at rest in r and θ , whereas if $\Delta < 0$, then no such geodesics exist. The boundary

$$\Delta = 0 \quad (9.14)$$

defines the location of horizons. With Δ given by equation (9.3), equation (9.14) gives **outer** and **inner horizons** at

$$\boxed{r_{\pm} = M \pm \sqrt{M^2 - Q^2 - a^2}}. \quad (9.15)$$

Between the horizons Δ is negative, and photons cannot be at rest. This is consistent with the picture that space is falling faster than light between the horizons.

9.6 Angular velocity of the horizon

The angular velocity of the horizon as observed by observers at rest at infinity can be read off directly from the Boyer-Lindquist metric (9.1). The horizon is at $dr = d\theta = 0$ and $\Delta = 0$, and then the null condition $ds^2 = 0$ implies that the angular velocity is

$$\frac{d\phi}{dt} = \frac{a}{R^2} . \quad (9.16)$$

The derivative is with respect to the proper time t of observers at rest at infinity, so this is the angular velocity observed by such observers.

9.7 Ergospheres

There are finite regions, just outside the outer horizon and just inside the inner horizon, within which the worldline of an object at rest, $dr = d\theta = d\phi = 0$, is spacelike. These regions, called **ergospheres**, are places where nothing can remain at rest (the place where little children come from). Objects can escape from within the outer ergosphere (whereas they cannot escape from within the outer horizon), but they cannot remain at rest there. A distant observer will see any object within the outer ergosphere being dragged around by the rotation of the black hole. The direction of dragging is the same as the rotation direction of the black hole in both outer and inner ergospheres.

The boundary of the ergosphere is at

$$g_{tt} = 0 , \quad (9.17)$$

which occurs where

$$R^2 \Delta = a^2 \sin^2 \theta . \quad (9.18)$$

Equation (9.18) has two solutions, the outer and inner ergospheres. The outer and inner ergospheres touch respectively the outer and inner horizons at the poles, $\theta = 0$ and π .

9.8 Turnaround radius

The turnaround radius is the radius inside the inner horizon at which infallers who fall from zero velocity and zero angular momentum at infinity turn around. The radius is at

$$g_{t\phi} = 0 , \quad (9.19)$$

which occurs where $\Delta = 1$, or equivalently at

$$r = \frac{Q^2}{2M} . \quad (9.20)$$

In the uncharged Kerr geometry, the turnaround radius is at zero radius, $r = 0$, but in the Kerr-Newman geometry the turnaround radius is at positive radius.

9.9 Antiverse

The surface at zero radius, $r = 0$, forms a disk bounded by the ring singularity. Objects can pass through this disk into the region at negative radius, $r < 0$, the **Antiverse**.

The Boyer-Lindquist metric (9.1) is unchanged by a symmetry transformation that simultaneously flips the sign both of the radius and mass, $r \rightarrow -r$ and $M \rightarrow -M$. Thus the Boyer-Lindquist geometry at negative r with positive mass is equivalent to the geometry at positive r with negative mass. In effect, the Boyer-Lindquist metric with negative r describes a rotating black hole of negative mass

$$M < 0 . \quad (9.21)$$

9.10 Sisytube

Inside the inner horizon there is a toroidal region around the ring singularity, which I call the **sisytube**, within which the light cone in t - ϕ coordinates opens up to the point that ϕ as well as t is a timelike coordinate. In the Wormhole, the direction of increasing proper time along t is t increasing, and along ϕ is ϕ decreasing, which is retrograde. In the Parallel Wormhole, the direction of increasing proper time along t is t decreasing, and along ϕ is ϕ increasing, which is again retrograde. Within the toroidal region, there exist timelike trajectories that go either forwards or backwards in coordinate time t as they wind retrograde around the toroidal tunnel. Because the ϕ coordinate is periodic, these timelike curves connect not only the past to the future (the usual case), but also the future to the past, which violates causality. In particular, as first pointed out by Carter (1968), there exist **closed timelike curves** (CTCs), trajectories that connect to themselves, connecting their own future to their own past, and repeating interminably, like Sisyphus pushing his rock up the mountain.

The boundary of the sisytube torus is at

$$g_{\phi\phi} = 0 , \quad (9.22)$$

which occurs where

$$R^2 = a^2 \sin^2 \theta \Delta . \quad (9.23)$$

In the uncharged Kerr geometry the sisytube is entirely at negative radius, $r < 0$, but in the Kerr-Newman geometry the sisytube extends to positive radius, Figure 9.1.

9.11 Extremal Kerr-Newman geometry

The Kerr-Newman geometry is called extremal when the outer and inner horizons coincide, $r_+ = r_-$, which occurs where

$$M^2 = Q^2 + a^2 . \quad (9.24)$$

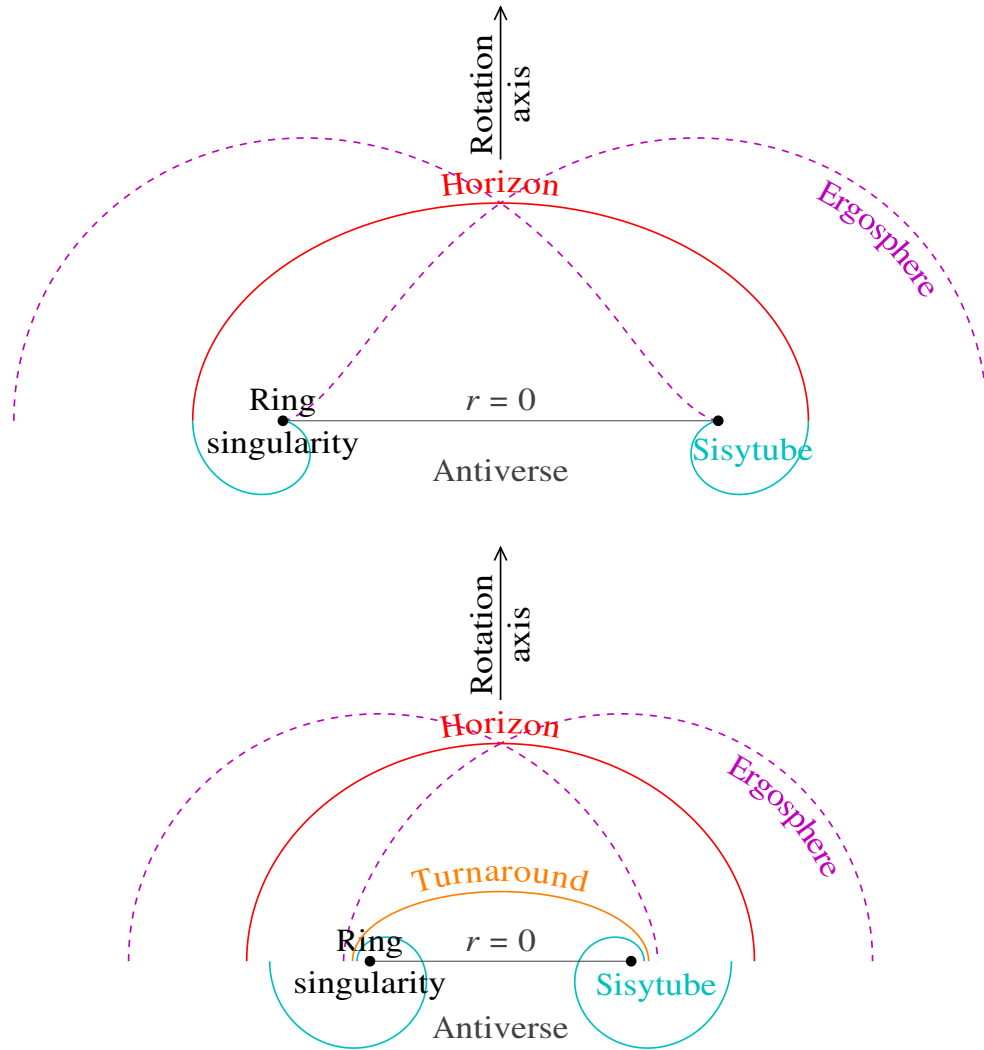


Figure 9.3 Spatial geometry of (upper) an extremal ($a = M$) Kerr black hole, and (lower) an extremal Kerr-Newman black hole with charge $Q = 0.8M$ and spin parameter $a = 0.6M$.

Figure 9.3 illustrates the structure of an extremal Kerr (uncharged) black hole, and an extremal Kerr-Newman (charged) black hole.

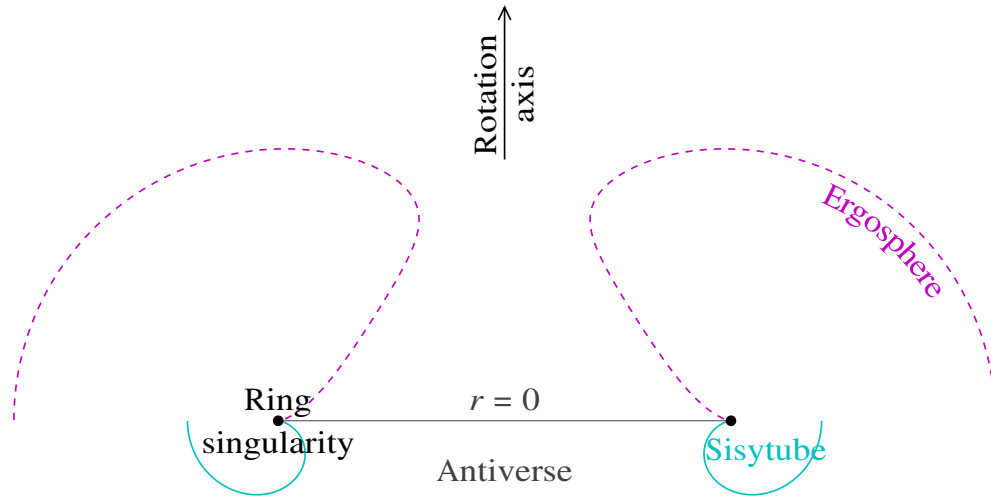


Figure 9.4 Spatial geometry of a super-extremal Kerr black hole with spin parameter $a = 1.04M$. A super-extremal black hole has no horizons.

9.12 Super-extremal Kerr-Newman geometry

If $M^2 < Q^2 + a^2$, then there are no horizons. The geometry is called super-extremal. Figure 9.4 illustrates the structure of a super-extremal Kerr black hole. A super-extremal black hole has a naked ring singularity, and CTCs in a sisytube unhidden by a horizon.

9.13 Energy-momentum tensor

The coordinate-frame Einstein tensor of the Kerr-Newman geometry in Boyer-Lindquist coordinates is a bit of a mess. The trick of raising one index, which for the Reissner-Nordström metric brought the Einstein tensor to diagonal form, equation (8.5), fails for Boyer-Lindquist (because the Boyer-Lindquist metric is not diagonal). The problem is endemic to the coordinate approach to general relativity. After tetrads it will emerge that, in the Boyer-Lindquist tetrad, the Einstein tensor is diagonal, and that the proper density ρ , the proper radial pressure p_r , and the proper transverse pressure $p_\perp$ in that frame are (do not confuse the notation ρ for proper density with the radial parameter ρ , equation (9.2), of the Boyer-Lindquist metric)

$$\rho = -p_r = p_\perp = \frac{Q^2}{8\pi\rho^4} . \quad (9.25)$$

This looks like the energy-momentum tensor (8.5) of the Reissner-Nordström geometry with the replacement $r \rightarrow \rho$. The energy-momentum is that of an electric field produced by a charge Q seemingly located at the ring singularity.

9.14 Weyl tensor

The Weyl tensor of the Kerr-Newman geometry in Boyer-Lindquist coordinates is likewise a mess. After tetrads, it will emerge that the 10 components of the Weyl tensor can be decomposed into 5 complex components of spin 0, ± 1 , and ± 2 . In the Boyer-Lindquist tetrad, the only non-vanishing component is the spin-0 component, the Weyl scalar C , but in contrast to the Schwarzschild and Reissner-Nordström geometries the spin-0 component is complex, not real:

$$C = -\frac{1}{(r - ia \cos \theta)^3} \left(M - \frac{Q^2}{r + ia \cos \theta} \right) . \quad (9.26)$$

9.15 Electromagnetic field

The expression for the electromagnetic field in Boyer-Lindquist coordinates is again a mess. After tetrads, it will emerge that, in the Boyer-Lindquist tetrad, the electromagnetic field is purely radial, and the electromagnetic potential has only a time component. For reference, the covariant electromagnetic potential A_μ in the Boyer-Lindquist coordinate (not tetrad) frame is

$$A_\mu = \frac{Qr}{\rho^2} \left\{ -1, 0, 0, \frac{a \sin \theta}{R\sqrt{\Delta}} \right\} . \quad (9.27)$$

9.16 Principal null congruences

The Kerr-Newman geometry admits a special set of space-filling, non-overlapping null geodesics called the principal outgoing and ingoing null congruences. These are the directions with respect to which the Weyl tensor and the electric field vector align. Photons that hold steady on the outer horizon are on the principal outgoing null congruence. The construction and special character of the principal null congruences will be demonstrated after tetrads, in §23.6.

Geodesics along the principal null congruences satisfy

$$d\theta = d\phi - \omega dt = 0 , \quad (9.28)$$

where $\omega = a/R^2$ is the azimuthal angular velocity of the geodesics through the coordinates. The Boyer-Lindquist line-element (9.1) is specifically constructed so that it aligns with the principal null congruences.

9.17 Finkelstein coordinates

Along the principal outgoing and ingoing null congruences, where equations (9.28) hold, the Boyer-Lindquist metric (9.1) reduces to

$$ds^2 = \frac{\rho^2 \Delta}{R^2} \left(-dt^2 + \frac{dr^2}{\Delta^2} \right) . \quad (9.29)$$

A tortoise coordinate r^* in the Kerr-Newman geometry may be defined analogously to that (8.16) in the Reissner-Nordström geometry,

$$r^* \equiv \int \frac{dr}{\Delta} , \quad (9.30)$$

which integrates to the same expressions (8.16) and (8.17) in terms of horizon radii $r_{\pm}$ and surface gravities $\kappa_{\pm}$ as in the Reissner-Nordström geometry. Principal outgoing and ingoing null geodesics follow

$$\begin{aligned} r^* - t &= \text{constant} & \text{outgoing} , \\ r^* + t &= \text{constant} & \text{ingoing} . \end{aligned} \quad (9.31)$$

A Finkelstein time coordinate t_F can be defined as in the Reissner-Nordström geometry, equation (8.19). Likewise, Kruskal-Szekeres coordinates can be defined as in the Reissner-Nordström geometry, equations (8.21) and (8.22). The Finkelstein and Kruskal spacetime diagrams for the Kerr-Newman geometry look identical to those of the Reissner-Nordström geometry (if the horizon radii $r_{\pm}$ are the same), Figures 8.3 and 8.4. The discussion in §§8.7–8.9 carries through essentially unchanged for the Kerr-Newman geometry.

The behaviour of geodesics in the angular direction is more complicated in the Kerr-Newman than Reissner-Nordström geometry, but this complexity is hidden in the Finkelstein and Kruskal diagrams.

9.18 Doran coordinates

For the Kerr-Newman geometry, the analogue of the Gullstrand-Painlevé metric is the Doran (2000) metric

$$ds^2 = -dt_{\text{ff}}^2 + \left[\frac{\rho}{R} dr - \beta \frac{R}{\rho} (dt_{\text{ff}} - a \sin^2 \theta d\phi_{\text{ff}}) \right]^2 + \rho^2 d\theta^2 + R^2 \sin^2 \theta d\phi_{\text{ff}}^2 , \quad (9.32)$$

where the free-fall time t_{ff} and azimuthal angle ϕ_{ff} are related to the Boyer-Lindquist time t and azimuthal angle ϕ by

$$dt_{\text{ff}} = dt - \frac{\beta}{1 - \beta^2} dr , \quad d\phi_{\text{ff}} = d\phi - \frac{a\beta}{R^2(1 - \beta^2)} dr . \quad (9.33)$$

The free-fall time t_{ff} is the proper time experienced by persons who free-fall from rest at infinity, with zero angular momentum. They follow trajectories of fixed θ and ϕ_{ff} , with radial velocity $dr/dt_{\text{ff}} = \beta R^2/\rho^2$. The 4-velocity $u^{\nu} \equiv dx^{\nu}/d\tau$ of such free-falling observers is

$$u^{t_{\text{ff}}} = 1 , \quad u^r = \frac{R^2 \beta}{\rho^2} , \quad u^{\theta} = 0 , \quad u^{\phi_{\text{ff}}} = 0 . \quad (9.34)$$

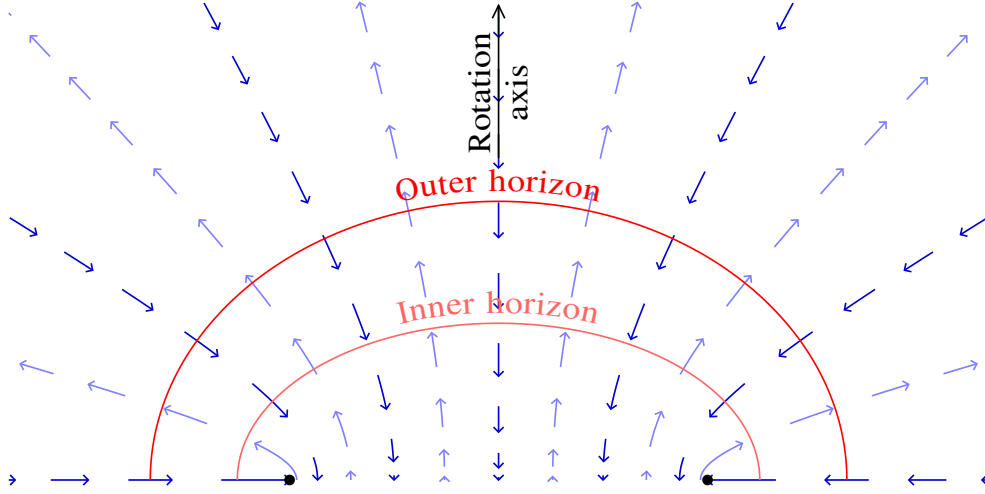


Figure 9.5 Spatial geometry of a Kerr black hole with spin parameter $a = 0.96M$. The arrows show the velocity β in the Doran metric. The flow follows lines of constant θ , which form nested hyperboloids orthogonal to and confocal with the nested spheroids of constant r .

For the Kerr-Newman geometry, the velocity β is

$$\beta = \mp \sqrt{1 - \Delta} = \mp \frac{\sqrt{2Mr - Q^2}}{R} \quad (9.35)$$

where the $\mp$ sign is $-$ (infalling) for black hole solutions, and $+$ (outfalling) for white hole solutions.

Horizons occur where the magnitude of the velocity β equals the speed of light

$$\beta = \mp 1 . \quad (9.36)$$

The boundaries of ergospheres occur where the velocity is

$$\beta = \mp \frac{\rho}{R} . \quad (9.37)$$

The turnaround radius is where the velocity is zero

$$\beta = 0 . \quad (9.38)$$

The sisytube is bounded by the imaginary velocity

$$\beta = i \frac{\rho}{a \sin \theta} . \quad (9.39)$$

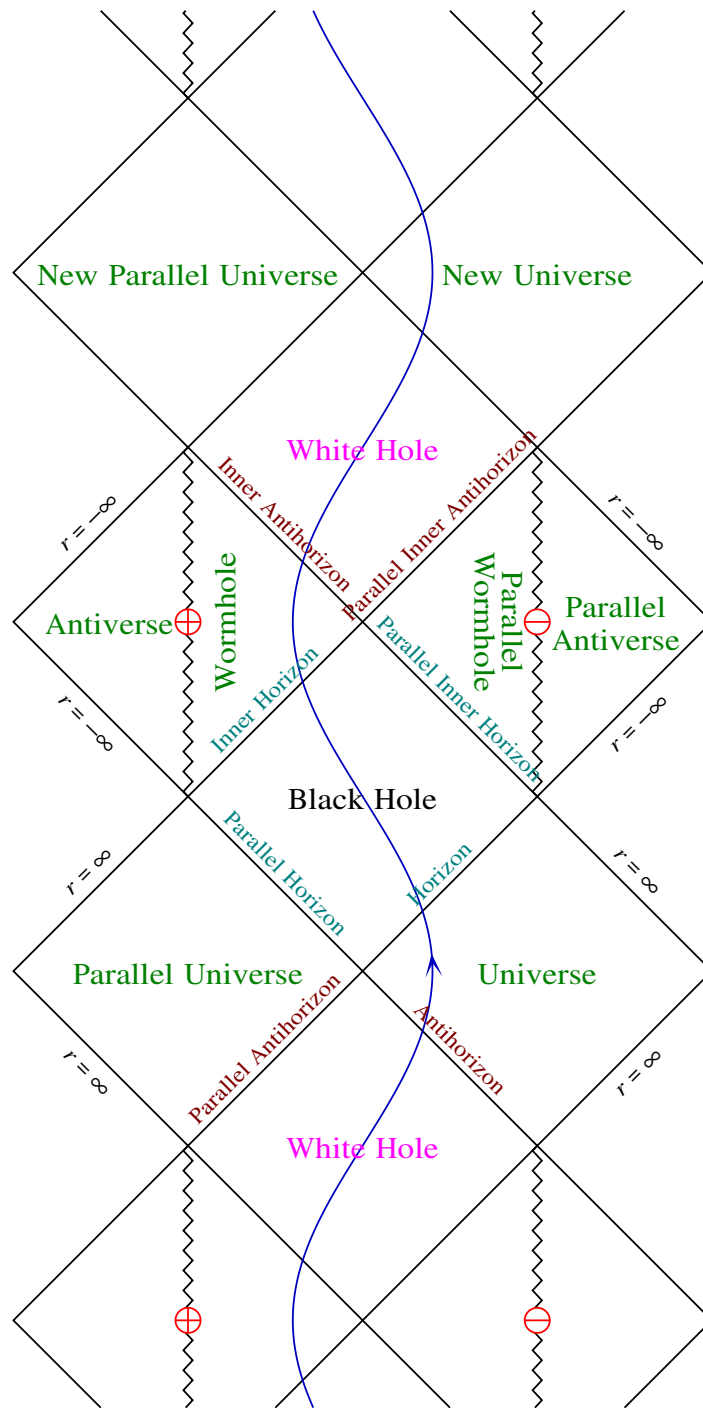


Figure 9.6 Penrose diagram of the Kerr-Newman geometry. The diagram is similar to that of the Reissner-Nordström geometry, except that it is possible to pass through the disk at $r = 0$ from the Wormhole region into the Antiverse region. This Penrose diagram, which represents a slice at fixed θ and ϕ , does not capture the full richness of the geometry, which contains closed timelike curves in a torus around the ring singularity, the sisytube.

9.19 Penrose diagram

The Penrose diagram of the Kerr-Newman geometry, Figure 9.6, resembles that of the Reissner-Nordström geometry, Figure 8.6, except that in the Kerr-Newman geometry an infaller can reach the Antiverse by passing through the disk at $r = 0$ bounded by the ring singularity. In the Reissner-Nordström geometry, the ring singularity shrinks to a point, and passing into the Antiverse would require passing through the singularity itself.

Concept Questions

1. What does it mean that the Universe is expanding?
2. Does the expansion affect the solar system or the Milky Way?
3. How far out do you have to go before the expansion is evident?
4. What is the Universe expanding into?
5. In what sense is the Hubble constant constant?
6. Does our Universe have a centre, and if so where is it?
7. What evidence suggests that the Universe at large is homogeneous and isotropic?
8. How can the Cosmic Microwave Background (CMB) be construed as evidence for homogeneity and isotropy given that it provides information only over a 2D surface on the sky?
9. What is thermodynamic equilibrium? What evidence suggests that the early Universe was in thermodynamic equilibrium?
10. What are cosmological parameters?
11. What cosmological parameters can or cannot be measured from the power spectrum of fluctuations of the CMB?
12. Friedmann-Lemaître-Robertson-Walker (FLRW) universes are characterized as closed, flat, or open. Does flat here mean the same as flat Minkowski space?
13. What is it that astronomers call dark matter?
14. What is the primary evidence for the existence of non-baryonic cold dark matter?
15. How can astronomers detect dark matter in galaxies or clusters of galaxies?
16. How can cosmologists claim that the Universe is dominated by not one but two distinct kinds of mysterious mass-energy, dark matter and dark energy, neither of which has been observed in the laboratory?
17. What key property or properties distinguish dark energy from dark matter?
18. A FLRW universe conserves entropy. Is that true? If so, can the entropy of the Universe increase?
19. Does the annihilation of electron-positron pairs into photons generate entropy in the early Universe, as its temperature cools through 1 MeV?
20. How does the wavelength of light change with the expansion of the Universe?
21. How does the temperature of the CMB change with the expansion of the Universe?

22. How does a blackbody (Planck) distribution change with the expansion of the Universe? What about a nonrelativistic distribution? What about a semi-relativistic distribution?
23. What is the horizon of our Universe? What is the Hubble distance?
24. What happens beyond the horizon of our Universe?
25. What caused the Big Bang?
26. What happened before the Big Bang?
27. What will be the fate of the Universe?

What's important?

1. The Cosmic Microwave Background (CMB) indicates that the early ($\approx 400,000$ year old) Universe was (a) uniform to a few $\times 10^{-5}$, and (b) in thermodynamic equilibrium. This indicates that the Universe was once very simple.

It is this simplicity that makes it possible to model the early Universe with some degree of confidence.

2. The power spectrum of fluctuations of the CMB has enabled precise measurements of cosmological parameters.
3. There is a remarkable concordance of evidence from a broad range of astronomical observations — supernovae, big bang nucleosynthesis, the clustering of galaxies, the abundances of clusters of galaxies, measurements of the Hubble constant from Cepheid variables and supernovae, and the ages of the oldest stars.
4. Observational evidence is consistent with the predictions of the theory of inflation in its simplest form — the expansion of the Universe, the spatial flatness of the Universe, the near uniformity of temperature fluctuations of the CMB (the horizon problem), the presence of acoustic peaks and troughs in the power spectrum of fluctuations of the CMB, the near power law shape of the power spectrum at large scales, its spectral index (tilt), the gaussian distribution of fluctuations at large scales.
5. What is non-baryonic dark matter?
6. What is dark energy? What is its equation of state $w \equiv p/\rho$, and how does w evolve with time?

Homogeneous, Isotropic Cosmology

10.1 Observational basis

Since 1998, observations have converged on a **Standard “ Λ CDM” Model of Cosmology**, a spatially flat Universe dominated by gravitationally repulsive dark energy whose equation of state is consistent with that of a cosmological constant (Λ), and by gravitationally attractive non-baryonic cold dark matter (CDM). The mass-energy of the Standard Model of the Universe consists of 70% dark energy, 25% non-baryonic cold dark matter (CDM), 5% baryonic matter, and a sprinkling of photons and neutrinos. The designation “baryonic” is conventional but misleading: it refers to all atomic matter, including not only baryons (nuclei), but also nonrelativistic charged leptons (electrons).

10.1.1 The expansion of the Universe

The **Hubble diagram**, a diagram of distance versus redshift of distant astronomical objects, indicates that the Universe is expanding.

Hubble’s law states that galaxies are receding with velocity proportional to distance, $v = H_0 d$, with constant of proportionality the Hubble constant H_0 (the 0 subscript signifies the present day value). Hubble’s law was first proposed by Georges Lemaître (1927) and by Edwin Hubble (1929) on the basis of observations.

The recession velocity v of an astronomical object can be determined with some precision from the redshift of its spectral lines, but its distance d is more difficult to measure, because astronomical objects, such as galaxies, typically have a wide range of intrinsic luminosities. Hubble estimated distances to galaxies using Cepheid variable stars, which had been discovered by Henrietta Leavitt (1912) to have periods proportional to their luminosities. A good distance estimator should be a “standard candle” of predictable luminosity, and it should be bright, so that it can be seen over cosmological distances.

The best modern Hubble diagram is that of Type Ia supernovae, illustrated in Figure 10.1, from data tabulated by Scolnic et al. (2018). A Type Ia supernova is thought to represent the thermonuclear explosion of a white dwarf star that through accretion from a companion star reaches the Chandrasekhar mass limit of $1.4 M_\odot$. Having a similar origin, such supernovae approximate standard candles (or standard bombs) having the same luminosity. Actually, some variation in luminosity is observed, which may be associated with the

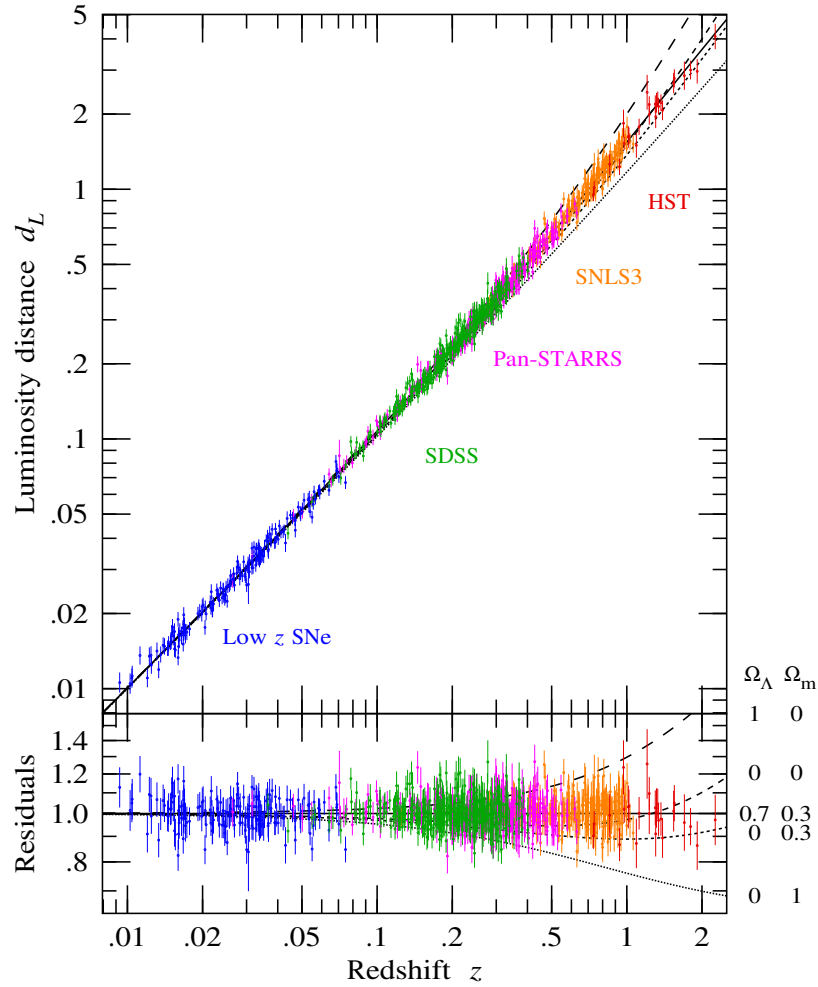


Figure 10.1 Hubble diagram of 1048 Type Ia supernovae from a compilation of surveys, from data tabulated by Scolnic et al. (2018). The vertical axis is the luminosity distance d_L in units of the present-day Hubble distance c/H_0 . The bottom panel shows residuals. The various smooth curves are 5 theoretical model Hubble diagrams, with parameters as indicated. The solid line is a flat Λ CDM model with $\Omega_\Lambda = 0.7$ and $\Omega_m = 0.3$.

amount of ^{56}Ni synthesized in the explosion, and which can be corrected at least in part through an empirical relation between luminosity and how rapidly the lightcurve decays (higher luminosity supernovae decay more slowly).

10.1.2 The acceleration of the Universe

Since light takes time to travel from distant parts of the Universe to astronomers here on Earth, the higher the redshift of an object, the further back in time astronomers are seeing.

In 1998 two teams, the Supernova Cosmology Project (Perlmutter, S. et al. (Supernova Cosmology Project, 32 authors), 1999), and the High- z Supernova Search team (Riess et al., 1998), precipitated the revolution that led to the Standard Model of Cosmology. They reported that observations of Type Ia supernova at high redshift indicated that the Universe is not only expanding, but also accelerating. The acceleration requires the mass-energy density of the Universe to be dominated at the present time by a gravitationally repulsive component, such as a cosmological constant Λ .

In the Hubble diagram of Type Ia supernova shown in Figure 10.1, the fitted curve is a best-fit flat cosmological model containing a cosmological constant and matter.

10.1.3 The Cosmic Microwave Background (CMB)

The single most powerful observational constraints on the Universe come from the Cosmic Microwave Background (CMB). Modern observations of the CMB have ushered in an era of precision cosmology, where key cosmological parameters are being measured with percent level uncertainties.

The CMB was discovered serendipitously by Arno Penzias & Robert Wilson (1965), who were puzzled by an apparently uniform excess temperature from a horn antenna, 6 metres in size, tuned to a wavelength of ~ 7 cm, that they had built to detect radio waves. They were unaware that Robert Dicke's group at Princeton had already realised that a hot Big Bang would have left a remnant of blackbody radiation filling the Universe, with a present-day temperature of a few Kelvin, and were setting about to try to detect it. When Penzias heard about Dicke's work, he and Wilson quickly realised that their observations fit what the Princeton group were predicting. The observations of Penzias and Wilson (1965) were published along with a theoretical explanation by Dicke et al. (1965) in back-to-back papers in an issue of the *Astrophysical Journal Letters*.

Dicke et al. (1965) argued that the temperature of the expanding Universe must have been higher in the past, and there must have been a time before which the temperature was high enough to ionize hydrogen, about 3,000 K. Before this time, called recombination, hydrogen and other elements would have been mostly ionized. The CMB comes to us from the time of recombination, when the Universe transitioned from being mainly ionized, and therefore opaque, to being mainly neutral, and therefore transparent. Recombination occurred when the Universe was about 400,000 years old, and the CMB has streamed essentially freely through the Universe since that time. Thus the CMB provides a snapshot of the Universe at recombination.

The CMB spectrum peaks in microwaves, which are absorbed by water vapour in the atmosphere. Modern observations of the CMB are therefore made using satellites, or with balloons, or at high-altitude sites with low water vapour, such as the South Pole, or the Atacama Desert in Chile.

The characteristics of the CMB measured from modern observations are as follows.

The CMB has a remarkably precise black body spectrum, Figure 10.2, with temperature (Fixsen, 2009)

$$T_0 = 2.72548 \pm 0.00057 \text{ K} . \quad (10.1)$$

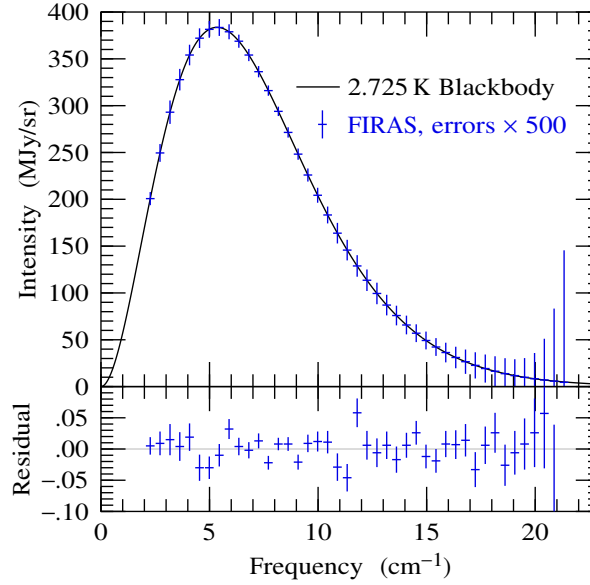


Figure 10.2 COBE/FIRAS observations of the (monopole) spectrum of the CMB. The observations (points with error bars multiplied by 500) fit extraordinarily well to a blackbody, or Planck, spectrum at a temperature of 2.725 K (solid line). In practice, the spectrum was observed by switching between the CMB sky and a blackbody calibrator. The lower graph shows the measured deviation from the blackbody calibrator. Data from https://lambda.gsfc.nasa.gov/product/cobe/firas_monopole_get.cfm.

The CMB shows a dipole anisotropy of $\Delta T = 3.355 \pm 0.008$ mK, implying that the solar system is moving through the CMB at velocity (Jarosik et al., 2011)

$$v = 369.1 \text{ km s}^{-1} \quad \text{in Galactic coordinate direction} \quad \{l, b\} = \{263^\circ 99 \pm 0.14, 48^\circ 26 \pm 0.03\}. \quad (10.2)$$

After dipole subtraction, the temperature of the CMB over the sky is uniform to a few parts in 10^5 .

The power spectrum of temperature T fluctuations shows a scale-invariant spectrum at large scales, and prominent acoustic peaks at smaller scales, Figure 10.3. The power spectrum fits astonishingly well to predictions based on the theory of inflation, §10.22, in its simplest form. The power spectrum yields precision measurements of some basic cosmological parameters, notably the densities of the principal contributions to the energy-density of the Universe: dark energy, non-baryonic cold dark matter, and baryons.

Figure 10.4, taken from the Planck Collaboration (2020, Fig. 18), shows a compilation of CMB power spectra as of 2018. There are two independent modes of polarization of opposite parity, electric “ E ” ($(-)^{\ell}$ parity) modes and magnetic “ B ” ($(-)^{\ell+1}$ parity) modes. There are corresponding E -mode and B -mode power spectra. The temperature fluctuation T has electric parity, so of the cross-power spectra between temperature T and E and B fluctuations, only the T - E cross-power is expected to be non-vanishing (if the Universe at large is not only homogeneous but also parity symmetric). The TE and EE power spectra come

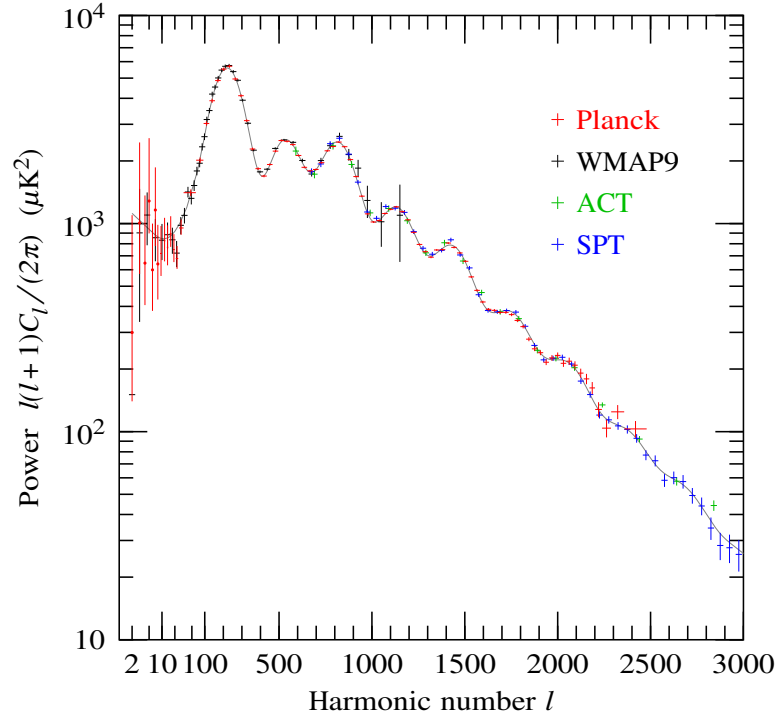


Figure 10.3 Power spectrum of fluctuations in the CMB from observations with the Planck satellite (Planck Collaboration, 2013), WMAP (Hinshaw et al., 2012), the Atacama Cosmology Telescope (Das et al., 2011), and the South Pole Telescope (Keisler et al., 2011). The plot is logarithmic in harmonic number l up to 100, linear thereafter. The fit is a best-fit flat Λ CDM model.

from Thomson scattering of CMB photons by ionized gas, mostly from around the epoch of recombination, but with some contribution from the subsequent epoch of reionization when the first stars formed.

10.1.4 The clustering of galaxies

The clustering of galaxies shows a power spectrum in good agreement with the Standard Model, Figure 10.5.

Historically, the principal evidence for non-baryonic cold dark matter was comparison between the power spectra of galaxies versus CMB. How can tiny fluctuations in the CMB grow into the observed fluctuations in matter today in only the age of the Universe? The answer was, non-baryonic dark matter that begins to cluster before recombination, when the CMB was released.

The interpretation of the power spectrum of galaxies is complicated by the facts that galaxies have undergone non-linear clustering at smaller scales, and that galaxies are a biased tracer of mass. However, the pattern of clustering at large, linear scales retains an imprint of baryonic acoustic oscillations (BAO) analogous to those in the CMB. Observations from large galaxy surveys have been able to measure the predicted

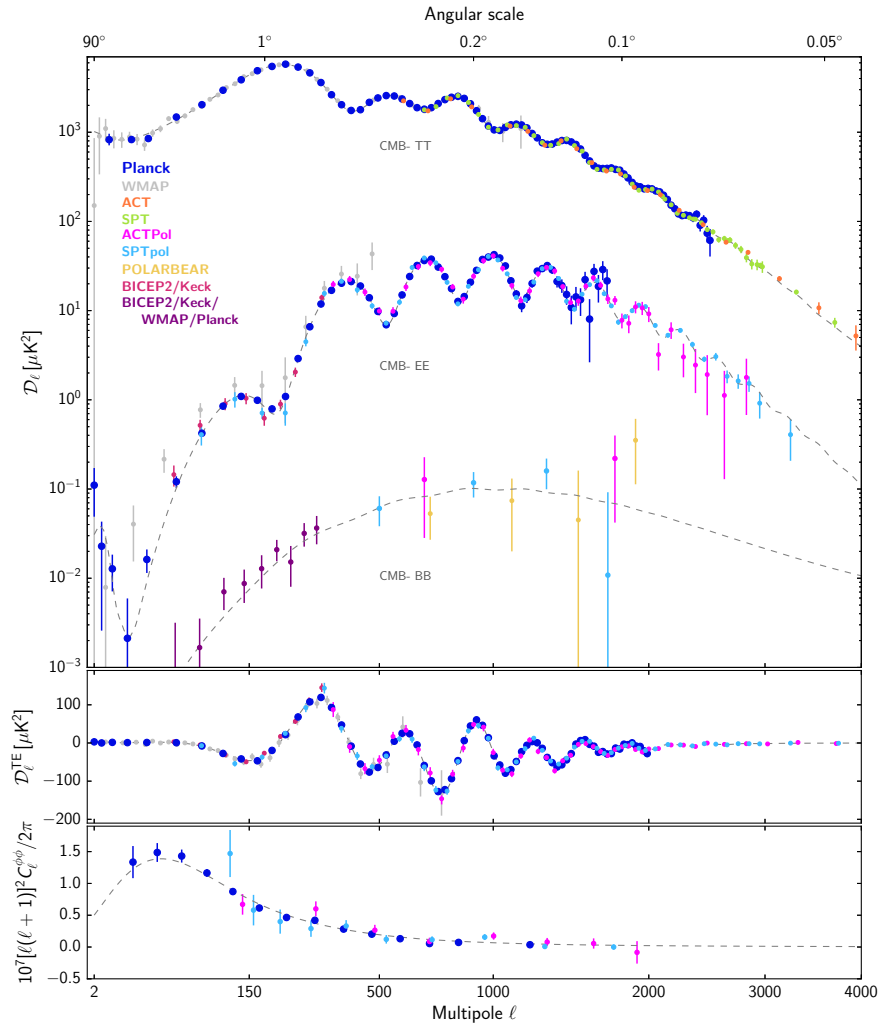


Figure 10.4 Compilation of CMB angular power spectra from satellite and ground-based observations (Planck Collaboration, 2020c, Fig. 18). (Upper) T -mode, E -mode, and B -mode power spectra. (Middle) Cross-power spectrum between T and E . (Lower) Lensing deflection power spectrum. The dashed lines show the best-fit Λ CDM model.

BAO, Figure 10.5. Comparison of the scales of acoustic oscillations in galaxies and the CMB allows the two scales to be matched, pinning the relative scales of galaxies today with those in the CMB at redshift $z \sim 1100$.

Major plans are underway to measure galaxy clustering as a function of redshift, with the primary aim to determine whether the evolution of dark energy is consistent with that of a cosmological constant. Such a

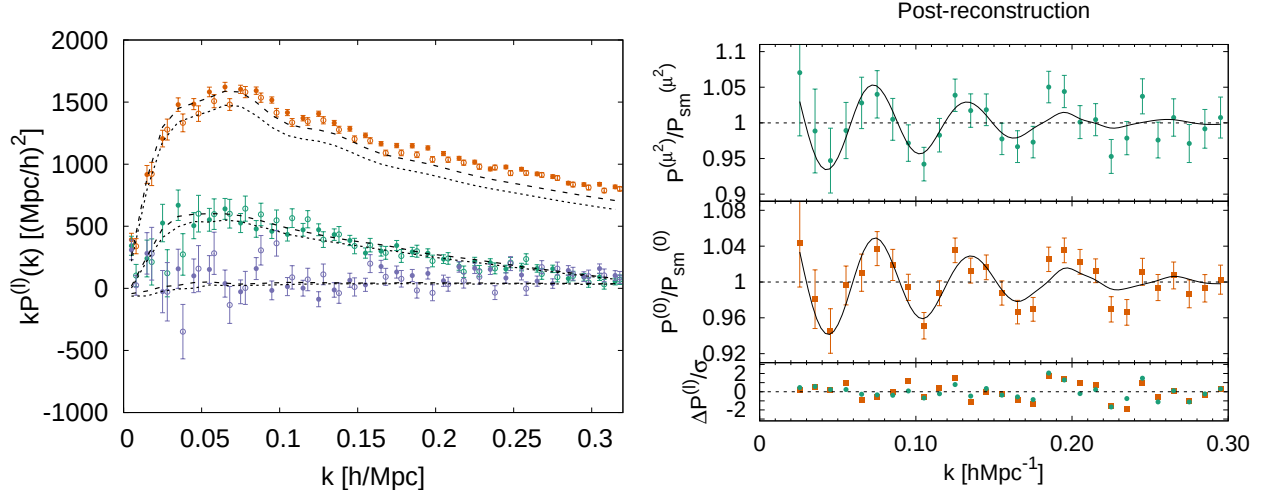


Figure 10.5 Monopole, dipole, and quadrupole power spectra of galaxies from the extended Baryon Oscillation Spectroscopic Survey (BOSS) of the Sloan Digital Sky Survey IV (SDSS-IV) (Gil-Marín et al., 2020). The analysis includes 377,458 luminous red galaxies covering approximately 18% of the sky over redshifts $z = 0.6$ – 1.0 . Filled and unfilled symbols are measurements from the north and south galactic caps respectively. The curves are flat Λ CDM model spectra calculated from simulations analyzed using the same footprint as the survey. The right panel shows spectra with a smooth component divided out to bring out the baryon acoustic oscillations (BAO).

measurement cannot be done with CMB observations, since the CMB offers only a snapshot of the Universe at high redshift.

10.1.5 Other supporting evidence

- The observed abundances of light elements H, D, ^3He , He, and Li are consistent with the predictions of big bang nucleosynthesis (BBN) provided that the baryonic density is $\Omega_b \approx 0.04$, in good agreement with measurements from the CMB.
- The ages of the oldest stars, in globular clusters, agree with the age of the Universe with dark energy, but are older than the Universe without dark energy.
- The existence of dark matter, possibly non-baryonic, is supported by ubiquitous evidence for unseen dark matter, deduced from sizes and velocities (or in the case of gravitational lensing, the gravitational potential) of various objects:
 - The Local Group of galaxies;
 - Rotation curves of spiral galaxies;
 - The temperature and distribution of x-ray gas in elliptical galaxies, and in clusters of galaxies;
 - Gravitational lensing by clusters of galaxies.
- The abundance of galaxy clusters as a function of redshift is consistent with a matter density $\Omega_m \approx 0.3$, but

not much higher. A low matter density slows the gravitational clustering of galaxies, implying relatively more and richer clusters at high redshift than at the present, as observed.

- The Bullet cluster is a rare example that supports the notion that the dark matter is non-baryonic. In the Bullet cluster, two clusters recently passed through each other. The baryonic matter, as measured from x-ray emission of hot gas, appears displaced from the dark matter, as measured from weak gravitational lensing.

10.2 Cosmological Principle

The **cosmological principle** (Milne, 1935) states that the Universe at large is

- **homogeneous** (has spatial translation symmetry),
- **isotropic** (has spatial rotation symmetry).

The primary evidence for this is the uniformity of the temperature of the CMB, which, after subtraction of the dipole produced by the motion of the solar system through the CMB, is constant over the sky to a few parts in 10^5 . Confirming evidence is the statistical uniformity of the distribution of galaxies over large scales.

The cosmological principle allows that the Universe evolves in time, as observations surely indicate — the Universe is expanding, galaxies, quasars, and galaxy clusters evolve with redshift, and the temperature of the CMB has undoubtedly decreased since recombination.

10.3 Friedmann-Lemaître-Robertson-Walker metric

Universes satisfying the cosmological principle are described by the Friedmann-Lemaître-Robertson-Walker (FLRW) metric, equation (10.28) below, discovered independently by Friedmann (1922; 1924) and Lemaître (1927) (English translation in Lemaître 1931). The FLRW metric was shown to be the unique metric for a homogeneous, isotropic universe by Robertson (1935; 1936; 1936) and Walker (1937). The metric, and the associated Einstein equations, which are known as the Friedmann equations, are set forward in the next several sections, §§10.4–10.9.

10.4 Spatial part of the FLRW metric: informal approach

The cosmological principle implies that

$$\boxed{\text{the spatial part of the FLRW metric is a 3D hypersphere}}. \quad (10.3)$$

In this context the term hypersphere is to be construed as including not only cases of positive curvature, which have finite positive radius of curvature, but also cases of zero and negative curvature, which have infinite and imaginary radius of curvature.

Figure 10.6 shows an embedding diagram of a 3D hypersphere in 4D Euclidean space. The horizontal

directions in the diagram represent the normal 3 spatial x, y, z dimensions, with one dimension z suppressed, while the vertical dimension represents the 4th spatial dimension w . The 3D hypersphere is a set of points $\{x, y, z, w\}$ satisfying

$$(x^2 + y^2 + z^2 + w^2)^{1/2} = R = \text{constant} . \quad (10.4)$$

An observer is sitting at the north pole of the diagram, at $\{0, 0, 0, 1\}$. A 2D sphere (which forms a 1D circle in the embedding diagram of Figure 10.6) at fixed distance surrounding the observer has **geodesic distance** $r_{\parallel}$ defined by

$$r_{\parallel} \equiv \text{proper distance to sphere measured along a radial geodesic} , \quad (10.5)$$

and **circumferential radius** r defined by

$$r \equiv (x^2 + y^2 + z^2)^{1/2} , \quad (10.6)$$

which has the property that the proper circumference of the sphere is $2\pi r$. In terms of $r_{\parallel}$ and r , the spatial metric is

$$dl^2 = dr_{\parallel}^2 + r^2 d\sigma^2 , \quad (10.7)$$

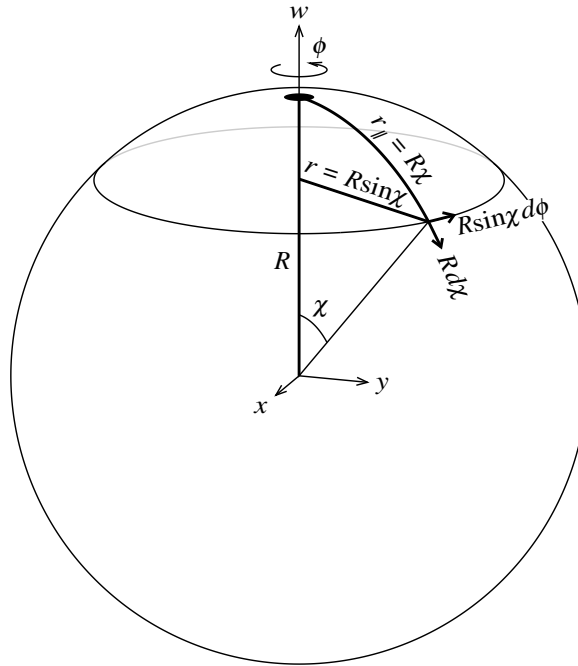


Figure 10.6 Embedding diagram of the FLRW geometry.

where $d\sigma^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$ is the metric of a unit 2-sphere.

Introduce the angle χ illustrated in the diagram. Evidently

$$\begin{aligned} r_{\parallel} &= R\chi , \\ r &= R\sin\chi . \end{aligned} \tag{10.8}$$

In terms of the angle χ , the spatial metric is

$$dl^2 = R^2 (d\chi^2 + \sin^2\chi d\sigma^2) , \tag{10.9}$$

which is one version of the spatial FLRW metric. The metric resembles the metric of a 2-sphere of radius R , which is not surprising since the same construction, with Figure 10.6 interpreted as the embedding diagram of a 2D sphere in 3D, yields the metric of a 2-sphere. Indeed, the construction iterates to give the metric of an N -dimensional sphere of arbitrarily many dimensions N .

Instead of the angle χ , the metric can be expressed in terms of the circumferential radius r . It follows from equations (10.8) that

$$r_{\parallel} = R \sin(r/R) , \tag{10.10}$$

whence

$$\begin{aligned} dr_{\parallel} &= \frac{dr}{\sqrt{1 - r^2/R^2}} \\ &= \frac{dr}{\sqrt{1 - Kr^2}} , \end{aligned} \tag{10.11}$$

where K is the curvature

$$K \equiv \frac{1}{R^2} . \tag{10.12}$$

In terms of r , the spatial FLRW metric is then

$$\boxed{dl^2 = \frac{dr^2}{1 - Kr^2} + r^2 d\sigma^2} . \tag{10.13}$$

The embedding diagram Figure 10.6 is a nice prop for the imagination, but it is not the whole story. The curvature K in the metric (10.13) may be not only positive, corresponding to real finite radius R , but also zero or negative, corresponding to infinite or imaginary radius R . The possibilities are called closed, flat, and open:

$$K \begin{cases} > 0 & \text{closed} & R \text{ real} , \\ = 0 & \text{flat} & R \rightarrow \infty , \\ < 0 & \text{open} & R \text{ imaginary} . \end{cases} \tag{10.14}$$

10.5 Comoving coordinates

The metric (10.13) is valid at any single instant of cosmic time t . As the Universe expands, the 3D spatial hypersphere (whether closed, flat, or open) expands. In cosmology it is highly advantageous to work in **comoving coordinates** that expand with the Universe. Why? Firstly, it is helpful conceptually and mathematically to think of the Universe as at rest in comoving coordinates. Secondly, linear perturbations, such as those in the CMB, have wavelengths that expand with the Universe, and are therefore fixed in comoving coordinates.

In practice, cosmologists introduce the **cosmic scale factor** $a(t)$

$$a(t) \equiv \text{measure of the size of the Universe, expanding with the Universe} , \quad (10.15)$$

which is proportional to but not necessarily equal to the radius R of the Universe. The cosmic scale factor a can be normalized in any arbitrary way. The most common convention adopted by cosmologists is to normalize it to unity at the present time,

$$a_0 = 1 , \quad (10.16)$$

where the 0 subscript conventionally signifies the present time.

Comoving geodesic and circumferential radial distances $x_{\parallel}$ and x are defined in terms of the proper geodesic and circumferential radial distances $r_{\parallel}$ and r by

$$ax_{\parallel} \equiv r_{\parallel} , \quad ax \equiv r . \quad (10.17)$$

Objects expanding with the Universe remain at fixed comoving positions $x_{\parallel}$ and x . In terms of the comoving circumferential radius x , the spatial FLRW metric is

$$dl^2 = a^2 \left(\frac{dx^2}{1 - \kappa x^2} + x^2 d\phi^2 \right) , \quad (10.18)$$

where the curvature constant κ , a constant in time and space, is related to the curvature K , equation (10.12), by

$$\kappa \equiv a^2 K . \quad (10.19)$$

Alternatively, in terms of the geodesic comoving radius $x_{\parallel}$, the spatial FLRW metric is

$$dl^2 = a^2 \left(dx_{\parallel}^2 + x^2 d\phi^2 \right) , \quad (10.20)$$

where

$$x = \begin{cases} \frac{\sin(\kappa^{1/2} x_{\parallel})}{\kappa^{1/2}} & \kappa > 0 \quad \text{closed} , \\ x_{\parallel} & \kappa = 0 \quad \text{flat} , \\ \frac{\sinh(|\kappa|^{1/2} x_{\parallel})}{|\kappa|^{1/2}} & \kappa < 0 \quad \text{open} . \end{cases} \quad (10.21)$$

Actually it is fine to use just the top expression of equations (10.21), which is mathematically equivalent to the bottom two expressions when $\kappa = 0$ or $\kappa < 0$ (because $\sin(ix)/i = \sinh(x)$).

For some purposes it is convenient to normalize the cosmic scale factor a so that $\kappa = 1, 0$, or -1 . In this case the spatial FLRW metric may be written

$$dl^2 = a^2 (d\chi^2 + x^2 d\phi^2) , \quad (10.22)$$

where

$$x = \begin{cases} \sin(\chi) & \kappa = 1 \quad \text{closed} , \\ \chi & \kappa = 0 \quad \text{flat} , \\ \sinh(\chi) & \kappa = -1 \quad \text{open} . \end{cases} \quad (10.23)$$

10.6 Spatial part of the FLRW metric: more formal approach

A more formal approach to the derivation of the spatial FLRW metric from the cosmological principle starts with the proposition that the spatial components $G_{\alpha\beta}$ of the Einstein tensor at fixed scale factor a (all time derivatives of a set to zero) should be proportional to the metric tensor

$$G_{\alpha\beta} = -K g_{\alpha\beta} \quad (\alpha, \beta = 1, 2, 3) . \quad (10.24)$$

Without loss of generality, the spatial metric can be taken to be of the form

$$dl^2 = f(r) dr^2 + r^2 d\phi^2 . \quad (10.25)$$

Imposing the condition (10.24) on the metric (10.25) recovers the spatial FLRW metric (10.13).

Exercise 10.1 Isotropic (Poincaré) form of the FLRW metric. By a suitable transformation of the comoving radial coordinate x , bring the spatial FLRW metric (10.18) to the “isotropic” form

$$dl^2 = \frac{4a^2}{(1 + \kappa X^2)^2} (dX^2 + X^2 d\phi^2) . \quad (10.26)$$

What is the relation between X and x ?

For an open geometry, $\kappa < 0$, the isotropic line-element (10.26) is also called the Poincaré ball, or in 2D the Poincaré disk, Figure 10.7. By construction, the isotropic line-element (10.26) is conformally flat, meaning that it equals the Euclidean line-element multiplied by a position-dependent conformal factor. Conformal transformations of a line-element preserve angles.

Solution.

$$X = \frac{x}{1 + \sqrt{1 - \kappa x^2}} = \frac{1}{\sqrt{\kappa}} \tan \left(\frac{\sqrt{\kappa} x}{2} \right) , \quad x = \frac{2X}{1 + \kappa X^2} = \frac{1}{\sqrt{\kappa}} \sin(\sqrt{\kappa} x) . \quad (10.27)$$

For an open geometry with $\kappa = -1$, X goes from 0 to 1 as x goes from 0 to ∞ .

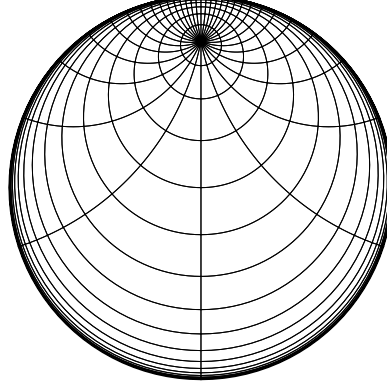


Figure 10.7 The Poincaré disk depicts the geometry of an open FLRW universe in isotropic coordinates. The lines are lines of latitude and longitude relative to a “pole” chosen here to be displaced from the centre of the disk. In isotropic coordinates, geodesics correspond to circles that intersect the boundary of the disk at right angles, such as the lines of constant longitude in this diagram. The lines of latitude remain unchanged under rotations about the pole.

10.7 FLRW metric

The full Friedmann-Lemaître-Robertson-Walker spacetime metric is

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dx^2}{1 - \kappa x^2} + x^2 d\phi^2 \right), \quad (10.28)$$

where t is **cosmic time**, which is the proper time experienced by comoving observers, who remain at rest in comoving coordinates $dx = d\theta = d\phi = 0$. Any of the alternative versions of the comoving spatial FLRW metric, equations (10.18), (10.20), (10.22), or (10.26), may be used as the spatial part of the FLRW spacetime metric (10.28). Conformal time η , equation (10.82), may also be used in place of the comoving proper time t .

10.8 Einstein equations for FLRW metric

The Einstein equations for the FLRW metric (10.28) are

$$-G^t_t = 3 \left(\frac{\dot{a}^2}{a^2} + \frac{\kappa}{a^2} \right) = 8\pi G\rho, \quad (10.29a)$$

$$G^x_x = G^\theta_\theta = G^\phi_\phi = -\frac{2\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{\kappa}{a^2} = 8\pi Gp, \quad (10.29b)$$

where overdots represent differentiation with respect to cosmic time t , so that for example $\dot{a} \equiv da/dt$. Note the trick of one index up, one down, to remove, modulo signs, the distorting effect of the metric on the

Einstein tensor. The Einstein equations (10.29) rearrange to give **Friedmann's equations**

$$\boxed{\frac{\dot{a}^2}{a^2} = \frac{8\pi G\rho}{3} - \frac{\kappa}{a^2}}, \quad (10.30a)$$

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p)}. \quad (10.30b)$$

Friedmann's two equations (10.30) are fundamental to cosmology. The first one relates the curvature κ of the Universe to the expansion rate $\dot{a}/a$ and the density ρ . The second one relates the acceleration $\ddot{a}/a$ to the density ρ plus 3 times the pressure p .

Exercise 10.2 FLRW equations in arbitrary dimensions. What are the FLRW equations in an arbitrary number N of spacetime dimensions? How does density ρ behave with cosmic scale factor a for a constant equation of state $w = p/\rho$?

Solution.

$$\frac{\dot{a}^2}{a^2} + \frac{\kappa}{a^2} = \frac{2\kappa_N\rho}{(N-1)(N-2)}, \quad (10.31a)$$

$$\frac{\ddot{a}}{a} = -\frac{\kappa_N}{(N-1)(N-2)} [(N-3)\rho + (N-1)p], \quad (10.31b)$$

where κ_N is the generalization of the gravitational coupling parameter $\kappa_4 \equiv 8\pi G$ to N spacetime dimensions (see Exercise 16.14). If $w = p/\rho$ is constant, then $\rho \propto a^{-(N-1)(1+w)}$.

10.9 Newtonian “derivation” of Friedmann equations

The Friedmann equations can be reproduced with a heuristic Newtonian argument.

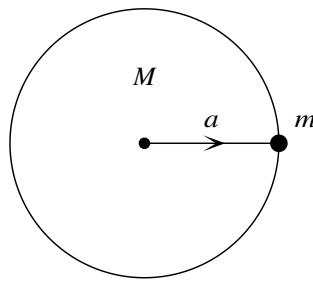


Figure 10.8 Newtonian picture in which the Universe is modeled as a uniform density sphere of radius a and mass M that gravitationally attracts a test mass m .

10.9.1 Energy equation

Model a piece of the Universe as a ball of radius a with uniform density ρ , hence of mass $M = \frac{4}{3}\pi\rho a^3$. Consider a small mass m attracted by this ball. Conservation of the kinetic plus potential energy of the small mass m implies

$$\frac{1}{2}m\dot{a}^2 - \frac{GMm}{a} = -\frac{\kappa mc^2}{2}, \quad (10.32)$$

where the quantity on the right is some constant whose value is not determined by this Newtonian treatment, but which GR implies is as given. The energy equation (10.32) rearranges to

$$\boxed{\frac{\dot{a}^2}{a^2} = \frac{8\pi G\rho}{3} - \frac{\kappa c^2}{a^2}}, \quad (10.33)$$

which reproduces the first Friedmann equation.

10.9.2 First law of thermodynamics

For adiabatic expansion, the first law of thermodynamics is

$$dE + p dV = 0. \quad (10.34)$$

With $E = \rho V$ and $V = \frac{4}{3}\pi a^3$, the first law (10.34) becomes

$$d(\rho a^3) + p da^3 = 0, \quad (10.35)$$

or, with the derivative taken with respect to cosmic time t ,

$$\dot{\rho} + 3(\rho + p)\frac{\dot{a}}{a} = 0. \quad (10.36)$$

Differentiating the first Friedmann equation in the form

$$\dot{a}^2 = \frac{8\pi G\rho a^2}{3} - \kappa c^2 \quad (10.37)$$

gives

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3}(\dot{\rho}a^2 + 2\rho a\dot{a}), \quad (10.38)$$

and substituting $\dot{\rho}$ from the first law (10.36) reduces this to

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3}a\dot{a}(-\rho - 3p). \quad (10.39)$$

Hence

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p)}, \quad (10.40)$$

which reproduces the second Friedmann equation.

10.9.3 Comment on the Newtonian derivation

The above Newtonian derivation of Friedmann's equations is only heuristic. A different result could have been obtained if different assumptions had been made. If for example the Newtonian gravitational force law $m\ddot{a} = -GMm/a^2$ were taken as correct, then it would follow that $\ddot{a}/a = -\frac{4}{3}\pi G\rho$, which is missing the all-important $3p$ contribution (without which there would be no inflation or dark energy) to Friedmann's second equation.

It is notable that the first law of thermodynamics is built in to the Friedmann equations. This implies that entropy is conserved in FLRW Universes (but see Concept question 30.5). This remains true even when the mix of particles changes, as happens for example during the epoch of electron-positron annihilation, or during big bang nucleosynthesis. How then does entropy increase in the real Universe? Through fluctuations away from the perfect homogeneity and isotropy assumed by the FLRW metric.

10.10 Hubble parameter

The **Hubble parameter** $H(t)$ is defined by

$$H \equiv \frac{\dot{a}}{a} . \quad (10.41)$$

The Hubble parameter H varies in cosmic time t , but is constant in space at fixed cosmic time t .

The value of the Hubble parameter today is called the **Hubble constant** H_0 (the subscript 0 signifies the present time). The Hubble constant measured from Cepheid variable stars and Type Ia supernova is (Riess et al., 2011; Riess et al., 2018).

$$H_0 = 73.5 \pm 1.7 \text{ km s}^{-1} \text{ Mpc}^{-1} . \quad (10.42)$$

The observed CMB power spectrum, Figure 10.3, provides an accurate measurement of the angular location of the first peak in the power spectrum, which determines the angular size of the sound horizon at recombination, Chapter 32. This cosmological yardstick translates into a measurement of the Hubble parameter H_0 , but only if a cosmological model is assumed. In particular, the angular location of the peak depends on the spatial curvature. The combination of CMB data with other data, notably Baryon Acoustic Oscillations in galaxy clustering, Figure 10.5, and the Hubble diagram of Type Ia supernovae, Figure 10.1, point consistently to a spatially flat cosmological model. If the Universe is taken to be spatially flat, then CMB data from the Planck satellite yield (Planck Collaboration, 2018)

$$H_0 = 67.4 \pm 0.5 \text{ km s}^{-1} \text{ Mpc}^{-1} . \quad (10.43)$$

The Cepheid and CMB measurements (10.42) and (10.43) of H_0 lie outside each other's error bars. One can either be impressed that two completely independent measurements of H_0 yield almost the same result, or be worried by the disagreement. I incline to the former view, since these kind of measurements tend to be beset with systematic uncertainties that can be difficult to get under control.

The distance d to an object that is receding with the expansion of the universe is proportional to the cosmic

scale factor, $d \propto a$, and its recession velocity v is consequently proportional to $\dot{a}$. The result is **Hubble's law** relating the recession velocity v and distance d of distant objects

$$\boxed{v = H_0 d} . \quad (10.44)$$

Since it takes light time to travel from a distant object, and the Hubble parameter varies in time, the linear relation (10.44) breaks down at cosmological distances.

We, in the Milky Way, reside in an overdense region of the Universe that has collapsed out of the general Hubble expansion of the Universe. The local overdense region of the Universe that has just turned around from the general expansion and is beginning to collapse for the first time is called the **Local Group** of galaxies. The Local Group consists of order 100 galaxies, mostly dwarf and irregular galaxies. It contains two major spiral galaxies, Andromeda (M31) and the Milky Way, and one mid-sized spiral galaxy Triangulum (M33). The Local Group is about 1 Mpc in radius.

Because of the ubiquity of the Hubble constant in cosmological studies, cosmologists often parameterize it by the quantity h defined by

$$h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}} . \quad (10.45)$$

The reciprocal of the Hubble constant gives an approximate estimate of the age of the Universe (c.f. Exercise 10.7),

$$\frac{1}{H_0} = 9.778 h^{-1} \text{ Gyr} = 14.0 h_{0.70}^{-1} \text{ Gyr} . \quad (10.46)$$

10.11 Critical density

The critical density ρ_{crit} is defined to be the density required for the Universe to be flat, $\kappa = 0$. According to the first of Friedmann equations (10.30), this sets

$$\boxed{\rho_{\text{crit}} \equiv \frac{3H^2}{8\pi G}} . \quad (10.47)$$

The critical density ρ_{crit} , like the Hubble parameter H , evolves with time.

10.12 Omega

Cosmologists designate the ratio of the actual density ρ of the Universe to the critical density ρ_{crit} by the fateful letter Ω , the final letter of the Greek alphabet,

$$\boxed{\Omega \equiv \frac{\rho}{\rho_{\text{crit}}}} . \quad (10.48)$$

Table 10.1: Cosmic inventory

Species		WMAP	Planck
		Hinshaw et al. (2012)	Planck Collaboration (2018)
Dark energy (Λ)	Ω_Λ	0.72 ± 0.01	0.685 ± 0.007
Non-baryonic cold dark matter (CDM)	Ω_c	0.24 ± 0.01	0.261 ± 0.002
Baryonic matter	Ω_b	0.047 ± 0.002	0.0490 ± 0.0005
Neutrinos	Ω_ν	< 0.02	< 0.004
Photons (CMB)	Ω_γ	5×10^{-5}	5×10^{-5}
Total	Ω	1.003 ± 0.004	0.999 ± 0.002
Curvature	Ω_k	-0.003 ± 0.004	0.001 ± 0.002

With no subscript, Ω denotes the total mass-energy density in all forms. A subscript x on Ω_x denotes mass-energy density of type x .

The curvature density ρ_k , which is not really a form of mass-energy but it is sometimes convenient to treat it as though it were, is defined by

$$\rho_k \equiv -\frac{3\kappa c^2}{8\pi G a^2}, \quad (10.49)$$

and correspondingly

$$\Omega_k \equiv \frac{\rho_k}{\rho_{\text{crit}}} = -\frac{\kappa c^2}{a^2 H^2}. \quad (10.50)$$

If the cosmic scale factor is normalized to unity at the present time, equation (10.16), then the relation between Ω_k and the curvature constant κ is $\Omega_k = -\kappa c^2/H_0^2$. According to the first of Friedmann's equations (10.30), the curvature density Ω_k satisfies

$$\boxed{\Omega_k = 1 - \Omega}. \quad (10.51)$$

Note that Ω_k has opposite sign from κ , so a closed universe has negative Ω_k .

Table 10.1 gives measurements of Ω in various species, as reported by Hinshaw et al. (2012) from the final analysis of the CMB power spectrum from WMAP, and by Planck Collaboration (2018) from the final analysis of the CMB power spectrum from Planck. Both sets of analyses incorporate measurements from a variety of other data, including CMB data at smaller scales, Figure 10.3, supernova data, Figure 10.1, galaxy clustering (Baryonic Acoustic Oscillation, or BAO) data, Figure 10.5, and local measurements of the Hubble constant H_0 (Riess et al., 2011; Riess et al., 2018). It is largely the CMB data that enable cosmological parameters to be measured to the level of precision given in the Table. However, the CMB data by themselves constrain tightly only a combination of the Hubble parameter H_0 and the curvature Ω_k , as illustrated in Figure 26 of Planck Collaboration (2018). Other data, in particular BAO and the supernova Hubble diagram, resolve this uncertainty, pointing to a flat Universe, $\Omega_k = 0$. Importantly, the various data

are consistent with each other, inspiring confidence in the correctness of the Standard Model. The neutrino limit implies an upper limit to the sum of the masses of all neutrino species (Planck Collaboration, 2018),

$$\sum_{\nu} m_{\nu} < 0.12 \text{ eV} . \quad (10.52)$$

Exercise 10.3 Omega in photons. Most of the energy density in electromagnetic radiation today is in CMB photons. Calculate Ω_{γ} in CMB photons. Note that photons may not be the only relativistic species today. Neutrinos with masses smaller than about 10^{-4} eV would be still be relativistic at the present time, Exercise 10.21.

Solution. CMB photons have a blackbody spectrum at temperature $T_0 = 2.725$ K, so their density can be calculated from the blackbody formula. The present day ratio Ω_{γ} of the mass-energy density ρ_{γ} of CMB photons to the critical density ρ_{crit} is

$$\Omega_{\gamma} \equiv \frac{\rho_{\gamma}}{\rho_{\text{crit}}} = \frac{8\pi G \rho_{\gamma}}{3H_0^2} = \frac{8\pi^3 G (kT_0)^4}{45H_0^2 c^5 \hbar^3} = 2.471 \times 10^{-5} h^{-2} T_{2.725 \text{ K}}^4 = 5.0 \times 10^{-5} h_{0.70}^{-2} T_{2.725 \text{ K}}^4 . \quad (10.53)$$

10.13 Types of mass-energy

The energy-momentum tensor $T_{\mu\nu}$ of an FLRW Universe is necessarily homogeneous and isotropic, by assumption of the cosmological principle, taking the form (note yet again the trick of one index up and one down to remove the distorting effect of the metric)

$$T_{\nu}^{\mu} = \begin{pmatrix} T_t^t & 0 & 0 & 0 \\ 0 & T_r^r & 0 & 0 \\ 0 & 0 & T_{\theta}^{\theta} & 0 \\ 0 & 0 & 0 & T_{\phi}^{\phi} \end{pmatrix} = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} . \quad (10.54)$$

Table 10.2 gives equations of state p/ρ for generic species of mass-energy, along with $(\rho + 3p)/\rho$, which

Table 10.2: Properties of universes dominated by various species

Species	p/ρ	$(\rho + 3p)/\rho$	$\rho \propto$
Radiation	1/3	2	a^{-4}
Matter	0	1	a^{-3}
Curvature	“−1/3”	“0”	a^{-2}
Vacuum	−1	−2	a^0

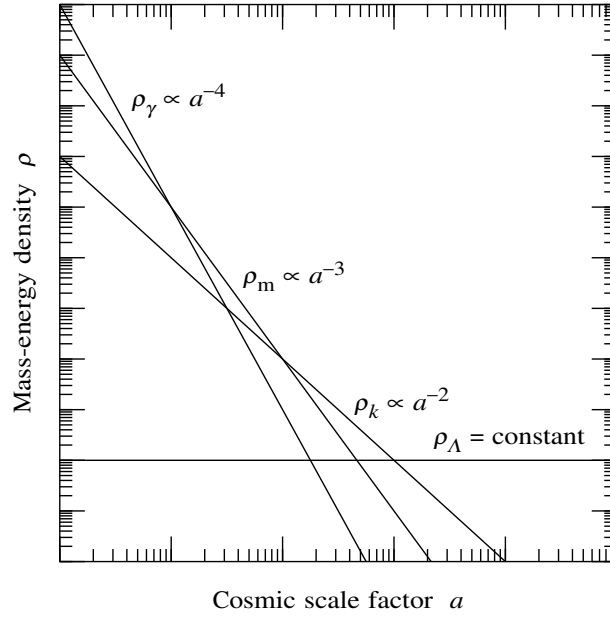


Figure 10.9 Behaviour of the mass-energy density ρ of various species as a function of cosmic time t .

determines the gravitational attraction (deceleration) per unit energy, and how the mass-energy varies with cosmic scale factor, $\rho \propto a^n$, Exercise 10.4.

As commented in §10.9.2, the first law of thermodynamics for adiabatic expansion is built into Friedmann's equations. In fact the law represents covariant conservation of energy-momentum for the system as a whole

$$D_\mu T^{\mu\nu} = 0 . \quad (10.55)$$

As long as species do not convert into each other (for example, no annihilation), covariant energy-momentum conservation holds individually for each species, so the first law applies to each species individually, determining how its energy density ρ varies with cosmic scale factor a . Figure 10.9 illustrates how the energy densities ρ of various species evolve as a function of scale factor a .

Vacuum energy is equivalent to a **cosmological constant**. Einstein originally introduced the cosmological constant Λ as a modification to the left hand side of his equations,

$$G_{\kappa\mu} + \Lambda g_{\kappa\mu} = 8\pi G T_{\kappa\mu} . \quad (10.56)$$

The cosmological constant term can be taken over to the right hand side and reinterpreted as vacuum energy $T_{\kappa\mu} = -\rho_\Lambda g_{\kappa\mu}$ with energy density ρ_Λ , satisfying

$$\Lambda = 8\pi G \rho_\Lambda . \quad (10.57)$$

Exercise 10.4 Mass-energy in an FLRW Universe.

1. **First law.** The first law of thermodynamics for adiabatic expansion is built into Friedmann's equations (= Einstein's equations for the FLRW metric):

$$d(\rho a^3) + p da^3 = 0 . \quad (10.58)$$

How does the density ρ evolve with cosmic scale factor for a species with equation of state $p/\rho = w$ with constant w ? You should get an answer of the form

$$\rho \propto a^n . \quad (10.59)$$

2. **Attractive or repulsive?** For what equation of state w is the mass-energy attractive or repulsive? Consider in particular the cases of “matter,” “radiation,” “curvature,” and “vacuum” energy.

Concept question 10.5 Mass of a ball of photons or of vacuum. What is the gravitational mass of a homogeneous, isotropic, spherical, ball of photons embedded in empty space, as measured by an observer outside the ball? Assume that the boundary of the ball is free to expand or contract. What if the ball of photons is bounded by a stationary reflecting spherical wall? What if the ball is a ball of vacuum energy instead of photons? **Answer.** The right way to address this question is to think about what happens at the boundary between the ball and empty space. See §20.17.

10.14 Redshifting

The spatial translation symmetry of the FLRW metric implies conservation of generalized momentum. As you will show in Exercise 10.6, a particle that moves along a geodesic in the radial direction, so that $d\theta = d\phi = 0$, has 4-velocity p^ν satisfying

$$p_{x_{\parallel}} = \text{constant} . \quad (10.60)$$

This conservation law implies that the coordinate momentum $p^{x_{\parallel}}$ of a radially moving particle decays as

$$p^{x_{\parallel}} = g^{x_{\parallel} x_{\parallel}} p_{x_{\parallel}} = \frac{p_{x_{\parallel}}}{a^2} \propto \frac{1}{a^2} , \quad (10.61)$$

so the proper momentum (the momentum measured in a comoving tetrad frame) decays as

$$p_{\text{proper}}^{x_{\parallel}} \equiv m \frac{dr_{\parallel}}{d\tau} = ma \frac{dx_{\parallel}}{d\tau} = ap^{x_{\parallel}} \propto \frac{1}{a} , \quad (10.62)$$

which is true for both massive and massless particles.

It follows from equation (10.62) that light observed on Earth from a distant object will be redshifted by a factor

$$1 + z = \frac{a_0}{a} , \quad (10.63)$$

where a_0 is the present day cosmic scale factor. Cosmologists often refer to the redshift of an epoch, since the cosmological redshift is an observationally accessible quantity that uniquely determines the cosmic time of emission.

Exercise 10.6 Geodesics in the FLRW geometry. The Friedmann-Lemaître-Robertson-Walker metric of cosmology is

$$ds^2 = -dt^2 + a(t)^2 \left[dx_{\parallel}^2 + \frac{\sin^2(\kappa^{1/2} x_{\parallel})}{\kappa} (d\theta^2 + \sin^2\theta d\phi^2) \right], \quad (10.64)$$

where κ is a constant, the curvature constant. Note that equation (10.64) is valid for all values of κ , including zero and negative values: there is no need to consider the cases separately.

1. **Conservation of generalized momentum.** Consider a particle moving with comoving 4-momentum $p^{\mu} \equiv dx^{\mu}/d\lambda$ along a geodesic in the radial direction, so that $d\theta = d\phi = 0$. Argue that the Lagrangian equations of motion

$$\frac{d}{d\lambda} \frac{\partial L}{\partial p^{x_{\parallel}}} = \frac{\partial L}{\partial x_{\parallel}} \quad (10.65)$$

with effective Lagrangian

$$L = \frac{1}{2} g_{\mu\nu} p^{\mu} p^{\nu} \quad (10.66)$$

imply that

$$p_{x_{\parallel}} = \text{constant} . \quad (10.67)$$

Argue further from the same Lagrangian equations of motion that the assumption of a radial geodesic is valid because

$$p_{\theta} = p_{\phi} = 0 \quad (10.68)$$

is a consistent solution. [Hint: The metric $g_{\mu\nu}$ depends on the coordinate $x_{\parallel}$. But for radial geodesics with $p^{\theta} = p^{\phi} = 0$, the possible contributions from derivatives of the metric vanish.]

2. **Proper momentum.** Argue that a proper interval of distance measured by comoving observers along the radial geodesic is $a dx_{\parallel}$. Hence show from equation (10.69) that the proper momentum $p^{x_{\parallel}}$ of the particle relative to comoving observers (who are at rest in the FLRW metric) evolves as

$$p^{x_{\parallel}} \equiv ma \frac{dx_{\parallel}}{d\lambda} \propto \frac{1}{a} . \quad (10.69)$$

3. **Redshift.** What relation does your result (10.69) imply between the redshift $1+z$ of a distant object observed on Earth and the expansion factor a since the object emitted its light? [Hint: Equation (10.69) is valid for massless as well as massive particles. Why?]
4. **Temperature of the CMB.** Argue from the above results that the temperature T of the CMB evolves with cosmic scale factor as

$$T \propto \frac{1}{a} . \quad (10.70)$$

10.15 Evolution of the cosmic scale factor

Given how the energy density ρ of each species evolves with cosmic scale factor a , the first Friedmann equation then determines how the cosmic scale factor $a(t)$ itself evolves with cosmic time t . If the Hubble parameter $H \equiv \dot{a}/a$ is expressed as a function of cosmic scale factor a , then cosmic time t can be expressed in terms of a as

$$t = \int \frac{da}{aH} . \quad (10.71)$$

The definition (10.47) of the critical density allows the Hubble parameter H to be written

$$\frac{H}{H_0} = \sqrt{\frac{\rho_{\text{crit}}}{\rho_{\text{crit}}(a_0)}} . \quad (10.72)$$

The critical density ρ_{crit} is itself the sum of the densities ρ of all species *including* the curvature density,

$$\rho_{\text{crit}} = \rho_k + \sum_{\text{species } x} \rho_x . \quad (10.73)$$

For example, in the case that the density is comprised of radiation, matter, and vacuum, the critical density is

$$\rho_{\text{crit}} = \rho_r + \rho_m + \rho_k + \rho_\Lambda , \quad (10.74)$$

and equation (10.72) is

$$\frac{H(t)}{H_0} = \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda} , \quad (10.75)$$

where Ω_x represents its value at the present time. For density comprised of radiation, matter, and vacuum, equation (10.75), the time t , equation (10.71), is

$$t = \frac{1}{H_0} \int \frac{da}{a \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda}} , \quad (10.76)$$

which is an elliptic integral of the third kind. The elliptic integral simplifies to elementary functions in some cases relevant to reality, Exercises 10.7 and 10.8.

If one single species in particular dominates the mass-energy density, then equation (10.76) integrates to give the results in Table 10.3.

10.16 Age of the Universe

The present age t_0 of the Universe since the Big Bang can be derived from equation (10.76) and cosmological parameters, Table 10.1. Planck Collaboration (2018) give the age of the Universe to be

$$t_0 = 13.80 \pm 0.02 \text{ Gyr} . \quad (10.77)$$

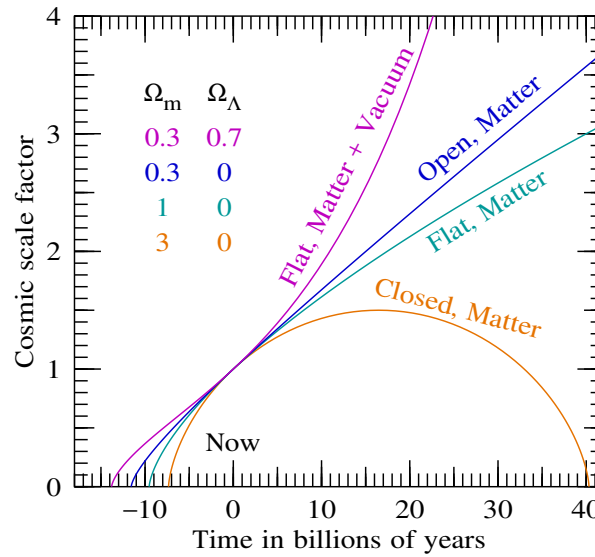
Figure 10.10 Cosmic scale factor as a function of time in universes with various Ω_m and Ω_Λ .

Table 10.3: Evolution of cosmic scale factor in universes dominated by various species

Dominant Species	$a \propto$
Radiation	$t^{1/2}$
Matter	$t^{2/3}$
Curvature	t
Vacuum	e^{Ht}

Exercise 10.7 Age of an FLRW universe containing matter and vacuum.

1. **Age of a universe dominated by matter and vacuum.** To a good approximation, the Universe today appears to be flat, and dominated by matter and a cosmological constant, with $\Omega_m + \Omega_\Lambda = 1$. Show that in this case the relation between age t and cosmic scale factor a is

$$t = \frac{2}{3H_0\sqrt{\Omega_\Lambda}} \operatorname{asinh} \sqrt{\frac{\Omega_\Lambda a^3}{\Omega_m}}. \quad (10.78)$$

2. **Age of our Universe.** Evaluate the age t_0 of the Universe today ($a_0 = 1$) in the approximation that the Universe is flat and dominated by matter and a cosmological constant. [Note: Astronomers define one Julian year to be exactly 365.25 days of $24 \times 60 \times 60 = 86,400$ seconds each. A parsec (pc) is the

distance at which a star has a parallax of 1 arcsecond, whence $1 \text{ pc} = (60 \times 60 \times 180/\pi) \text{ au}$, where 1 au is one Astronomical Unit, the Earth-Sun distance. One Astronomical Unit was officially defined by the International Astronomical Union (IAU) in 2012 to be $1 \text{ au} \equiv 149,597,870,700 \text{ m}$, with official abbreviation au.]

Exercise 10.8 Age of an FLRW universe containing radiation and matter. The Universe was dominated by radiation and matter over many decades of expansion including the time of recombination. Show that for a flat Universe containing radiation and matter the relation between age t and cosmic scale factor a is

$$t = \frac{2\Omega_r^{3/2}}{3H_0\Omega_m^2} \frac{\hat{a}^2(2 + \sqrt{1 + \hat{a}})}{(1 + \sqrt{1 + \hat{a}})^2} , \quad (10.79)$$

where $\hat{a}$ is the cosmic scale factor scaled to 1 at matter-radiation equality,

$$\hat{a} \equiv \frac{a}{a_{\text{eq}}} = \frac{\Omega_m a}{\Omega_r} . \quad (10.80)$$

You may well find a formula different from (10.79), but you should be able to recover the latter using the identity $\sqrt{1 + \hat{a}} - 1 = \hat{a}/(\sqrt{1 + \hat{a}} + 1)$. Equation (10.79) has the virtue that it is numerically stable to evaluate for all $\hat{a}$, including tiny $\hat{a}$.

10.17 Conformal time

It is often convenient to use **conformal time** η defined by (with units c temporarily restored)

$$a d\eta \equiv c dt , \quad (10.81)$$

with respect to which the FLRW metric is

$$\boxed{ds^2 = a(\eta)^2 \left(-d\eta^2 + dx_{\parallel}^2 + x^2 d\phi^2 \right)} , \quad (10.82)$$

with x given by equation (10.21). The term conformal refers to a metric that is multiplied by an overall factor, the conformal factor (squared). In the FLRW metric (10.82), the cosmic scale factor a is the conformal factor.

Conformal time η is constructed so that radial null geodesics move at unit velocity in conformal coordinates. Light moving radially, with $d\theta = d\phi = 0$, towards an observer at the origin $x_{\parallel} = 0$ satisfies

$$\frac{dx_{\parallel}}{d\eta} = -1 . \quad (10.83)$$

Exercise 10.9 Relation between conformal time and cosmic scale factor. What is the relation between conformal time η and cosmic scale factor a if the energy-momentum is dominated by a species with equation of state $p/\rho = w = \text{constant}$?

Solution. The conformal time η is related to cosmic scale factor a by (units $c = 1$)

$$\eta = \int \frac{da}{a^2 H} . \quad (10.84)$$

For $p/\rho = w = \text{constant}$, a possible choice of integration constant for η is: if $w > -1/3$ (decelerating), set $\eta = 0$ at $a = 0$, so that $\eta \rightarrow \infty$ at $a \rightarrow \infty$; if $w < -1/3$ (accelerating), set $\eta = 0$ at $a \rightarrow \infty$, so that $\eta \rightarrow -\infty$ at $a \rightarrow 0$. Then

$$\eta = \frac{2}{(1+3w)aH} \propto \pm a^{(1+3w)/2} , \quad (10.85)$$

in which the sign is positive for $w > -1/3$, negative for $w < -1/3$, ensuring that conformal time η always increases with cosmic scale factor a . For the special case of a curvature-dominated universe, $w = -1/3$,

$$\eta = \frac{\ln a}{aH} \propto \ln a , \quad (10.86)$$

which goes to $\eta \rightarrow -\infty$ as $a \rightarrow 0$ and $\eta \rightarrow \infty$ as $a \rightarrow \infty$.

10.18 Looking back along the lightcone

Since light moves radially at unit velocity in conformal coordinates, an object at geodesic distance $x_{\parallel}$ that emits light at conformal time η_{em} is observed at conformal time η_{obs} given by

$$\boxed{x_{\parallel} = \eta_{\text{obs}} - \eta_{\text{em}}} . \quad (10.87)$$

The comoving geodesic distance $x_{\parallel}$ to an object is

$$x_{\parallel} = \int_{\eta_{\text{em}}}^{\eta_{\text{obs}}} d\eta = \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{c dt}{a} = \int_{a_{\text{em}}}^{a_{\text{obs}}} \frac{c da}{a^2 H} = \int_0^z \frac{c dz}{H} , \quad (10.88)$$

where the last equation assumes the relation $1+z = 1/a$, valid as long as a is normalized to unity at the observer (us) at the present time $a_{\text{obs}} = a_0 = 1$. In the case that the density is comprised of (curvature and) radiation, matter, and vacuum, equation (10.88) gives

$$x_{\parallel} = \frac{c}{H_0} \int_{1/(1+z)}^1 \frac{da}{a^2 \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda}} , \quad (10.89)$$

which is an elliptic integral of the first kind. Given the geodesic comoving distance $x_{\parallel}$, the circumferential comoving distance x then follows as

$$x = \frac{\sinh(\sqrt{\Omega_k} H_0 x_{\parallel}/c)}{\sqrt{\Omega_k} H_0/c} . \quad (10.90)$$

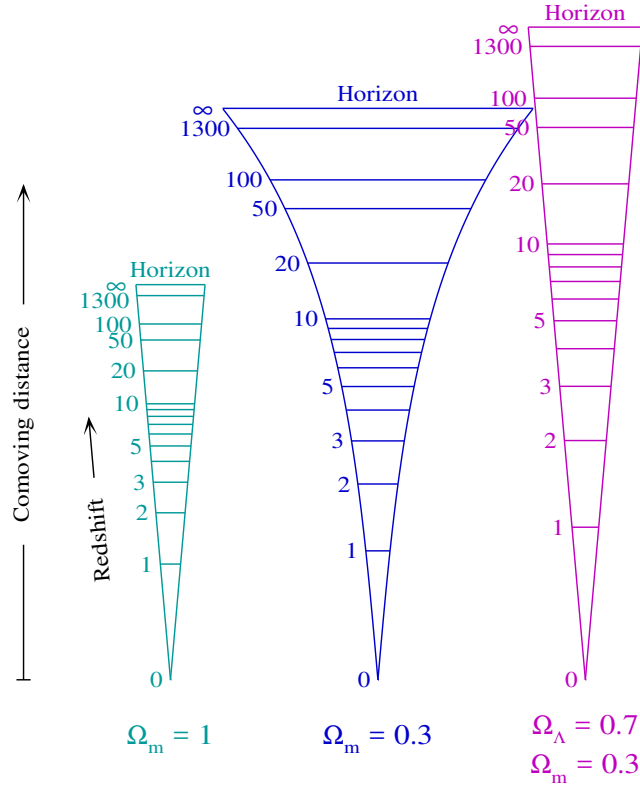


Figure 10.11 In this diagram, each wedge represents a cone of fixed opening angle, with the observer (us) at the point of the cone, at zero redshift. The wedges show the relation between physical sizes, namely the comoving distances $x_{\parallel}$ in the radial (vertical) and x in the transverse (horizontal) directions, and observable quantities, namely redshift and angular separation, in three different cosmological models: (left) a flat matter-dominated universe, (middle) an open matter-dominated universe, and (right) a flat Λ CDM universe.

To second order in redshift z ,

$$x \approx x_{\parallel} \approx \frac{c}{H_0} \left[z - z^2 \left(\Omega_r + \frac{3}{4}\Omega_m + \frac{1}{2}\Omega_k \right) + \dots \right] . \quad (10.91)$$

The geodesic and circumferential distances $x_{\parallel}$ and x differ at order z^3 .

Figure 10.11 illustrates the relation between the comoving geodesic and circumferential distances $x_{\parallel}$ and x , equations (10.89) and (10.90), and redshift z , equation (10.63), in three cosmological models, including the standard flat Λ CDM model.

10.19 Hubble diagram

The Hubble diagram of Type Ia supernova shown in Figure 10.1 is a plot of (log) luminosity distance $\log d_L$ versus (log) redshift $\log z$. The luminosity distance is explained in §10.19.1 immediately following.

10.19.1 Luminosity distance

Astronomers conventionally define the **luminosity distance** d_L to a celestial object so that the observed flux F from the object (energy observed per unit time per unit collecting area of the telescope) is equal to the intrinsic luminosity L of the object (energy per unit time emitted by the object in its rest frame) divided by $4\pi d_L^2$,

$$F = \frac{L}{4\pi d_L^2} . \quad (10.92)$$

In other words, the luminosity distance d_L is defined so that flux F and luminosity L are related by the usual inverse square law of distance. Objects at cosmological distances are redshifted, so the luminosity at some emitted wavelength λ_{em} is observed at the redshifted wavelength $\lambda_{\text{obs}} = (1+z)\lambda_{\text{em}}$. The luminosity distance (10.92) is defined so that the flux $F(\lambda_{\text{obs}})$ on the left hand side is at the observed wavelength, while the luminosity $L(\lambda_{\text{em}})$ on the right hand side is at the emitted wavelength. The observed flux and emitted luminosity are then related by

$$F = \frac{L}{(1+z)^2 4\pi x^2} , \quad (10.93)$$

where x is the comoving circumferential radius, normalized to $a_0 = 1$ at the present time. The factor of $1/(4\pi x^2)$ expresses the fact that the luminosity is spread over a sphere of proper area $4\pi x^2$. Equation (10.93) involves two factors of $1+z$, one of which come from the fact that the observed photon energy is redshifted, and the other from the fact that the observed number of photons detected per unit time is redshifted by $1+z$. Equations (10.92) and (10.93) imply that the luminosity distance d_L is related to the circumferential distance x and the redshift z by

$$d_L = (1+z)x . \quad (10.94)$$

Why bother with the luminosity distance if it can be reduced to the circumferential distance x by dividing by a redshift factor? The answer is that, especially historically, fluxes of distant astronomical objects are often measured from images without direct spectral information. If the intrinsic luminosity of the object is treated as “known” (as with Cepheid variables and Type Ia supernovae), then the luminosity distance $d_L = \sqrt{L/(4\pi F)}$ can be inferred without knowledge of the redshift. In practice objects are often measured with a fixed colour filter or set of filters, and some additional correction, historically called the *K*-correction, is necessary to transform the flux in an observed filter to a common band.

10.19.2 Magnitudes

The Hubble diagram of Type Ia supernova shown in Figure 10.1 has for its vertical axis the astronomers' system of magnitudes, a system that dates back to the 2nd century BC Greek astronomer Hipparchus.

A magnitude is a logarithmic measure of brightness, defined such that an interval of 5 magnitudes m corresponds to a factor of 100 in linear flux F . Following Hipparchus, the magnitude system is devised such that the brightest stars in the sky have apparent magnitudes of approximately 0, while fainter stars have larger magnitudes, the faintest naked eye stars in the sky being about magnitude 6. Traditionally, the system is tied to the star Vega, which is defined to have magnitude 0. Thus the apparent magnitude m of a star is

$$m = m_{\text{Vega}} - 2.5 \log(F/F_{\text{Vega}}) . \quad (10.95)$$

The absolute magnitude M of an object is defined to equal the apparent magnitude m that it would have if it were 10 parsecs away, which is the approximate distance to the star Vega. Thus

$$m - M = 5 \log(d_L/10 \text{ pc}) , \quad (10.96)$$

where d_L is the luminosity distance. The difference $m - M$ is called the **distance modulus**.

Exercise 10.10 Hubble diagram. Draw a theoretical Hubble diagram, a plot of luminosity distance d_L versus redshift z , for universes with various values of Ω_Λ and Ω_m . The relation between d_L and z is an elliptic integral of the first kind, so you will need to find a program that does elliptic integrals (alternatively, you can do the integral numerically). The elliptic integral simplifies to elementary functions in simple cases where the mass-energy density is dominated by a single component (either mass $\Omega_m = 1$, or curvature $\Omega_k = 1$, or a cosmological constant $\Omega_\Lambda = 1$).

Solution. Your model curves should look similar to those in Figure 10.1.

10.20 Recombination

The CMB comes to us from the epoch of **recombination**, when the Universe transitioned from being mostly ionized, and therefore opaque, to mostly neutral, and therefore transparent. As the Universe expands, the temperature of the cosmic background decreases as $T \propto a^{-1}$. Given that the CMB temperature today is $T_0 \approx 3 \text{ K}$, the temperature would have been about 3,000 K at a redshift of about 1,000. This temperature corresponds to the temperature at which hydrogen, the most abundant element in the Universe, ionizes. Not coincidentally, the temperature of recombination is comparable to the $\approx 5,800 \text{ K}$ surface temperature of the Sun. The CMB and Sun temperatures differ because the baryon-to-photon number density is much greater in the Sun.

The transition from mostly ionized to mostly neutral takes place over a fairly narrow range of redshifts, just as the transition from ionized to neutral at the photosphere of the Sun is rather sharp. Thus recombination can be approximated as occurring almost instantaneously. Planck Collaboration (2018) give the redshift of

last scattering, where the photon-electron scattering (Thomson) optical depth was 1,

$$z_* = 1089.8 \pm 0.2 . \quad (10.97)$$

Hinshaw et al. (2012, supplementary data) give the age of the Universe at recombination,

$$t_* = 376,000 \pm 4,000 \text{ yr} . \quad (10.98)$$

10.21 Horizon

Light can come from no more distant point than the Big Bang. This distant point defines what cosmologists traditionally refer to as the **horizon** (or particle horizon) of our Universe, located at infinite redshift, $z = \infty$. Equation (10.88) gives the geodesic distance between us at redshift zero and the horizon as

$$x_{\parallel}(\text{horizon}) = \int_0^{\infty} \frac{c \, dz}{H} . \quad (10.99)$$

The standard Λ CDM paradigm is based in part on the proposition that the Universe had an early inflationary phase, §10.22. If so, then there is no place where the redshift reaches infinity. However, the redshift is large at reheating, when inflation ends, and cosmologists call this the horizon,

$$x_{\parallel}(\text{horizon}) = \int_0^{\text{huge}} \frac{c \, dz}{H} . \quad (10.100)$$

Figure 10.12 shows a spacetime diagram of an FLRW Universe with cosmological parameters consistent with those of (Planck Collaboration, 2018). In this model, the comoving horizon distance to reheating is

$$x_{\parallel}(\text{horizon}) = 3.333 \, c/H_0 = 14.5 \text{ Gpc} = 47.2 \text{ Glyr} . \quad (10.101)$$

The redshift of reheating in this model has been taken at $z = 10^{28}$, but the horizon distance is insensitive to the choice of reheating redshift.

The horizon should be distinguished from the future horizon, which Hawking and Ellis (1973) define to be the farthest that an observer will ever be able to see in the indefinite future. If the Universe continues accelerating, as it is currently, then our future horizon will be finite, as illustrated in Figure 10.12.

A quantity that cosmologists sometimes refer to loosely as the horizon is the **Hubble distance**, defined to be

$$\text{Hubble distance} \equiv \frac{c}{H} , \quad (10.102)$$

The Hubble distance sets the characteristic scale over which two observers can communicate and influence each other, which is smaller than the horizon distance.

The standard Λ CDM model has the curious property that the Universe is switching from a matter-dominated period of deceleration to a vacuum-dominated period of acceleration. During deceleration, objects appear over the horizon, while during acceleration, they disappear over the horizon. Figure 10.13 illustrates the evolution of the observed redshifts of objects at fixed comoving distances. In the past decelerating phase,

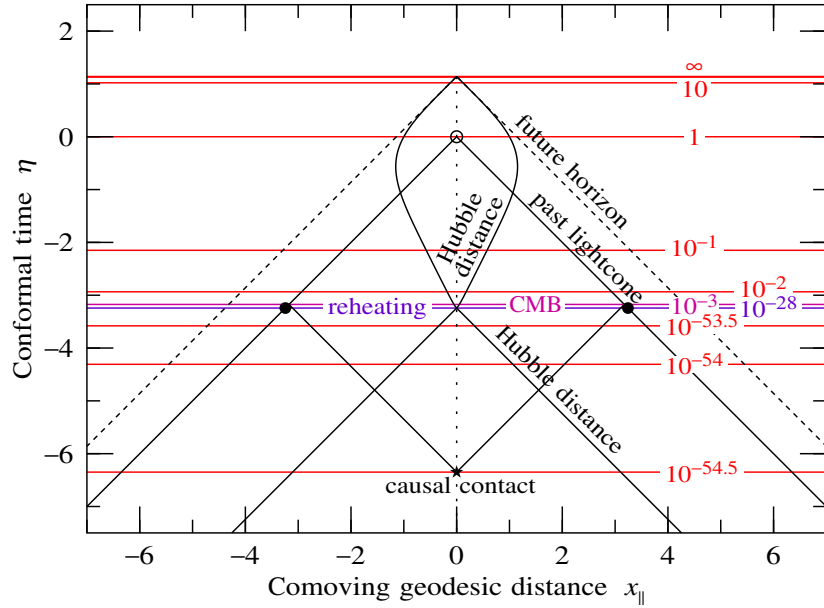


Figure 10.12 Spacetime diagram of an FLRW Universe in conformal coordinates η and $x_{||}$, in units of the present day Hubble distance c/H_0 . The unfilled circle marks our position, which is taken to be the origin of the conformal coordinates. In conformal coordinates, light moves at 45° on the spacetime diagram. The diagram is drawn for a flat Λ CDM model with $\Omega_\Lambda = 0.7$, $\Omega_m = 0.3$, and a radiation density such that the redshift of matter-radiation equality is 3400, consistent with Planck Collaboration (2018). Horizontal lines are lines of constant cosmic scale factor a , labelled by their values relative to the present, $a_0 = 1$. Reheating, at the end of inflation, has been taken to be at redshift 10^{28} . Filled dots mark the place that cosmologists traditionally call the horizon, at reheating, which is a place of large, but not infinite, redshift. Inflation offers a solution to the horizon problem because all points on the CMB within our past lightcone could have been in causal contact at an early stage of inflation. If dark energy behaves like a cosmological constant into the indefinite future, then we will have a future horizon.

the redshift of objects appearing over the horizon decreased rapidly from some huge value. In the future accelerating phase, the redshift of objects disappearing over the horizon will increase in proportion to the cosmic scale factor.

10.22 Inflation

Part of the Standard Model of Cosmology is the hypothesis that the early Universe underwent a period of **inflation**, when the mass-energy density was dominated by “vacuum” energy, and the Universe expanded exponentially, with $a \propto e^{Ht}$. The idea of inflation was originally motivated around 1980 by the idea that early in the Universe the forces of nature would be unified, and that there is energy associated with that unification. For example, the inflationary energy could be the energy associated with Grand Unification of

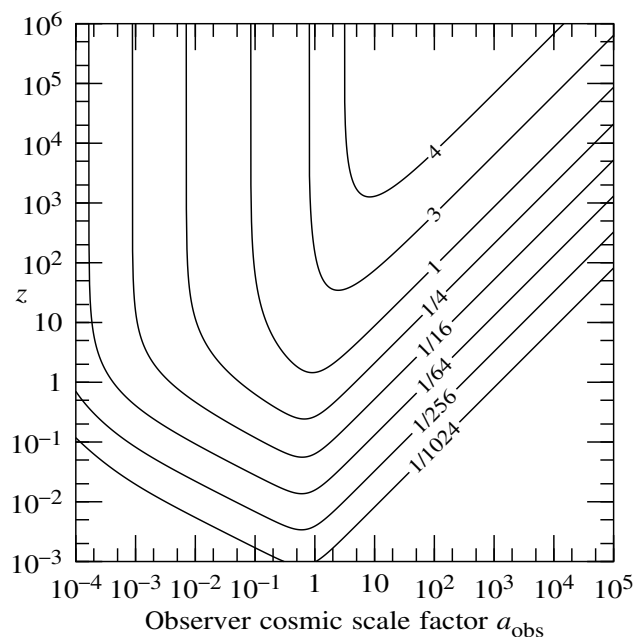


Figure 10.13 Redshift z of objects at fixed comoving distances as a function of the epoch a_{obs} at which an observer observes them. The label on each line is the comoving distance $x_{\parallel}$ in units of c/H_0 . The diagram is drawn for a flat Λ CDM model with $\Omega_{\Lambda} = 0.7$, $\Omega_m = 0.3$, and a radiation density such that the redshift of matter-radiation equality is 3400. The present-day Universe, at $a_{\text{obs}} = 1$, is transitioning from a decelerating, matter-dominated phase to an accelerating, vacuum-dominated phase. Whereas in the past redshifts tended to decrease with time, in the future redshifts will tend to increase with time.

the $U(1) \times SU(2) \times SU(3)$ forces of the standard model. The three coupling constants of the standard model vary slowly with energy, appearing to converge (almost) at an energy of around $m_{\text{GUT}} \sim 10^{16}$ GeV, not much less than the Planck energy of $m_P \sim 10^{19}$ GeV. The associated vacuum energy density would be of order $\rho_{\text{GUT}} \sim m_{\text{GUT}}^4$ in Planck units.

Alan Guth (1981) pointed out that, regardless of theoretical arguments for inflation, an early inflationary epoch would solve a number of observational conundra. The most important observational problem is the **horizon problem**, Exercise 10.12. If the Universe has always been dominated by radiation and matter, and therefore always decelerating, then up to the time of recombination light could only have travelled a distance corresponding to about 1 degree on the cosmic microwave background sky, Exercise 10.11. If that were the case, then how come the temperature at points in the cosmic microwave background more than a degree apart, indeed even 180° apart, on opposite sides of the sky, have the same temperature, even though they could never have been in causal contact? Guth pointed out that inflation could solve the horizon problem by allowing points to be initially in causal contact, then driven out of causal contact by the acceleration and consequent exponential expansion induced by vacuum energy, provide that the inflationary expansion

continued over a sufficient number e -folds, Exercise 10.12. Guth's solution is illustrated in the spacetime diagram in Figure 10.12.

Guth pointed out that inflation could solve some other problems, such as the **flatness problem**. However, most of these problems are essentially equivalent to the horizon problem, Exercise 10.13.

A distinct basic problem that inflation solves is the **expansion problem**. If the Universe has always been dominated by a gravitationally attractive form of mass-energy, such as matter or radiation, then how come the Universe is expanding? Inflation solves the problem because an initial period dominated by gravitationally repulsive vacuum energy could have accelerated the Universe into enormous expansion.

Inflation also offers an answer to the question of where the matter and radiation seen in the Universe today came from. Inflation must have come to an end, since the present day Universe does not contain the enormously high vacuum energy density that dominated during inflation (the vacuum energy during inflation was vast compared to the present-day cosmological constant). The vacuum energy must therefore have decayed into other forms of gravitationally attractive energy, such as matter and radiation. The process of decay is called **reheating**. Reheating is not well understood, because it occurred at energies well above those accessible to experiment today. Nevertheless, if inflation occurred, then so also did reheating.

Compelling evidence in favour of the inflationary paradigm comes from the fact that, in its simplest form, inflationary predictions for the power spectrum of fluctuations of the CMB fit astonishingly well to observational data, which continue to grow ever more precise.

Exercise 10.11 Horizon size at recombination.

1. **Comoving horizon distance.** Assume for simplicity a flat, matter-dominated Universe. From equation (10.99), what is the comoving horizon distance $x_{\parallel}$ as a function of cosmic scale factor a ?
2. **Angular size on the CMB of the horizon at recombination.** For a flat Universe, the angular size on the CMB of the horizon at recombination equals the ratio of the comoving horizon distance at recombination to the comoving distance between us and recombination. Recombination occurs at sufficiently high redshift that the latter distance approximates the comoving horizon at the present time. Estimate the angular size on the CMB of the horizon at recombination if the redshift of recombination is $z_{\text{rec}} \approx 1100$.

Exercise 10.12 The horizon problem.

1. **Expansion factor.** The temperature of the CMB today is $T_0 \approx 3\text{ K}$. By approximately what factor has the Universe expanded since the temperature was some initial high temperature, say the GUT temperature $T_i \approx 10^{29}\text{ K}$, or the Planck temperature $T_i \approx 10^{32}\text{ K}$?
2. **Hubble distance.** By what factor has the Hubble distance c/H increased during the expansion of part 1? Assume for simplicity that the Universe has been mainly radiation-dominated during this period, and that the Universe is flat. [Hint: For a flat Universe $H^2 \propto \rho$, and for radiation-dominated Universe $\rho \propto a^{-4}$.]
3. **Comoving Hubble distance.** Hence determine by what factor the comoving Hubble distance $x_H = c/(aH)$ has increased during the expansion of part 1.
4. **Comoving Hubble distance during inflation.** During inflation the Hubble distance c/H remained

constant, while the cosmic scale factor a expanded exponentially. What is the relation between the comoving Hubble distance $x_H = c/(aH)$ and cosmic scale factor a during inflation? [You should obtain an answer of the form $x_H \propto a^2$.]

5. **Number of e -foldings to solve the horizon problem.** By how many e -foldings must the Universe have inflated in order to solve the horizon problem? Assume again, as in part 1, that the Universe has been mainly radiation-dominated during expansion from the Planck temperature to the current temperature, and that this radiation-dominated epoch was immediately preceded by a period of inflation. [Hint: Inflation solves the horizon problem if the currently observable Universe was within the Hubble distance at the beginning of inflation, i.e. if the comoving $x_{H,0}$ now is less than the comoving Hubble distance $x_{H,i}$ at the beginning of inflation. The ‘number of e -foldings’ is $\ln(a_f/a_i)$, where $\ln$ is the natural logarithm, and a_i and a_f are the cosmic scale factors at the beginning (i for initial) and end (f for final) of inflation.]

Exercise 10.13 Relation between horizon and flatness problems. Show that Friedmann’s equation (10.30a) can be written in the form

$$\Omega - 1 = \kappa x_H^2, \quad (10.103)$$

where $x_H \equiv c/(aH)$ is the comoving Hubble distance. Use this equation to argue in your own words how the horizon and flatness problems are related.

10.23 Evolution of the size and density of the Universe

Figure 10.14 shows the evolution of the cosmic scale factor a as a function of time t predicted by the standard flat Λ CDM model, coupled with a plausible depiction of the early inflationary epoch. The parameters of the model are the same as those for Figure 10.12. In the model, the Universe starts with an inflationary phase, and transitions instantaneously at reheating to a radiation-dominated phase. Not long before recombination, the Universe goes over to a matter-dominated phase, then later to the dark-energy-dominated phase of today. The relation between cosmic time t and cosmic scale factor a is given by equation (10.71), and some relevant analytic results are in Exercises 10.7 and 10.8.

Figure 10.14 also shows the evolution of the Hubble distance c/H , which sets the approximate scale within which regions are in causal contact. The Hubble distance is constant during vacuum-dominated phases, but is approximately proportional to the age of the Universe at other times. The Figure illustrates that regions that are in causal contact prior to inflation can fly out of causal contact during the accelerated expansion of inflation. Once the Universe transitions to a decelerating radiation- or matter-dominated phase, regions that were out of causal contact can come back into causal contact, inside the Hubble distance.

Since inflation occurred at high energies inaccessible to experiment, the energy scale of inflation is unknown, and the number of e -folds during which inflation persisted is unknown. Figure 10.14 illustrates the case where the energy scale of inflation is around the GUT scale, and the number of e -folds is only slightly greater than

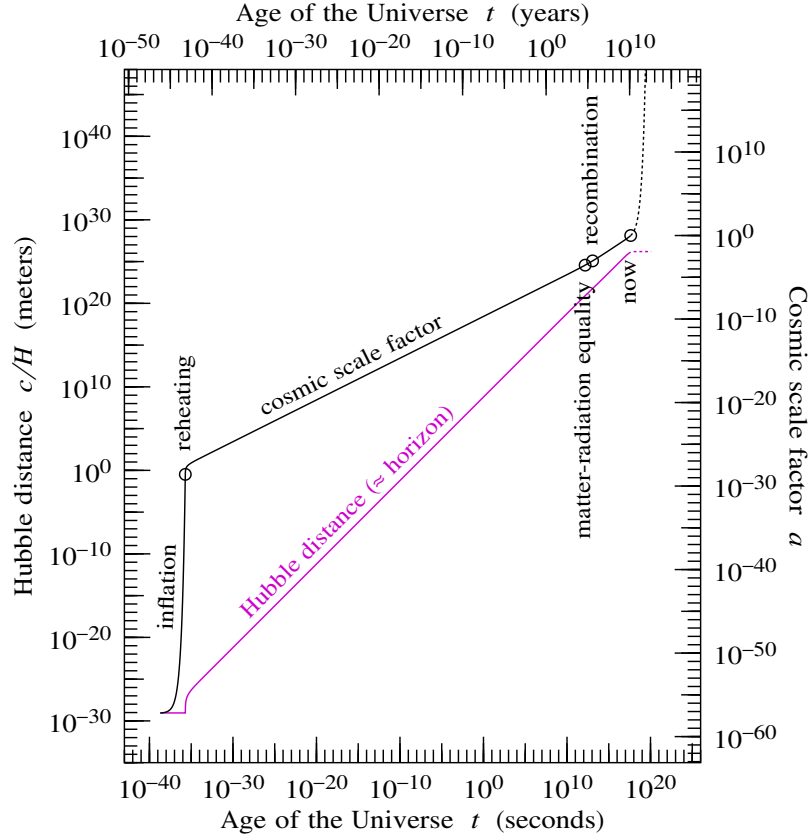


Figure 10.14 Cosmic scale factor a and Hubble distance c/H as a function of cosmic time t , for a flat Λ CDM model with the same parameters as in Figure 10.12. In this model, the Universe began with an inflationary epoch where the density was dominated by constant vacuum energy, the Hubble parameter H was constant, and the cosmic scale factor increased exponentially, $a \propto e^{Ht}$. The initial inflationary phase came to an end when the vacuum energy decayed into radiation energy, an event called reheating. The Universe then became radiation-dominated, evolving as $a \propto t^{1/2}$. At a redshift of $z_{\text{eq}} \approx 3400$ the Universe passed through the epoch of matter-radiation equality, where the density of radiation equalled that of (non-baryonic plus baryonic) matter. Matter-radiation equality occurred just prior to recombination, at $z_{\text{rec}} \approx 1090$ (Planck Collaboration, 2018). The Universe remained matter-dominated, evolving as $a \propto t^{2/3}$, until relatively recently (from a cosmological perspective). The Universe transitioned through matter-dark energy equality at $z_{\Lambda} \approx 0.4$. The dotted line shows how the cosmic scale factor and Hubble distance will evolve in the future, if the dark energy is a cosmological constant, and if it does not decay into some other form of energy.

the number necessary to solve the horizon problem. Figure 10.14 does not attempt to extrapolate to what might possibly have happened prior to inflation.

Figure 10.15 shows the mass-energy density ρ as a function of time t for the same flat Λ CDM model as shown in Figure 10.14. Since the Universe here is taken to be flat, the density equals the critical density

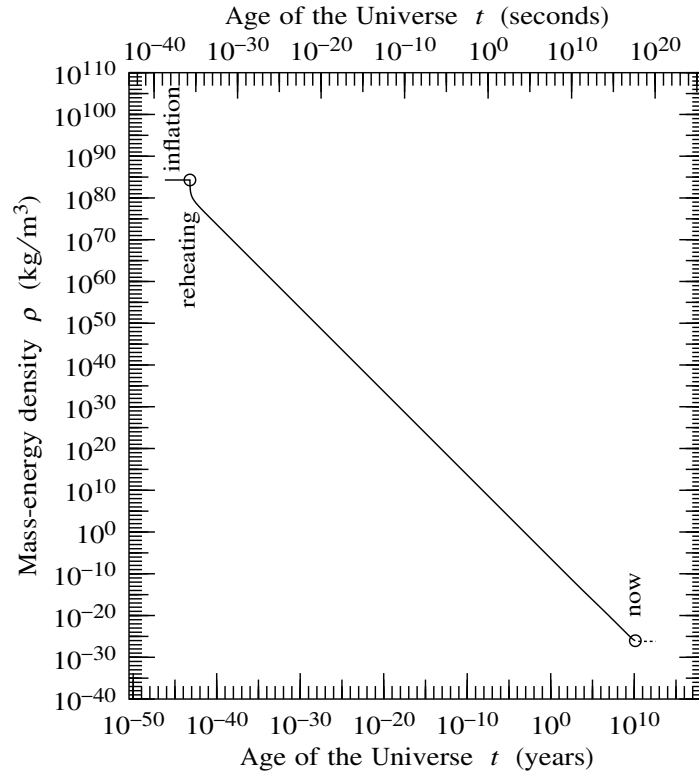


Figure 10.15 Mass-energy density ρ of the Universe as a function of cosmic time t corresponding to the evolution of the cosmic scale factor shown in Figure 10.14.

at all times, and is proportional to the inverse square of the Hubble distance c/H plotted in Figure 10.14. The energy density is constant during epochs dominated by vacuum energy, but decreases approximately as $\rho \propto t^{-2}$ at other times.

10.24 Evolution of the temperature of the Universe

Figure 10.16 shows the radiation (photon) temperature T as a function of time t corresponding to the evolution of the scale factor a and temperature T shown in Figures 10.14 and 10.15.

A system of photons in thermodynamic equilibrium has a blackbody distribution of energies. The CMB has a precise blackbody spectrum, not because it is in thermodynamic equilibrium today, but rather because the CMB was in thermodynamic equilibrium with electrons and nuclei at the time of recombination, and the CMB has streamed more or less freely through the Universe since recombination. A thermal distribution of

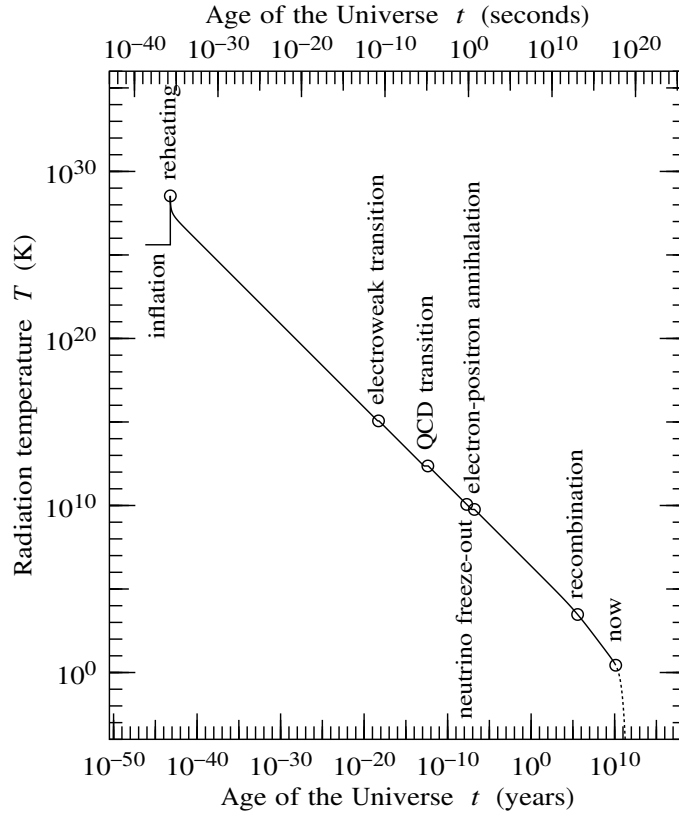


Figure 10.16 Radiation temperature T of the Universe as a function of cosmic time t corresponding to the evolution of the cosmic scale factor shown in Figure 10.14. The temperature during inflation was the Hawking temperature, equal to $H/(2\pi)$ in Planck units. After inflation and reheating, the temperature decreases as $T \propto a^{-1}$, modified by a factor depending on the effective entropy-weighted number g_s of particle species, equation (10.105). In this plot, the effective number g_s of relativistic particle species has been approximated as changing abruptly at three discrete temperatures, electron-positron annihilation, the QCD phase transition, and the electroweak phase transition, Table 10.4.

relativistic particles retains its thermal distribution in an expanding FLRW universe (albeit with a changing temperature), Exercise 10.14.

The evolution of the temperature of photons in the Universe can be deduced from conservation of entropy. The Friedmann equations imply the first law of thermodynamics, §10.9.2, and thus enforce conservation of entropy per comoving volume (but see Concept question 30.5). Entropy is conserved in an FLRW universe even when particles annihilate with each other. For example, electrons and positrons annihilated with each other when the temperature fell through $T \approx m_e = 511 \text{ keV}$, but the entropy lost by electrons and positrons was gained by photons, for no net change in entropy, Figure 10.17 and Exercise 10.22.

In the real Universe, entropy increases as a result of fluctuations away from the perfect homogeneity and

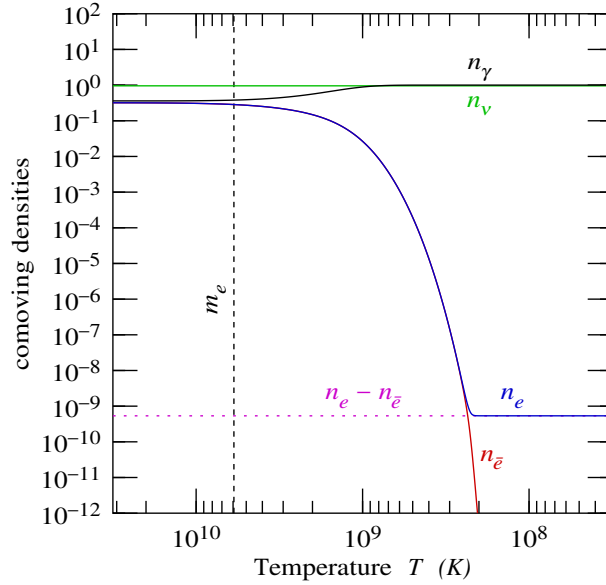


Figure 10.17 Comoving number densities $a^3 n$ of photons γ , neutrinos ν , electrons e , and positrons $\bar{e}$ as a function of temperature T around the temperature m_e near which electrons and positrons annihilate. Annihilating electrons and positrons dump their entropy into photons, increasing the comoving density of photons, while conserving total entropy per comoving volume. The comoving densities are normalized to $a^3 n_\gamma = 1$ at the present time. The calculations are described in Exercise 10.22.

isotropy assumed by the FLRW geometry. By far the biggest repositories of entropy in today's Universe are black holes, principally supermassive black holes. However, black holes are irrelevant to the CMB, since the CMB has propagated essentially unchanged since recombination. It is fine to compute the temperature of cosmological radiation from conservation of cosmological entropy.

The entropy of a system in thermodynamic equilibrium is approximately one per particle, Exercise 10.19. The number of particles in the Universe today is dominated by particles that were relativistic at the time they decoupled, namely photons and neutrinos, and these therefore dominate the cosmological entropy. The ratio $\eta_b \equiv n_b/n_\gamma$ of baryon to photon number in the Universe today is less than a billionth,

$$\eta_b \equiv \frac{n_b}{n_\gamma} = \frac{\epsilon_\gamma \Omega_b}{m_b \Omega_\gamma} = 6.1 \times 10^{-10} \frac{\Omega_b h^2}{0.0224} \left(\frac{T_0}{2.725 \text{ K}} \right)^{-3}, \quad (10.104)$$

where $\epsilon_\gamma = \pi^4 T_0 / (30\zeta(3)) = 2.701 T_0$ is the mean energy per photon (Exercise 10.16), and $m_b = 939 \text{ MeV}$ is the approximate mass per baryon. The value is as reported by the Planck team (Planck Collaboration, 2018).

Conservation of entropy per comoving volume implies that the photon temperature T at redshift z is

related to the present day photon temperature T_0 by (Exercise 10.20)

$$\frac{T}{T_0} = (1+z) \left(\frac{g_{s,0}}{g_s} \right)^{1/3}, \quad (10.105)$$

where g_s is the entropy-weighted effective number of relativistic particle species.

The other major contributors to cosmological entropy today, besides photons, are neutrinos and antineutrinos. Neutrinos decoupled at a temperature of about $T \approx 1$ MeV. Above that temperature weak interactions were fast enough to keep neutrinos and antineutrinos in thermodynamic equilibrium with protons and neutrons, hence with photons, but below that temperature neutrinos and antineutrinos froze out.

Neutrino oscillation data indicate that at least 2 of the 3 neutrino types have masses that would make cosmic neutrinos nonrelativistic at the present time, §10.25. Neutrino oscillations fix only differences in squared masses of neutrinos, leaving unconstrained the absolute mass levels. If the lightest neutrino has mass $m_\nu \lesssim 10^{-4}$ eV, equation (10.112), then it would remain relativistic at the present time, and it would produce a Cosmic Neutrino Background (CNB) analogous to the CMB. Because neutrinos froze out before $e\bar{e}$ -annihilation, annihilating electrons and positrons dumped their entropy into photons, increasing the temperature of photons relative to that of neutrinos. The temperature of the CNB today would be, Exercise 10.21,

$$T_\nu = \left(\frac{4}{11} \right)^{1/3} 2.725 \text{ K} = 1.945 \text{ K}. \quad (10.106)$$

Sadly, neutrinos interact too weakly for such a background to be detectable with current technology. Like the CMB, the CNB should have a (redshifted) thermal distribution inherited from being in thermodynamic equilibrium at $T \sim 1$ MeV.

Table 10.4 gives approximate values of the effective entropy-weighted number g_s of relativistic particle species over various temperature ranges. The extra factor of two for g_s in the final column of Table 10.4 arises because every particle species has an antiparticle (the two spin states of a photon can be construed as each other's antiparticle). The entropy of a relativistic fermionic species is 7/8 that of a bosonic species, Exercise 10.17, equation (10.142). The difference in photon and neutrino temperatures leads to an extra factor of 4/11 in the value of g_s today, which, with 1 bosonic species (photons) and 3 fermionic species (neutrinos), together with their antiparticles, is, equation (10.153),

$$g_{s,0} = 2 \left(1 + \frac{7}{8} \frac{4}{11} \cdot 3 \right) = \frac{43}{11} = 3.91. \quad (10.107)$$

A more comprehensive evaluation of g_s is given by Kolb and Turner (1990, Fig. 3.5), and Planck Collaboration (2018, Fig. 36). Over the range of energies $T \lesssim 1$ TeV covered by the standard model of physics, there are four principal epochs in the evolution of the effective number g_s of relativistic species, punctuated by electron-positron annihilation at $T \approx 0.5$ MeV, the QCD phase transition from bound nuclei to free quarks and gluons at $T \approx 200$ MeV, and the electroweak phase transition above which all standard model particles are relativistic at $T \approx 100$ GeV. There could well be further changes in the number of relativistic species at higher temperatures, for example at grand unification, or perhaps if supersymmetry becomes unbroken at some energy, but at present experimental data do not unambiguously constrain the possibilities.

Table 10.4: Effective entropy-weighted number of relativistic particle species

Temperature T	particles	spin	chiralities	generations	flavours	colours	multiplicity	g_s
$T \lesssim 0.5 \text{ MeV}$	photon γ	1					1	$2 \left(1 + \frac{7}{8} \frac{4}{11} 3 \right) = 3.91$
	neutrinos ν_e, ν_μ, ν_τ	$\frac{1}{2}$	1	3			3	
$0.5 \text{ MeV} \lesssim T \lesssim 200 \text{ MeV}$	photon γ	1					1	$2 \left(1 + \frac{7}{8} 5 \right) = 10.75$
	neutrinos ν_e, ν_μ, ν_τ	$\frac{1}{2}$	1	3			3	
	electron e	$\frac{1}{2}$	2	1			2	
$200 \text{ MeV} \lesssim T \lesssim 100 \text{ GeV}$	photon γ	1					1	$2 \left(9 + \frac{7}{8} 25 \right) = 61.75$
	SU(3) gluons	1					8	
	neutrinos ν_e, ν_μ, ν_τ	$\frac{1}{2}$	1	3			3	
	leptons e, μ	$\frac{1}{2}$	2	2			4	
	quarks u, d, s	$\frac{1}{2}$	2	$\frac{3}{2}$	2	3	18	
$T \gtrsim 100 \text{ GeV}$	SU(2) $\times$ U(1) $_Y$ bosons	1					3 + 1	$2 \left(14 + \frac{7}{8} 45 \right) = 106.75$
	SU(3) gluons	1					8	
	complex Higgs	0					2	
	neutrinos ν_e, ν_μ, ν_τ	$\frac{1}{2}$	1	3			3	
	leptons e, μ, τ	$\frac{1}{2}$	2	3			6	
	quarks u, d, c, s, t, b	$\frac{1}{2}$	2	3	2	3	36	

10.25 Neutrino mass

Neutrinos are created naturally by nucleosynthesis in the Sun, and by interaction of cosmic rays with the atmosphere. When a neutrino is created (or annihilated) by a weak interaction, it is created in a weak eigenstate. Observations of solar and atmospheric neutrinos indicate that neutrino species oscillate into each other, implying that the weak eigenstates are not mass eigenstates. The weak eigenstates are denoted ν_e, ν_μ , and ν_τ , while the mass eigenstates are denoted ν_1, ν_2 , and ν_3 . Oscillation data yield mass squared differences between the three mass eigenstates (Forero, Tortola, and Valle, [2012](#))

$$|\Delta m_{21}|^2 = (7.6 \pm 0.2) \times 10^{-5} \text{ eV}^2 \quad \text{solar neutrinos ,} \quad (10.108a)$$

$$|\Delta m_{31}|^2 = (2.4 \pm 0.1) \times 10^{-3} \text{ eV}^2 \quad \text{atmospheric neutrinos .} \quad (10.108b)$$

The data imply that at least two of the neutrino types have mass. The squared mass difference between m_1 and m_2 implies that at least one of them must have a mass

$$m_{\nu_1} \text{ or } m_{\nu_2} \geq \sqrt{7.6 \times 10^{-5} \text{ eV}^2} \approx 0.01 \text{ eV} . \quad (10.109)$$

The squared mass difference between m_1 and m_3 implies that at least one of them must have a mass

$$m_{\nu_1} \text{ or } m_{\nu_3} \geq \sqrt{2.4 \times 10^{-3} \text{ eV}^2} \approx 0.05 \text{ eV} . \quad (10.110)$$

The ordering of masses is undetermined by the data. The natural ordering is $m_1 < m_2 < m_3$, but an inverted hierarchy $m_3 < m_1 \approx m_2$ is possible. Constraints from the CMB impose an upper limit on the sum of the masses of the three neutrino types (Planck Collaboration, 2018),

$$\sum m_\nu < 0.12 \text{ eV} . \quad (10.111)$$

A direct measurement by the KATRIN experiment yields an upper limit of $m_{\nu_e} < 1.1 \text{ eV}$ on the mass of the electron neutrino (Aker et al., 2019).

The CNB temperature, equation (10.106), is $T_\nu = 1.945 \text{ K} = 1.676 \times 10^{-4} \text{ eV}$. The redshift at which a neutrino of mass m_ν becomes nonrelativistic is then

$$1 + z_\nu = \frac{m_\nu}{T_\nu} = \frac{m_\nu}{1.676 \times 10^{-4} \text{ eV}} . \quad (10.112)$$

Neutrinos of masses 0.01 eV and 0.05 eV would have become nonrelativistic at $z_\nu \approx 60$ and 300 respectively. Only a neutrino of mass $\lesssim 10^{-4} \text{ eV}$ would remain relativistic at the present time.

The masses from neutrino oscillation data suggest that at least two species of cosmological neutrinos are nonrelativistic today. If so, then the neutrino density Ω_ν today is related to the sum $\sum m_\nu$ of neutrino masses by

$$\Omega_\nu = \frac{8\pi G \sum m_\nu n_\nu}{3H_0^2} = 5.4 \times 10^{-4} \left(\frac{\sum m_\nu}{0.05 \text{ eV}} \right) h_{0.70}^{-2} . \quad (10.113)$$

The number and entropy densities of neutrinos today are unaffected by whether they are relativistic, so the effective number- and entropy-weighted numbers $g_{n,0}$ and $g_{s,0}$ are unaffected. On the other hand the energy density of neutrinos today does depend on whether or not they are relativistic. If just one neutrino type is relativistic and the other two are nonrelativistic, then the effective energy-weighted number $g_{\rho,0}$ of relativistic species today is

$$g_{\rho,0} = 2 + \left(\frac{4}{11} \right)^{4/3} \frac{7}{8} 2 = 2.45 . \quad (10.114)$$

The density Ω_r of relativistic particles today is $\Omega_r = (g_{\rho,0}/2)\Omega_\gamma$.

10.25.1 The neutrino mass puzzle

The experimental fact that neutrinos have mass is puzzling. The other salient experimental property of neutrinos is that they are left-handed (and anti-neutrinos are right-handed). A particle whose spin and

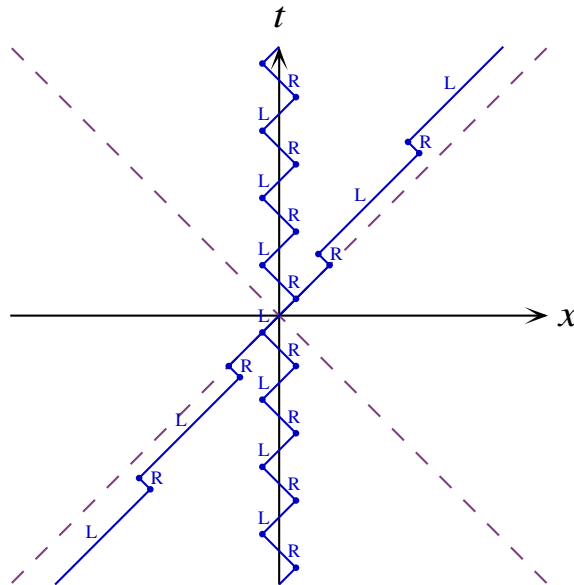


Figure 10.18 According to the Standard Model of physics, a massive fermion acquires its mass by interacting with the Higgs field. The interaction flips the fermion between left- and right-handed chiralities as it propagates through spacetime, as illustrated schematically in this spacetime diagram. In the fermion's rest frame, its wavefunction is a linear combination of left- and right-handed chiralities with equal amplitudes (in absolute value). Boosting the fermion in a direction opposite to its spin amplifies the left-handed component by a boost factor $e^{\theta/2}$ and deamplifies the right-handed component by $e^{-\theta/2}$, so a fermion moving relativistically appears almost entirely left-handed.

momentum point in the same direction is called right-handed, while a particle whose spin and momentum point in opposite directions is called left-handed. The handedness of a particle is also called its chirality. For massless particles, chirality is Lorentz-invariant: a massless particle that is purely left-handed in one frame remains purely left-handed in any Lorentz-transformed frame.

The problem is that a particle cannot be both massive and purely left- or right-handed. A massive particle that looks left-handed, spin anti-aligned with its momentum, in one frame, looks right-handed to an observer who overtakes the particle from behind. This does not immediately contradict the experimental fact that neutrinos are both massive and left-handed, since in all experiments neutrinos are highly relativistic, in which case the left-handed components are boosted exponentially compared to the right-handed components, as illustrated in Figure 10.18. But in principle an observer could overtake a left-handed neutrino, which the observer would then see as right-handed. But then where are the right-handed neutrinos? It is not enough to say that right-handed neutrinos are too weakly interacting to have been observed. A right-handed neutrino observed from behind would look like a left-handed neutrino and thereby become interacting, so right-handed neutrinos should make themselves felt in cosmology.

A leading idea to solve the problem of neutrino mass is that neutrinos are so-called Majorana fermions,

which have the defining property that when observed from behind they not only switch from left- to right-handed, but also from particle to antiparticle. Thus a left-handed neutrino observed from behind looks like a right-handed antineutrino. Switching from particle to antiparticle would violate charge conservation, so other fermions, namely electrons and quarks, cannot be Majorana fermions because they possess conserved charges (electric charge and color charge). Left-handed neutrinos have weak isospin and weak hypercharge, but those charges are not strictly conserved at energies below the ~ 100 GeV scale at which the electroweak $U(1)_Y \times SU(2)$ symmetry breaks down to the $U(1)_{\text{em}}$ electromagnetic symmetry. Thus at energies below the electroweak scale, neutrinos can be massive Majorana fermions without violating any strict conservation law.

The problem of neutrino mass is resumed in §48.4.1.

10.26 Occupation number, number density, and energy-momentum

A careful treatment of the evolution of the number and energy-momentum densities of species in an FLRW universe requires consideration of their momentum distributions.

In this section, including the Exercises, units c , $\hbar$, and G are kept explicit, but the Boltzmann constant is set to unity, $k = 1$, which is equivalent to measuring temperature T in units of energy.

10.26.1 Occupation number

Choose a locally inertial frame attached to an observer. The distribution of a particle species in the observer's frame is described by a dimensionless scalar occupation number $f(t, \mathbf{x}, \mathbf{p})$ that specifies the number dN of particles at the observer's position $x^\mu \equiv \{t, \mathbf{x}\}$ with momentum $p^k \equiv \{E, \mathbf{p}\}$ in a dimensionless Lorentz-invariant 6-dimensional volume of phase space,

$$dN = f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3}, \quad (10.115)$$

with g being the number of spin states of the particle. Here d^3r and d^3p denote the proper spatial and momentum 3-volume elements in the observer's locally inertial frame. The quantum mechanical normalization factor $(2\pi\hbar)^3$ ensures that f counts the number of particles per free-particle quantum state. If the particle species has rest mass m , then its energy E is related to its momentum by $E^2 - p^2 c^2 = m^2 c^4$, which explains why the occupation number is treated as a function only of momentum $\mathbf{p}$.

The phase-space volume element $d^3r d^3p$ is a scalar, invariant under Lorentz transformations of the observer's frame. In fact, as shown in §4.22.1, the phase-space volume element is invariant under any canonical transformation of coordinates and momenta, which includes not only Lorentz transformations but also a broad range of other transformations. For example, in place of $d^3r d^3p$ it would be possible to use the phase-space volume element $d^3x d^3\pi$ formed out of the spatial comoving coordinates x^α and their conjugate generalized momenta π_α .

The Lorentz invariance of the phase-space volume element $d^3r d^3p$ can be demonstrated more simplistically

as follows. First, the 3-volume element d^3r is related to the scalar 4-volume element $dt d^3r$ by

$$\frac{dt d^3r}{d\lambda} = E d^3r , \quad (10.116)$$

since $dt/d\lambda = E$. The left hand side of equation (10.116) is the derivative of the observer's 4-volume $dt d^3r$ with respect to the affine parameter $d\lambda \equiv d\tau/m$, with τ the observer's proper time. Since both the 4-volume and affine parameter are scalars, it follows that $E d^3r$ is a scalar (actually, $dt d^3r = dt dr^1 dr^2 dr^3$ is a pseudoscalar, not a scalar, as is $E d^3r = p^0 dr^1 dr^2 dr^3$; see Chapter 15). Second, the momentum 3-volume element d^3p is related to the scalar 4-volume element $dE d^3p$ by

$$\delta(E^2 - p^2 c^2 - m^2 c^4) dE d^3p = \frac{d^3p}{2E} , \quad (10.117)$$

where the Dirac delta-function enforces conservation of the particle rest mass m . The 4-volume d^4p is a scalar, and the delta-function is a function of a scalar argument, hence d^3p/E is likewise a scalar (again, $dE d^3p = -dp_0 dp_1 dp_2 dp_3$ and $d^3p/E = dp_1 dp_2 dp_3/p^0$ are actually pseudoscalars, not scalars). Since $E d^3r$ and d^3p/E are both Lorentz-invariant (pseudo-)scalars, so is their product, the phase space volume $d^3r d^3p$ (which is a genuine scalar).

10.26.2 Occupation number in an FLRW universe

The homogeneity and isotropy of an FLRW universe imply that, for a comoving observer, the occupation number f is independent of position and direction,

$$f(t, \mathbf{x}, \mathbf{p}) = f(t, p) . \quad (10.118)$$

10.26.3 Number density

In the locally inertial frame of an observer, the number density and flux of a particle species form a 4-vector n^k ,

$$n^k = \int p^k f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3p}{E(2\pi\hbar)^3} . \quad (10.119)$$

In particular, the number density n^0 , with units number of particles per unit proper volume, is the time component of the number current,

$$n^0 = \int f \frac{g d^3p}{(2\pi\hbar)^3} . \quad (10.120)$$

In an FLRW universe, the spatial components of the number flux vanish by isotropy, so the only non-vanishing component is the time component n^0 , which is just the proper number density n of the particle species,

$$n \equiv n^0 = \int f(t, p) \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} . \quad (10.121)$$

10.26.4 Energy-momentum tensor

In the locally inertial frame of an observer, the energy-momentum tensor T^{kl} of a particle species is

$$T^{kl} = \int p^k p^l f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3 p}{E (2\pi\hbar)^3} . \quad (10.122)$$

For an FLRW universe, homogeneity and isotropy imply that the energy-momentum tensor in the locally inertial frame of a comoving observer is diagonal, with time component $T^{00} = \rho$, and isotropic spatial components $T^{ab} = p \delta_{ab}$. The proper energy density ρ of a particle species is

$$\rho = \int E f(t, p) \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} , \quad (10.123)$$

and the proper isotropic pressure p is (don't confuse pressure p on the left hand side with momentum p on the right hand side)

$$p = \int \frac{p^2}{3E} f(t, p) \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} . \quad (10.124)$$

10.27 Occupation numbers in thermodynamic equilibrium

Frequent collisions tend to drive a system towards thermodynamic equilibrium. Electron-photon scattering keeps photons in near equilibrium with electrons, while Coulomb scattering keeps electrons in near equilibrium with ions, primarily hydrogen ions (protons) and helium nuclei. Thus photons and baryons can be treated as having unperturbed distributions in mutual thermodynamic equilibrium.

In thermodynamic equilibrium at temperature T , the occupation numbers of fermions, which obey an exclusion principle, and of bosons, which obey an anti-exclusion principle, are

$$f = \begin{cases} \frac{1}{e^{(E-\mu)/T} + 1} & \text{fermion} , \\ \frac{1}{e^{(E-\mu)/T} - 1} & \text{boson} , \end{cases} \quad (10.125)$$

where μ is the chemical potential of the species. In the limit of small occupation numbers, $f \ll 1$, equivalent to large negative chemical potential, $\mu \rightarrow -\text{large}$, both fermion and boson distributions go over to the Boltzmann distribution

$$f = e^{(-E+\mu)/T} \quad \text{Boltzmann} . \quad (10.126)$$

Chemical potential is the thermodynamic potential associated with conservation of number. There is a distinct potential for each conserved species. For example, radiative recombination and photoionization of hydrogen,



separately preserves proton and electron number, hydrogen being composed of one proton and one electron.

In thermodynamic equilibrium, the chemical potential μ_H of hydrogen is the sum of the chemical potentials μ_p and μ_e of protons and electrons,

$$\mu_p + \mu_e = \mu_H . \quad (10.128)$$

Photon number is not conserved, so photons have zero chemical potential,

$$\mu_\gamma = 0 , \quad (10.129)$$

which is closely associated with the fact that photons are their own antiparticles. For photons, which are bosons, the thermodynamic distribution (10.125) becomes the Planck distribution,

$$f = \frac{1}{e^{E/T} - 1} \quad \text{Planck} . \quad (10.130)$$

Exercise 10.14 Distribution of non-interacting particles initially in thermodynamic equilibrium. The number dN of a particle species in an interval $d^3r d^3p$ of phase space (proper positions r and proper momenta p , not to be confused with the same symbol p for pressure) for an ideal gas of free particles (nonrelativistic, relativistic, or anything in between) in thermodynamic equilibrium at temperature T and chemical potential μ is

$$dN = f \frac{g d^3r d^3p}{(2\pi\hbar)^3} , \quad (10.131)$$

where the occupation number f is (units $k = 1$, where k is the Boltzmann constant)

$$f = \frac{1}{e^{(E-\mu)/T} \pm 1} , \quad (10.132)$$

with a $+$ sign for fermions and a $-$ sign for bosons. The energy E and momentum p of particles of mass m are related by $E^2 = p^2 c^2 + m^2 c^4$. For bosons, the chemical potential is constrained to satisfy $\mu \leq E$, but for fermions μ may take any positive or negative value, with $\mu \gg E$ corresponding to a degenerate Fermi gas. As the Universe expands, proper distance increase as $r \propto a$, while proper momenta decrease as $p \propto a^{-1}$, so the phase space volume $d^3r d^3p$ remains constant.

1. **Occupation number.** Write down an expression for the occupation number $f(p)$ of a distribution of particles that start in thermodynamic equilibrium and then remain non-interacting while the Universe expands by a factor a .
2. **Relativistic particles.** Conclude that a distribution of non-interacting relativistic particles initially in thermodynamic equilibrium retains its thermodynamic equilibrium distribution in an FLRW universe as long as the particles remain relativistic. How do the temperature T and chemical potential μ of the relativistic distribution vary with cosmic scale factor a ?
3. **Nonrelativistic particles.** Show similarly that a distribution of non-interacting nonrelativistic particles initially in thermodynamic equilibrium remains thermal. How do the temperature T and chemical potential $\mu - m$ of the nonrelativistic distribution vary with cosmic scale factor a ?

4. **Transition from relativistic to nonrelativistic.** What happens to a distribution of non-interacting particles that are relativistic in thermodynamic equilibrium, but redshift to being nonrelativistic?

Exercise 10.15 The first law of thermodynamics with non-conserved particle number. As seen in §10.9.2, the first law of thermodynamics in the form

$$T d(a^3 s) = d(a^3 \rho) + p d(a^3) = 0 \quad (10.133)$$

is built into Friedmann's equations. But what happens when for example the temperature falls through the temperature $T \approx 0.5 \text{ MeV}$ at which electrons and positrons annihilate? Won't there be entropy production associated with $e\bar{e}$ annihilation? Should not the first law of thermodynamics actually say

$$T d(a^3 s) = d(a^3 \rho) + p d(a^3) - \sum_X \mu_X d(a^3 n_X) , \quad (10.134)$$

with the last term taking into account the variation in the number $a^3 n_X$ of various species X ?

Solution. Each distinct chemical potential μ_X is associated with a conserved number, so the additional terms contribute zero change to the entropy,

$$\sum_X \mu_X d(a^3 n_X) = 0 , \quad (10.135)$$

as long as the species are in mutual thermodynamic equilibrium. For example, positrons and electrons in thermodynamic equilibrium satisfy $\mu_{\bar{e}} = -\mu_e$, and

$$\mu_e d(a^3 n_e) + \mu_{\bar{e}} d(a^3 n_{\bar{e}}) = \mu_e d(a^3 n_e - a^3 n_{\bar{e}}) = 0 , \quad (10.136)$$

which vanishes because the difference $a^3 n_e - a^3 n_{\bar{e}}$ between electron and positron number is conserved. Thus the entropy conservation equation (10.133) remains correct in an FLRW universe even when number changing processes are occurring.

Exercise 10.16 Number, energy, pressure, and entropy of a relativistic ideal gas at zero chemical potential. The number density n , energy density ρ , and pressure p of an ideal gas of a single species of free particles are given by equations (10.121), (10.123), and (10.124), with occupation number (10.132). Show that for an ideal relativistic gas of g bosonic species in thermodynamic equilibrium at temperature T and zero chemical potential, $\mu = 0$, the number density n , energy density ρ , and pressure p are (units $k = 1$; number density n in units 1/volume, energy density ρ and pressure p in units energy/volume)

$$n = g \frac{\zeta(3) T^3}{\pi^2 c^3 \hbar^3} , \quad \rho = 3p = g \frac{\pi^2 T^4}{30 c^3 \hbar^3} , \quad (10.137)$$

where $\zeta(3) = 1.2020569$ is a Riemann zeta function. The entropy density s of an ideal gas of free particles in thermodynamic equilibrium at zero chemical potential is

$$s = \frac{\rho + p}{T} . \quad (10.138)$$

Conclude that the entropy density s of an ideal relativistic gas of g bosonic species in thermodynamic equilibrium at temperature T and zero chemical potential is (units 1/volume)

$$s = g \frac{2\pi^2 T^3}{45c^3 \hbar^3} . \quad (10.139)$$

Exercise 10.17 A relation between thermodynamic integrals. Prove that

$$\int_0^\infty \frac{x^{n-1} dx}{e^x + 1} = (1 - 2^{1-n}) \int_0^\infty \frac{x^{n-1} dx}{e^x - 1} . \quad (10.140)$$

[Hint: Use the fact that $(e^x + 1)(e^x - 1) = (e^{2x} - 1)$.] Hence argue that the ratios of number, energy, and entropy densities of relativistic fermionic (f) to relativistic bosonic (b) species in thermodynamic equilibrium at the same temperature are

$$\frac{n_f}{n_b} = \frac{3}{4} , \quad \frac{\rho_f}{\rho_b} = \frac{s_f}{s_b} = \frac{7}{8} . \quad (10.141)$$

Conclude that if the number n , energy ρ , and entropy s of a mixture of bosonic and fermionic species in thermodynamic equilibrium at the same temperature T are written in the form of equations (10.137) and (10.139), then the effective number-, energy-, and entropy-weighted numbers g of particle species are, in terms of the number g_b of bosonic and g_f of fermionic species,

$$g_n = g_b + \frac{3}{4}g_f , \quad g_\rho = g_s = g_b + \frac{7}{8}g_f . \quad (10.142)$$

Exercise 10.18 Relativistic particles in the early Universe had approximately zero chemical potential. Show that the small particle-antiparticle symmetry of our Universe implies that to a good approximation relativistic particles in thermodynamic equilibrium in the early Universe had zero chemical potential.

Solution. The chemical potentials of particles X and antiparticles $\bar{X}$ in thermodynamic equilibrium are necessarily related by

$$\mu_{\bar{X}} = -\mu_X . \quad (10.143)$$

If the particle-antiparticle asymmetry is denoted η , defined for relativistic particles by

$$n_X - n_{\bar{X}} = \eta n_X , \quad (10.144)$$

then $\mu_X/T \sim \eta$. More accurately, to linear order in η ,

$$\frac{\mu_X}{T} \approx \eta \frac{\pi^2}{3\zeta(3)} \times \begin{cases} 1 & \text{(bosons)} \\ \frac{2}{3} & \text{(fermions)} \end{cases} . \quad (10.145)$$

Exercise 10.19 Entropy per particle. The entropy of an ideal gas of free particles in thermodynamic equilibrium is

$$s = \frac{\rho + p - \mu n}{T} . \quad (10.146)$$

Argue that the entropy per particle s/n is a quantity of order unity, whether particles are relativistic or nonrelativistic.

Solution. For relativistic bosons with zero chemical potential, equations (10.137) and (10.139) imply that the entropy per particle is

$$\frac{s}{n} = \frac{2\pi^4}{45\zeta(3)} \times \begin{cases} 1 & = 3.6 \quad (\text{bosons}) , \\ \frac{7}{6} & = 4.2 \quad (\text{fermions}) . \end{cases} \quad (10.147)$$

For a nonrelativistic species, the number density n is related to the temperature T and chemical potential μ by

$$n = g \left(\frac{mT}{2\pi\hbar^2} \right)^{3/2} e^{(\mu-m)/T} . \quad (10.148)$$

Under cosmological conditions, the occupation number of nonrelativistic species was small, $e^{(\mu-m)/T} \ll 1$. However, tiny occupation numbers correspond to values of $(\mu - m)/T$ that are only logarithmically large (negative). The entropy per particle of a nonrelativistic species is

$$\frac{s}{n} = \frac{5}{2} + \ln \left[\frac{g}{n} \left(\frac{mT}{2\pi\hbar^2} \right)^{3/2} \right] , \quad (10.149)$$

which remains modest even if the argument of the logarithm is huge.

Exercise 10.20 Photon temperature at high redshift versus today. Use entropy conservation, $a^3 s = \text{constant}$, to argue that the ratio of the photon temperature T at redshift z in the early Universe to the photon temperature T_0 today is as given by equation (10.105).

Exercise 10.21 Cosmic Neutrino Background. Neutrino oscillation data imply mass squared differences that indicate that at least 2 of the 3 neutrino types are massive today, equations (10.109) and (10.110). The oscillation data do not constrain the offset from zero mass. A neutrino of mass $\lesssim 10^{-4}$ eV would remain relativistic at the present time, equation (10.112), and would produce a Cosmic Neutrino Background. Neutrinos that are nonrelativistic today would have clustered gravitationally, similar to collisionless non-baryonic dark matter, except that the fermionic character of neutrinos means that they could become degenerate (occupation number almost 1) in regions of high density, such as in the cores of galaxies.

1. **Temperature of the CNB.** Weak interactions were fast enough to keep neutrinos in thermodynamic equilibrium with protons and neutrons, hence with photons, electrons, and positrons up to just before $e\bar{e}$ annihilation, but then neutrinos decoupled. When electrons and positrons annihilated, they dumped their entropy into that of photons, leaving the entropy of neutrinos unchanged. Argue that conservation of comoving entropy implies

$$a^3 T^3 \left(g_\gamma + \frac{7}{8} g_e \right) = T_\gamma^3 g_\gamma , \quad (10.150a)$$

$$a^3 T^3 g_\nu = T_\nu^3 g_\nu , \quad (10.150b)$$

where the left hand sides refer to quantities before $e\bar{e}$ annihilation, which happened at $T \sim 0.5$ MeV, and

the right hand sides to quantities after $e\bar{e}$ annihilation (including today). Deduce the ratio of neutrino to photon temperatures today,

$$\frac{T_\nu}{T_\gamma} . \quad (10.151)$$

Does the temperature ratio (10.151) depend on the number of neutrino types? What is the neutrino temperature today in K, if the photon temperature today is 2.725 K?

2. **Effective number of relativistic particle species.** Because the temperatures of photons and neutrinos are different, the effective number g of relativistic species today is not given by equations (10.142). What are the effective number-, energy-, and entropy-weighted numbers $g_{n,0}$, $g_{\rho,0}$, and $g_{s,0}$ of relativistic particle species today? What are their arithmetic values if the relativistic species consist of photons and three species of neutrino? How are these values altered if, as is likely, neutrinos today are nonrelativistic?

Solution. The ratio of neutrino to photon temperatures after $e\bar{e}$ annihilation is

$$\frac{T_\nu}{T_\gamma} = \left(\frac{g_\gamma}{g_\gamma + \frac{7}{8} g_e} \right)^{1/3} = \left(\frac{4}{11} \right)^{1/3} = 0.714 . \quad (10.152)$$

No, the temperature ratio does not depend on the number of neutrino types. The ratio depends on neutrinos having decoupled a short time before $e\bar{e}$ -annihilation. Equation (10.152) implies that the CNB temperature is given by equation (10.106). With 2 bosonic degrees of freedom from photons, and 6 fermionic degrees of freedom from 3 relativistic neutrino types, the effective number-, energy-, and entropy-weighted number of relativistic degrees of freedom is

$$g_{n,0} = g_\gamma + \left(\frac{T_\nu}{T_\gamma} \right)^3 \frac{3}{4} g_\nu = 2 + \frac{4}{11} \frac{3}{4} 6 = \frac{40}{11} = 3.64 , \quad (10.153a)$$

$$g_{\rho,0} = g_\gamma + \left(\frac{T_\nu}{T_\gamma} \right)^4 \frac{7}{8} g_\nu = 2 + \left(\frac{4}{11} \right)^{4/3} \frac{7}{8} 6 = 3.36 , \quad (10.153b)$$

$$g_{s,0} = g_\gamma + \left(\frac{T_\nu}{T_\gamma} \right)^3 \frac{7}{8} g_\nu = 2 + \frac{4}{11} \frac{7}{8} 6 = \frac{43}{11} = 3.91 . \quad (10.153c)$$

Neutrinos today interact too weakly to annihilate, so their number and entropy today is that of relativistic species even if they are nonrelativistic today. However, their energy density today is not that of relativistic particles.

Exercise 10.22 Abundance of electrons and positrons in thermodynamic equilibrium. Calculate and plot the comoving number densities $a^3 n$ of photons, electrons and positrons in thermodynamic equilibrium as the temperature T cooled through the electron mass m_e .

Solution. The results are shown in Figure 10.17.

Start by considering the more general situation of an ideal gas of any species, either fermionic or bosonic, rest mass m , in thermodynamic equilibrium at temperature T and chemical potential μ in a volume V . The

logarithm of the grand partition function Z_G of such an ideal gas is (units $c = k = 1$)

$$\ln Z_G = V \int \pm \ln \left[1 \pm e^{(-E+\mu)/T} \right] \frac{g d^3p}{(2\pi\hbar)^3}, \quad (10.154)$$

where the $\pm$ signs are $+$ for fermions, $-$ for bosons. The laws of thermodynamics state that energy density ρ , number density n , and pressure p (not to be confused with same symbol for momentum p) in thermodynamic equilibrium are given by partial derivatives of $\ln Z_G$ with respect to $-1/T$, μ/T , and V ,

$$d \ln Z_G = \rho V d \left(-\frac{1}{T} \right) + n V d \left(\frac{\mu}{T} \right) + \frac{p}{T} dV. \quad (10.155)$$

For an ideal gas, $\ln Z_G$ is proportional to volume V , and

$$\frac{\ln Z_G}{V} = \frac{p}{T} = s - \frac{\rho}{T} + \frac{\mu n}{T}, \quad (10.156)$$

where s is the entropy density.

At the present time, the observed small baryon-to-photon ratio n_b/n_γ of the Universe implies a similarly small electron-to-photon ratio n_e/n_γ , from equations (10.104) and (31.8),

$$\frac{n_e}{n_\gamma} = \frac{f_+ n_b}{n_\gamma} = 5.4 \times 10^{-10}. \quad (10.157)$$

The small electron-to-photon ratio today implies a small electron-positron asymmetry before electron-positron annihilation, implying $\mu_e/T \ll 1$ before electron-positron annihilation.

As long as the particle-antiparticle symmetry is small when relativistic, an approximation to the grand partition function that holds asymptotically at both high and low temperatures, and is accurate to better than 5% at intermediate temperatures, is

$$\ln Z_G \approx \frac{gVT^3}{2\pi^2\hbar^3} e^{(\mu-m)/T} c_0 \left(1 + c_1 \frac{m}{T} \right)^{3/2}, \quad (10.158)$$

where the constants c_0 and c_1 for respectively fermions and bosons are

$$c_0 \equiv \left\{ \frac{7}{8}, 1 \right\} \frac{\pi^4}{45} \approx \{1.894, 2.165\}, \quad c_1 \equiv \left(\frac{\pi}{2c_0^2} \right)^{1/3} = \{0.759, 0.695\}. \quad (10.159)$$

The partial derivatives (10.155) of the approximate logarithmic grand partition function (10.158) yield the number density n , energy density ρ , and pressure p ,

$$n \approx \frac{gT^3}{2\pi^2\hbar^3} e^{(\mu-m)/T} c_0 \left(1 + c_1 \frac{m}{T} \right)^{3/2}, \quad (10.160a)$$

$$\rho \approx n(m + qT), \quad (10.160b)$$

$$p \approx nT, \quad (10.160c)$$

where the factor q is

$$q \equiv \frac{3 + \frac{3}{2}c_1 m/T}{1 + c_1 m/T}, \quad (10.161)$$

which varies from $q = 3$ at $T \gg m$ to $q = \frac{3}{2}$ at $T \ll m$. The entropy density s is

$$s = \frac{\rho + p - \mu n}{T} \approx \left(1 + q + \frac{m - \mu}{T}\right) n. \quad (10.162)$$

The entropy in photons, which have $q = 3$ and $m = \mu = 0$, is $s_\gamma = 4n_\gamma$. The total entropy in all particles can be written

$$s = 2g_s n_\gamma, \quad (10.163)$$

which defines the effective entropy-weighted number g_s of relativistic species. The total comoving entropy $a^3 s$ is conserved. Conservation of comoving entropy implies that the cube of the product of scale factor a and temperature T is inversely proportional to the effective entropy-weighted number g_s of relativistic species,

$$\left(\frac{aT}{a_0 T_0}\right)^3 = \frac{g_{s,0}}{g_s}. \quad (10.164)$$

In the problem being considered, when electrons and positrons annihilate, they dump their entropy into photons, conserving the total comoving entropy of photons, electrons, and positrons as the Universe expands. The effective entropy-weighted number g_s of photons γ , electrons e , and positrons $\bar{e}$ through electron-positron annihilation is approximately

$$\begin{aligned} g_s \equiv \frac{s}{2n_\gamma} &\approx \frac{1}{2n_\gamma} \left[4n_\gamma + \left(1 + q_e + \frac{m_e - \mu_e}{T}\right) n_e + \left(1 + q_e + \frac{m_e + \mu_e}{T}\right) n_{\bar{e}} \right] \\ &\approx 2 + \frac{7}{8} \left[\left(1 + q_e + \frac{m_e}{T}\right) \cosh(\mu_e/T) - \frac{\mu_e}{T} \sinh(\mu_e/T) \right] e^{-m_e/T} \left(1 + c_1 \frac{m_e}{T}\right)^{3/2}. \end{aligned} \quad (10.165)$$

For the purposes of calculating how the cosmic scale factor a changes with temperature T during electron-positron annihilation, it suffices to approximate the electron chemical potential as zero, $\mu_e \approx 0$, since before annihilation, when electrons and positrons are relativistic, the chemical potential is much less than the temperature, $\mu_e/T \ll 1$, and after annihilation electrons (and positrons) contribute little to the entropy, and the value of the chemical potential ceases to make much difference. Thus the effective entropy-weighted number g_s of photons, electrons, and positrons is approximately

$$g_s \approx 2 + \frac{7}{8} \left(1 + q_e + \frac{m_e}{T}\right) e^{-m_e/T} \left(1 + c_1 \frac{m_e}{T}\right)^{3/2}. \quad (10.166)$$

Inserting the expression (10.166) into equation (10.164) yields the cosmic scale factor a in terms of temperature T through electron-positron annihilation.

An expression for chemical potential μ_e is needed to calculate the number densities of electrons and positrons through electron-positron annihilation. The chemical potential can be deduced from conservation of the comoving difference $a^3(n_e - n_{\bar{e}})$ in the number densities of electrons and positrons.

The approximation (10.160a), coupled with the thermodynamic equilibrium condition $\bar{\mu} = -\mu$, implies that the difference $n - \bar{n}$ between the number densities of particles and antiparticles in thermodynamic equilibrium approximates

$$n - \bar{n} \approx \frac{gT^3}{\pi^2 \hbar^3} \sinh\left(\frac{\mu}{T}\right) e^{-m/T} c'_0 \left(1 + c'_1 \frac{m}{T}\right)^{3/2}. \quad (10.167)$$

In the approximation (10.160a), the constants c'_0 and c'_1 in equation (10.167) are the same as the constants c_0 and c_1 given by equations (10.159); but a more accurate approximation for the difference $n - \bar{n}$, equation (10.167), uses instead the constants c'_0 and c'_1 defined by, for respectively fermions and bosons,

$$c'_0 \equiv \left\{ \frac{3}{4}, 1 \right\} 2\zeta(3) \approx \{1.803, 2.404\}, \quad c'_1 \equiv \left(\frac{\pi}{2c_0^2} \right)^{1/3} = \{0.785, 0.648\}, \quad (10.168)$$

with $\zeta(3) \approx 1.202$ the Riemann zeta function. The approximation (10.167) with constants given by equations (10.168) is asymptotically correct at both high and low temperatures, and is accurate to better than 5% at intermediate temperatures. Putting together equations (10.167), (10.164), and (10.166) yields

$$\left. \frac{n_e}{n_\gamma} \right|_0 = \frac{g_{s,0}(n_e - n_{\bar{e}})}{g_s n_\gamma} = \frac{g_{s,0}}{g_s} 2 \sinh\left(\frac{\mu_e}{T}\right) e^{-m_e/T} 0.833 \left(1 + 0.785 \frac{m_e}{T}\right)^{3/2}, \quad (10.169)$$

where $0.833 = \frac{3}{2}\zeta(3)/(\pi^4/45)$ and 0.785 are relevant constants from equations (10.159) and (10.168). Equation (10.169) can be solved for μ_e/T in terms of temperature T , given the present day value electron-to-photon ratio $n_e/n_\gamma|_0$ from equation (10.157), the effective number of degrees of freedom g_s from equation (10.166), and its present day value $g_{s,0} = 2$.

With μ_e/T determined from equation (10.169), comoving number densities $a^3 n$ in terms of temperature T follow from equation (10.160a), along with equations (10.164) and (10.166).

10.28 Maximally symmetric spaces

By construction, the FLRW metric is spatially homogeneous and isotropic, which means it has maximal spatial symmetry. A special subclass of FLRW metrics is in addition stationary, satisfying time translation invariance. As you will show in Exercise 10.24, stationary FLRW metrics may have curvature and a cosmological constant, but no other source. You will also show that a coordinate transformation brings such FLRW metrics to the explicitly stationary form

$$ds^2 = - \left(1 - \frac{1}{3}\Lambda r_s^2\right) dt_s^2 + \frac{dr_s^2}{1 - \frac{1}{3}\Lambda r_s^2} + r_s^2 d\sigma^2, \quad (10.170)$$

where the time t_s and radius r_s are subscripted s for stationary to distinguish them from FLRW time t and radius r . The cosmological constant Λ is related to the mass-energy density ρ_Λ by, equation (2.137),

$$\Lambda = 8\pi G \rho_\Lambda. \quad (10.171)$$

Spacetimes that are homogeneous, isotropic, and stationary, and are therefore described by the metric (10.170), are called **maximally symmetric**. A maximally symmetric space with a positive cosmological constant, $\Lambda > 0$, is called **de Sitter** (dS) space, while that with a negative cosmological constant, $\Lambda < 0$, is called **anti de Sitter** (AdS) space. The maximally symmetric space with zero cosmological constant is just Minkowski space. Thanks to their high degree of symmetry, de Sitter and anti de Sitter spaces play a prominent role in theoretical studies of quantum gravity.

de Sitter space has a horizon at radius

$$r_H = \sqrt{\frac{3}{\Lambda}} = \sqrt{\frac{3}{8\pi G\rho_\Lambda}} . \quad (10.172)$$

Whereas inside the horizon the time coordinate t_s is timelike and the radial coordinate r_s is spacelike, outside the horizon the time coordinate t_s is spacelike and the radial coordinate r_s is timelike.

The Riemann tensor, Ricci tensor, Ricci scalar, and Einstein tensor of maximally symmetric spaces are

$$R_{\kappa\lambda\mu\nu} = \frac{1}{3}\Lambda(g_{\kappa\mu}g_{\lambda\nu} - g_{\kappa\nu}g_{\lambda\mu}) , \quad R_{\kappa\mu} = \Lambda g_{\kappa\mu} , \quad R = 4\Lambda , \quad G_{\kappa\mu} = -\Lambda g_{\kappa\mu} . \quad (10.173)$$

Exercise 10.23 Maximally symmetric spaces in arbitrary dimensions. Generalize equations (10.170)–(10.173) to an arbitrary number N of spacetime dimensions.

Solution. The line-element (10.170) generalizes to

$$ds^2 = - \left(1 - \frac{2\Lambda r_s^2}{(N-1)(N-2)} \right) dt_s^2 + \frac{dr_s^2}{1 - \frac{2\Lambda r_s^2}{(N-1)(N-2)}} + r_s^2 d\omega^2 . \quad (10.174)$$

The horizon radius r_H is

$$r_H = \sqrt{\frac{2}{(N-1)(N-2)\Lambda}} = \sqrt{\frac{2}{(N-1)(N-2)\kappa_N\rho_\Lambda}} . \quad (10.175)$$

The Riemann tensor, Ricci tensor, Ricci scalar, and Einstein tensor are

$$R_{\kappa\lambda\mu\nu} = \frac{2\Lambda}{(N-1)(N-2)}(g_{\kappa\mu}g_{\lambda\nu} - g_{\kappa\nu}g_{\lambda\mu}) , \quad R_{\kappa\mu} = \frac{2\Lambda}{N-2}g_{\kappa\mu} , \quad R = \frac{2N\Lambda}{N-2} , \quad G_{\kappa\mu} = -\Lambda g_{\kappa\mu} . \quad (10.176)$$

10.28.1 de Sitter spacetime as a closed FLRW spacetime

Just as it was possible to conceive the spatial part of the FLRW geometry as a 3D hypersphere embedded in 4D Euclidean space, §10.6, so also it is possible to conceive a maximally symmetric space as a 4D hyperboloid embedded in 5D space.

For de Sitter space, the parent 5D space is a Minkowski space with metric $ds^2 = -du^2 + dx^2 + dy^2 + dz^2 + dw^2$, and the embedded 4D hyperboloid is a set of points

$$-u^2 + x^2 + y^2 + z^2 + w^2 = r_H^2 = \text{constant} , \quad (10.177)$$

with r_H is the horizon radius (10.175). The de Sitter hyperboloid is illustrated in Figure 10.19. Let $r \equiv (x^2 + y^2 + z^2)^{1/2}$, and introduce the boost angle ψ and rotation angle χ defined by

$$u = r_H \sinh \psi , \quad (10.178a)$$

$$r = r_H \cosh \psi \sin \chi , \quad (10.178b)$$

$$w = r_H \cosh \psi \cos \chi . \quad (10.178c)$$

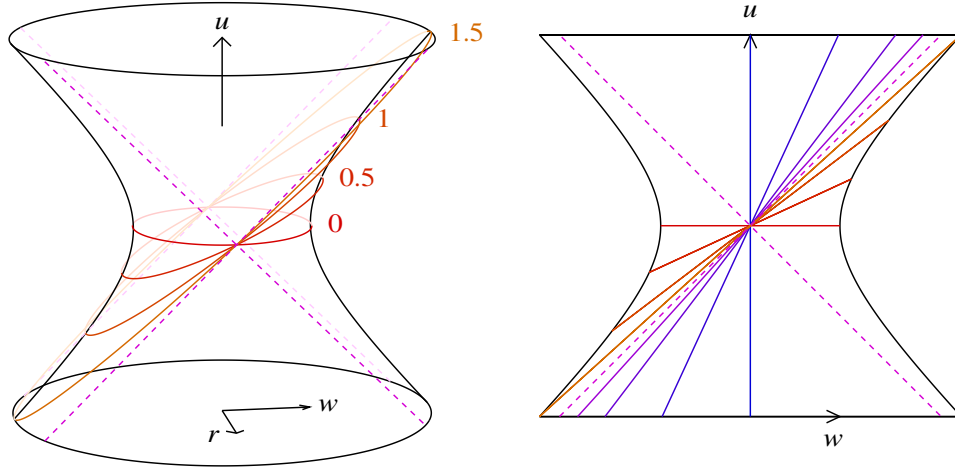


Figure 10.19 Embedding spacetime diagram of de Sitter space, shown on the left in 3D, on the right in a 2D projection on to the u - w plane. Objects are confined to the surface of the embedded hyperboloid. The vertical direction is timelike, while the horizontal directions are spacelike. The position of a non-accelerating observer defines a “north pole” at $r = 0$ and $w > 0$, traced by the (black) line at the right edge of each diagram. Antipodeal to the north pole is a “south pole” at $r = 0$ and $w < 0$, traced by the (black) line at the left edge of each diagram. The (reddish) skewed circles on the 3D diagram, which project to straight lines in the 2D diagram, are lines of constant stationary time t_s , labelled with their value in units of the horizon radius r_H , as measured by the observer at rest at the north pole $r = 0$. Lines of constant stationary time t_s transform into each other under a Lorentz boost in the uw -plane. The 45° dashed lines are null lines constituting the past and future horizons of the north pole observer (or of the south pole observer). The 2D diagram on the right shows in addition a sample of (bluish) timelike geodesics that pass through $u = w = 0$ (these lines are omitted from the 3D diagram).

The radius r defined by equation (10.178b) is the same as the radius r_s in the stationary metric (10.170). In terms of the angles ψ and χ , the metric on the de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 \left[-d\psi^2 + \cosh^2\psi (d\chi^2 + \sin^2\chi d\phi^2) \right]. \quad (10.179)$$

The metric (10.179) is of FLRW form (10.28) with $t = r_H\psi$ and $a(t) = r_H \cosh\psi$ and a closed spatial geometry. The de Sitter space describes a spatially closed FLRW universe that contracts, reaches a minimum size at $t = 0$, then reexpands. Comoving observers, those with $\chi = \text{constant}$ and fixed angular position, move vertically upward on the embedded hyperboloid in Figure 10.19.

The spatial position at $r = 0$ and $w > 0$ defines a “north pole” of de Sitter space. Antipodeal to the north pole is a “south pole” at $r = 0$ and $w < 0$. The surface $u = w$ is a future horizon for an observer at the north pole, and a past horizon for an observer at the south pole. Similarly the surface $u = -w$ is a past horizon for an observer at the north pole, and a future horizon for an observer at the south pole. The causal diamond of any observer is the region of spacetime bounded by the observer’s past and future horizons. The north polar observer’s causal diamond is the region $w > |u|$, while the south polar observer’s causal diamond is the region $w < -|u|$.

The radial coordinate r is spacelike within the causal diamonds of either the north or south polar observers, where $|w| > |u|$, but timelike outside those causal diamonds, where $|w| < |u|$.

The de Sitter hyperboloid possesses a symmetry under Lorentz boosts in the uw -plane. The time t_s in the stationary metric (10.170) is, modulo a factor of r_H , the boost angle of this Lorentz boost, which is

$$t_s = \begin{cases} r_H \operatorname{atanh}(u/w) & |w| > |u| , \\ r_H \operatorname{atanh}(w/u) & |w| < |u| . \end{cases} \quad (10.180)$$

The stationary time coordinate t_s is timelike inside the causal diamonds of either the north or south pole observers, $|w| > |u|$, but spacelike outside those causal diamonds, $|w| < |u|$.

10.28.2 de Sitter spacetime as an open FLRW spacetime

An alternative coordinatization of the same embedded hyperboloid (10.177) for de Sitter space yields a metric in FLRW form but with an open spatial geometry. Let $r \equiv (x^2 + y^2 + z^2)^{1/2}$ as before, and define ψ and χ by

$$u = r_H \sinh \psi \cosh \chi , \quad (10.181a)$$

$$r = r_H \sinh \psi \sinh \chi , \quad (10.181b)$$

$$w = r_H \cosh \psi . \quad (10.181c)$$

The r defined by equation (10.181b) is *not* the same as the r_s in the stationary metric (10.170); rather, it is w that equals r_s . In terms of the angles ψ and χ defined by equations (10.181), the metric on the de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 [-d\psi^2 + \sinh^2 \psi (d\chi^2 + \sinh^2 \chi d\phi^2)] . \quad (10.182)$$

The metric (10.182) is in FLRW form (10.28) with $t = r_H \psi$ and $a(t) = r_H \sinh \psi$ and an open spatial geometry. Whereas the coordinates $\{\psi, \chi\}$, equation (10.178), for de Sitter with closed spatial geometry cover the entire embedded hyperboloid shown in Figure 10.19, the coordinates $\{\psi, \chi\}$, equation (10.181), for de Sitter with open spatial geometry cover only the region of the hyperboloid with $|u| \geq |r|$ and $w \geq r_H$. The region of positive cosmic scale factor, $\psi \geq 0$, corresponds to $u \geq 0$. Conceptually, for de Sitter with open spatial geometry, there is a Big Bang at $\{u, r, w\} = \{0, 0, 1\}r_H$, comoving observers from which fill the region $u \geq |r|$ and $w \geq r_H$. Comoving observers, those with $\chi = \text{constant}$, follow straight lines in the ur -plane, bounded by the null cone at $u = |r|$.

In the open FLRW metric (10.182) for de Sitter space, the coordinates t_s and r_s of the stationary metric (10.170) are respectively spacelike and timelike. Lines of constant stationary time t_s , equation (10.180), coincide with geodesics of comoving observers, at constant χ , while lines of constant stationary radius $r_s = w$, equation (10.181c), coincide with lines of constant FLRW time ψ ,

$$t_s/r_H = \chi , \quad (10.183a)$$

$$r_s/r_H = w/r_H = \cosh \psi . \quad (10.183b)$$

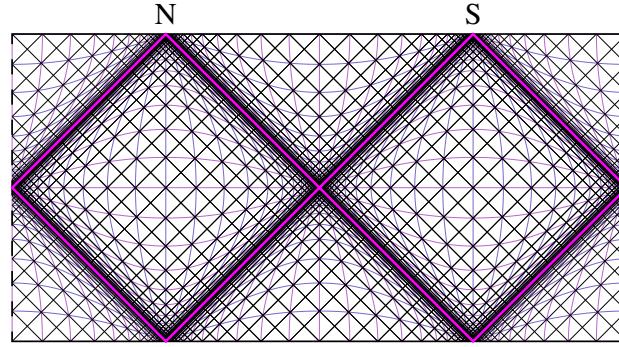


Figure 10.20 Penrose diagram of de Sitter space. The left and right edges are identified. The topology is that of a 3-sphere in the horizontal (spatial) direction times the real line in the vertical (time) direction. The thick (pink) null lines are past and future horizons for observers who follow (vertical) geodesics at the “north” and “south” poles at $r = 0$, marked N and S. The approximately horizontal and vertical contours are contours of constant stationary time t_s and radius r_s in the stationary form (10.170) of the de Sitter metric. The contours are uniformly spaced by 0.4 in t_s/r_H and the tortoise coordinate r_s^*/r_H , equation (10.185). The stationary coordinates t_s and r_s are respectively timelike (vertical) and spacelike (horizontal) inside the causal diamonds of the north and south pole observers, but switch to being respectively spacelike and timelike outside the causal diamonds, in the lower and upper wedges. The lower and upper wedges correspond to the open FLRW version (10.182) of the de Sitter metric. In the lower wedges, comoving observers collapse to a Big Crunch where their future horizons converge, while in the upper wedges, comoving observers expand away from a Big Bang from which their past horizons diverge.

10.28.3 Penrose diagram of de Sitter space

Figure 10.20 shows a Penrose diagram of de Sitter space. A natural choice of Penrose coordinates comes from requiring that vertical lines on the embedded de Sitter hyperboloid 10.19 become vertical lines on the Penrose diagram. These vertical lines are geodesics for comoving observers, lines of constant χ , in the closed FLRW form (10.179) form of the de Sitter metric. The corresponding Penrose time coordinate t_P follows from solving for the radial null geodesics of the metric (10.179), whence $t_P = \int d\psi / \cosh \psi$. The resulting Penrose coordinates for de Sitter space are

$$t_P = \text{atan}(\sinh \psi) = \text{atan}(u/r_H) , \quad (10.184a)$$

$$r_P = \chi = \text{atan}(r/w) . \quad (10.184b)$$

The radial coordinate r in both the closed and open FLRW forms (10.179) and (10.182) of the de Sitter metric was chosen so that a comoving observer at the origin was at $r = 0$, at either the north or the south pole. The Penrose diagram 10.20 depicts both closed and open FLRW geometries, but the open geometry is shifted by 90° to the equator, so that it appears to interleave with the closed geometry instead of overlapping it. The thick (pink) null lines at 45° outline the causal diamonds of north and south polar observers in the closed FLRW geometry. The null lines also outline the causal wedges of equatorial observers in the open FLRW geometry. The lower wedges correspond to collapsing spacetimes that terminate in a Big Crunch where the null lines cross. The upper wedges correspond to expanding spacetimes that begin in a Big Bang

where the null lines cross. Note that the causal diamonds of any non-accelerating observer are spherically symmetric about the observer. Thus the causal diamonds of the closed and open observers touch only along one-dimensional lines, not along three-dimensional hypersurfaces as the Penrose diagram might suggest. The causal diamonds of observers in de Sitter and anti de Sitter spacetimes are different for different observers, and there is no reason to expect that the spacetime could be tiled fully by the causal diamonds of some set of observers.

There is no physical singularity, no divergence of the Riemann tensor, at the Big Crunch and Big Bang points of the collapsing and expanding open FLRW forms of the de Sitter geometry. Does that mean that the collapsing de Sitter spacetime evolves smoothly into an expanding spacetime? As long as the spacetime is pure vacuum, there is no way to tell whether spacetime is expanding or collapsing. Only when the spacetime contains matter of some kind, as our Universe does, can a preferred set of comoving coordinates be defined. When matter is present, Big Crunches and Big Bangs are, setting aside quantum gravity, genuine singularities that cannot be removed by a coordinate transformation.

The horizontal and vertical contours in the Penrose diagram 10.20 are contours of constant stationary time t_s and radius r_s . Translation in t_s is a symmetry of de Sitter spacetime, and to exhibit this symmetry, the contours of t_s in the Penrose diagram are chosen to be uniformly spaced. A similarly symmetric appearance for the radial coordinate is achieved by choosing contours of r_s to be uniformly spaced in the tortoise coordinate r_s^* ,

$$r_s^* \equiv \int \frac{dr_s}{1 - r_s^2/r_H^2} = r_H \operatorname{atanh}(r_s/r_H) . \quad (10.185)$$

The contours in the Penrose diagram 10.20 are uniformly spaced by 0.4 in t_s/r_H and r_s^*/r_H . In terms of the time and tortoise coordinates t_s and r_s^* , the Penrose time and radial coordinates are

$$t_P \pm r_P = \operatorname{atan} \left[\sinh \left(\frac{t_s \pm r_s^*}{r_H} \right) \right] . \quad (10.186)$$

10.28.4 Anti de Sitter space

For anti de Sitter space, the parent 5D space is a Minkowski space with signature $--++++$, metric $ds^2 = -du^2 - dv^2 + dx^2 + dy^2 + dz^2$, and the embedded 4D hyperboloid is a set of points

$$-u^2 - v^2 + x^2 + y^2 + z^2 = -r_H^2 = \text{constant} , \quad (10.187)$$

with $r_H \equiv \sqrt{-3/\Lambda}$. The anti de Sitter hyperboloid is illustrated in Figure 10.21. Let $r \equiv (x^2 + y^2 + z^2)^{1/2}$, and introduce the boost angle χ and rotation angle ψ defined by

$$u = r_H \cosh \chi \cos \psi , \quad (10.188a)$$

$$v = r_H \cosh \chi \sin \psi , \quad (10.188b)$$

$$r = r_H \sinh \chi . \quad (10.188c)$$

The time coordinate ψ defined by equations (10.188) appears to be periodic, with period 2π , but this is an artefact of the embedding. In a causal spacetime, the time coordinate would not loop back on itself. Rather,

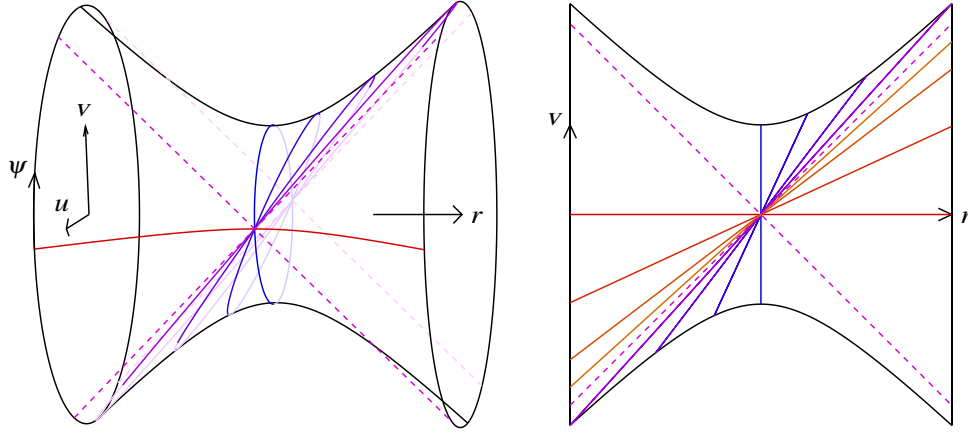


Figure 10.21 Embedding spacetime diagram of anti de Sitter space, shown on the left in 3D, on the right in a 2D projection on to the v - r plane. The vertical direction winding around the hyperboloid is timelike, while the horizontal direction is spacelike. The position of a non-accelerating observer defines a spatial “pole” at $r = 0$. In the 3D diagram, the (red) horizontal line is an example line of constant stationary time t_s for the observer at the pole. Lines of constant stationary time t_s transform into each other under a rotation in the uv -plane. The (bluish) lines at less than 45° from vertical are a sample of geodesics that pass through the pole at $r = 0$ at time $\psi = 0$. In anti de Sitter space, all timelike geodesics that pass through a spatial point boomerang back to the spatial point in a proper time πr_H . The 2D diagram on the right shows in addition (reddish) lines of constant stationary time t_s for observers on the various geodesics.

the coordinate ψ can be taken to increase monotonically as it loops around the hyperboloid, extending from $-\infty$ to ∞ . In terms of the angles ψ and χ , the metric on the anti de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 (-\cosh^2 \chi d\psi^2 + d\chi^2 + \sinh^2 \chi d\phi^2) . \quad (10.189)$$

The metric (10.189) is of stationary form (10.170) with $t_s = r_H \psi$ and $r_s = r_H \sinh \chi$.

10.28.5 Anti de Sitter spacetime as an open FLRW spacetime

An alternative coordinatization of the same embedded hyperboloid 10.21 for anti de Sitter space,

$$u = r_H \cos \psi , \quad (10.190a)$$

$$v = r_H \sin \psi \cosh \chi , \quad (10.190b)$$

$$r = r_H \sin \psi \sinh \chi . \quad (10.190c)$$

yields a metric in FLRW form with an open spatial geometry,

$$ds^2 = r_H^2 [-d\psi^2 + \sin^2 \psi (d\chi^2 + \sinh^2 \chi d\phi^2)] . \quad (10.191)$$

Whereas the coordinates (10.188) cover all of the anti de Sitter hyperboloid 10.21, the open coordinates (10.190) cover only the regions with $|u| \leq r_H$. These are the upper and lower diamonds bounded by the (pink) dashed

null lines in the hyperboloid 10.21. In each diamond, the open spacetime undergoes a Big Bang at the earliest vertex of the diamond, expands to a maximum size, turns around, and collapses to a Big Crunch at the latest vertex of the diamond.

10.28.6 Anti de Sitter spacetime as a Rindler space

Anti de Sitter spacetime possesses symmetry under Lorentz boosts in any time-space plane, such as the vx -plane. In the open FLRW form (10.191) of anti de Sitter geometry, such boosts transform geodesics of comoving observers into each other. Outside the open causal diamonds on the other hand, these boosts generate the worldlines of a certain set of “Rindler” observers who accelerate with constant acceleration in the vx -plane. Rindler time and space coordinates $\{\chi, \psi\}$ are defined by

$$u = r_H \cosh \psi , \quad (10.192a)$$

$$v = r_H \sinh \psi \sinh \chi , \quad (10.192b)$$

$$x = r_H \sinh \psi \cosh \chi , \quad (10.192c)$$

yielding the AdS Rindler metric

$$ds^2 = r_H^2 (-\sinh^2 \psi d\chi^2 + d\psi^2) + dy^2 + dz^2 . \quad (10.193)$$

10.28.7 Penrose diagram of anti de Sitter space

Figure 10.22 shows a Penrose diagram of anti de Sitter space. A natural choice of Penrose coordinates comes from requiring that horizontal lines on the embedded anti de Sitter hyperboloid 10.21 become horizontal lines on the Penrose diagram. These horizontal lines are lines of constant stationary time $t_s = r_H \psi$ in the stationary form (10.189) form of the anti de Sitter metric. The corresponding Penrose radial coordinate r_P follows from solving for the radial null geodesics of the metric (10.189), whence $r_P = \int d\chi / \cosh \chi$. The resulting Penrose coordinates for de Sitter space are

$$t_P = \psi = t_s / r_H , \quad (10.194a)$$

$$r_P = \text{atan}(\sinh \chi) = r_s^* / r_H , \quad (10.194b)$$

where r_s^* is the tortoise radial coordinate

$$r_s^* \equiv \int \frac{dr_s}{1 + r_s^2 / r_H^2} = r_H \text{atan}(r_s / r_H) . \quad (10.195)$$

Thus, for anti de Sitter, lines of constant time t_s and radius r_s in the stationary metric (10.170) correspond also to lines of constant Penrose time and radius t_P and r_P .

The thick (pink) null lines in the Penrose diagram 10.22 outline the causal diamonds of comoving observers in the open FLRW (10.191) form of the anti de Sitter metric. The null lines also outline the causal wedges of Rindler observers in the Rindler (10.193) form of the anti de Sitter metric.

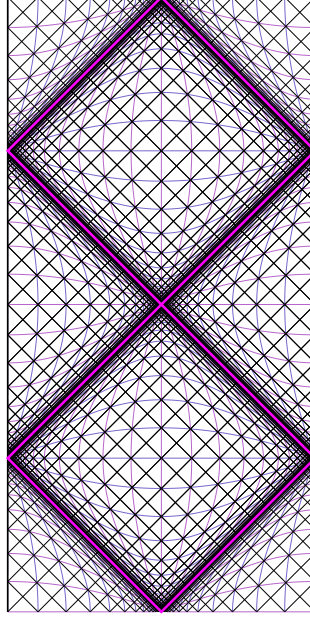


Figure 10.22 Penrose diagram of anti de Sitter space. The diagram repeats vertically indefinitely. The topology is that of Euclidean 3-space in the horizontal (spatial) direction times the real line in the vertical (time) direction. The thick (pink) null lines outline the causal diamonds of observers in the open FLRW form (10.191) of the anti de Sitter spacetime. The spacetime of the open FLRW geometry expands from a Big Bang at a crossing point of the null lines, and collapses to a Big Crunch at the next crossing point. The thick null lines also outline the causal wedges of Rindler observers, equation (10.193), at the left and right edges of the diagram. The approximately horizontal and vertical contours are lines of constant ψ and χ , uniformly spaced by 0.4 in χ and ψ^* , equation (10.196), both in the open FLRW diamonds and in the left and right Rindler wedges, equations (10.191) and (10.193). The coordinates ψ and χ are respectively timelike (vertical) and spacelike (horizontal) in the open diamonds, and respectively spacelike and timelike in the Rindler wedges.

The approximately horizontal and vertical contours in the Penrose diagram 10.22 are lines of constant ψ and χ in both the open FLRW (10.191) and Rindler (10.193) forms of the anti de Sitter metric. In the open FLRW causal diamonds, the horizontal lines are lines of constant cosmic time ψ , while the vertical contours are geodesics, lines of constant χ . In the Rindler causal wedges, the horizontal contours are lines of constant boost angle χ , while the vertical contours are worldlines of Rindler observers, lines of constant ψ .

Anti de Sitter space is symmetric under boosts in the vx -plane, corresponding to translations of the coordinate χ in either of the open FLRW (10.191) or Rindler (10.193) forms of the anti de Sitter metric. The contours in the Penrose diagram 10.22 are uniformly spaced in χ by 0.4 so as to manifest this symmetry. A similarly symmetric appearance for the ψ coordinate is achieved by choosing contours to be uniformly

spaced by 0.4 in the tortoise coordinate ψ^*

$$\psi^* \equiv \begin{cases} \int \frac{d\psi}{\sin \psi} = \ln \tan(\psi/2) & \text{open ,} \\ \int \frac{d\psi}{\sinh \psi} = \ln \tanh(\psi/2) & \text{Rindler .} \end{cases} \quad (10.196)$$

Exercise 10.24 Maximally symmetric spaces.

1. Argue that in a stationary spacetime, every scalar quantity must be independent of time. In particular, the Riemann scalar R , and the contracted Ricci product $R^{\mu\nu}R_{\mu\nu}$ must be independent of time. Conclude that the density ρ and pressure p of a stationary FLRW spacetime must be constant.
2. Conclude that a stationary FLRW spacetime may have curvature and a cosmological constant, but no other source. Show that the FLRW metric then takes the form (10.28) with cosmic scale factor

$$a(t) = \begin{cases} H_0 t & \Omega_k = 1, \Omega_\Lambda = 0, \\ \exp(H_0 t) & \Omega_k = 0, \Omega_\Lambda = 1, \\ \sqrt{-\Omega_k/\Omega_\Lambda} \cosh(\sqrt{\Omega_\Lambda} H_0 t) & \Omega_k < 0, \Omega_\Lambda > 0, \\ \sqrt{\Omega_k/\Omega_\Lambda} \sinh(\sqrt{\Omega_\Lambda} H_0 t) & \Omega_k > 0, \Omega_\Lambda > 0, \\ \sqrt{-\Omega_k/\Omega_\Lambda} \sin(\sqrt{-\Omega_\Lambda} H_0 t) & \Omega_k > 0, \Omega_\Lambda < 0, \end{cases} \quad (10.197)$$

with, in N spacetime dimensions,

$$\Omega_\Lambda H_0^2 = \frac{2\Lambda}{(N-1)(N-2)}, \quad \Omega_k H_0^2 = -\kappa. \quad (10.198)$$

As elsewhere in this chapter, $H = H_0$ at $a = 1$, and the Ω 's sum to unity, $\Omega_k + \Omega_\Lambda = 1$.

3. Show that the FLRW metric transforms into the explicitly stationary form (10.170) under a coordinate transformation to proper radius $r_s = a(t)x$ and stationary time t_s given by

$$t_s = \begin{cases} \sqrt{1 - \kappa x^2} t & \Omega_k = 1, \Omega_\Lambda = 0, \\ t - \frac{1}{H_0} \ln \sqrt{1 - H_0^2 r_s^2} & \Omega_k = 0, \Omega_\Lambda = 1, \\ \frac{1}{\sqrt{\Omega_\Lambda} H_0} \operatorname{acoth} \left[\sqrt{1 - \kappa x^2} \coth \left(\sqrt{\Omega_\Lambda} H_0 t \right) \right] & \Omega_k < 0, \Omega_\Lambda > 0, \\ \frac{1}{\sqrt{\Omega_\Lambda} H_0} \operatorname{atanh} \left[\sqrt{1 - \kappa x^2} \tanh \left(\sqrt{\Omega_\Lambda} H_0 t \right) \right] & \Omega_k > 0, \Omega_\Lambda > 0, \\ \frac{1}{\sqrt{-\Omega_\Lambda} H_0} \operatorname{atan} \left[\sqrt{1 - \kappa x^2} \tan \left(\sqrt{-\Omega_\Lambda} H_0 t \right) \right] & \Omega_k > 0, \Omega_\Lambda < 0. \end{cases} \quad (10.199)$$

Note that in all cases $t_s = t$ at the origin $r_s = 0$.

Concept question 10.25 Milne Universe. In Exercise 10.24 you found that the FLRW metric for an open universe with zero energy-momentum content ($\Omega_k = 1$, $\Omega_\Lambda = 0$), also known as the Milne metric, is equivalent to flat Minkowski space. How can an open universe be equivalent to flat space? Draw a spacetime diagram of Minkowski space showing (a) worldlines of observers at constant comoving FLRW position x , and (b) hypersurfaces of constant FLRW time t .

Concept question 10.26 Stationary FLRW metrics with different curvature constants describe the same spacetime. How can it be that stationary FLRW metrics with different curvature constants κ (but the same cosmological constant Λ) describe the same spacetime?

PART TWO

TETRAD APPROACH TO GENERAL RELATIVITY

Concept Questions

1. The vierbein has 16 degrees of freedom instead of the 10 degrees of freedom of the metric. What do the extra 6 degrees of freedom correspond to?
2. Tetrad transformations are defined to be Lorentz transformations. Don't general coordinate transformations already include Lorentz transformations as a particular case, so aren't tetrad transformations redundant?
3. What does coordinate gauge-invariant mean? What does tetrad gauge-invariant mean?
4. Is the coordinate metric $g_{\mu\nu}$ tetrad gauge-invariant?
5. What does a directed derivative ∂_m mean physically?
6. Is the directed derivative ∂_m coordinate gauge-invariant?
7. Is the tetrad metric γ_{mn} coordinate gauge-invariant? Is it tetrad gauge-invariant?
8. What is the tetrad-frame 4-velocity u^m of a person at rest in an orthonormal tetrad frame?
9. If the tetrad frame is accelerating (not in free-fall), which of the following is true/false?
 - a. Does the tetrad-frame 4-velocity u^m of a person continuously at rest in the tetrad frame change with time? $\partial_0 u^m = 0$? $D_0 u^m = 0$?
 - b. Do the tetrad axes γ_m change with time? $\partial_0 \gamma_m = 0$? $D_0 \gamma_m = 0$?
 - c. Does the tetrad metric γ_{mn} change with time? $\partial_0 \gamma_{mn} = 0$? $D_0 \gamma_{mn} = 0$?
 - d. Do the covariant components u_m of the 4-velocity of a person continuously at rest in the tetrad frame change with time? $\partial_0 u_m = 0$? $D_0 u_m = 0$?
10. Suppose that $\mathbf{p} = \gamma_m p^m$ is a 4-vector. Is the proper rate of change of the proper components p^m measured by an observer equal to the directed time derivative $\partial_0 p^m$ or to the covariant time derivative $D_0 p^m$? What about the covariant components p_m of the 4-vector? [Hint: The proper contravariant components of the 4-vector measured by an observer are $p^m \equiv \gamma^m \cdot \mathbf{p}$ where γ^m are the contravariant locally inertial rest axes of the observer. Similarly the proper covariant components are $p_m \equiv \gamma_m \cdot \mathbf{p}$.]
11. A person with two eyes separated by proper distance $\delta\xi^n$ observes an object. The observer observes the photon 4-vector from the object to be p^m . The observer uses the difference δp^m in the two 4-vectors detected by the two eyes to infer the binocular distance to the object. Is the difference δp^m in photon 4-vectors detected by the two eyes equal to the directed derivative $\delta\xi^n \partial_n p^m$ or to the covariant derivative $\delta\xi^n D_n p^m$?

12. Suppose that p^m is a tetrad 4-vector. Parallel-transport the 4-vector by an infinitesimal proper distance $\delta\xi^n$. Is the change in p^m measured by an ensemble of observers at rest in the tetrad frame equal to the directed derivative $\delta\xi^n\partial_n p^m$ or to the covariant derivative $\delta\xi^n D_n p^m$? [Hint: What if “rest” means that the observer at each point is separately at rest in the tetrad frame at that point? What if “rest” means that the observers are mutually at rest relative to each other in the rest frame of the tetrad at one particular point?]
13. What is the physical significance of the fact that directed derivatives fail to commute?
14. Physically, what do the tetrad connection coefficients Γ_{kmn} mean?
15. What is the physical significance of the fact that Γ_{kmn} is antisymmetric in its first two indices (if the tetrad metric γ_{mn} is constant)?
16. Are the tetrad connections Γ_{kmn} coordinate gauge-invariant?

What's important?

This chapter describes the tetrad formalism of general relativity.

1. Why tetrads? Because physics is clearer in a locally inertial frame than in a coordinate frame.
2. The primitive object in the tetrad formalism is the vierbein $e^m{}_\mu$, in place of the metric in the coordinate formalism.
3. Written suitably, for example as equation (11.9), a metric ds^2 encodes not only the metric coefficients $g_{\mu\nu}$, but a full vierbein $e^m{}_\mu$, through $ds^2 = \gamma_{mn} e^m{}_\mu dx^\mu e^n{}_\nu dx^\nu$.
4. The tetrad road from vierbein to energy-momentum is similar to the coordinate road from metric to energy-momentum, albeit a little more complicated.
5. In the tetrad formalism, the directed derivative ∂_m is the analogue of the coordinate partial derivative $\partial/\partial x^\mu$ of the coordinate formalism. Directed derivatives ∂_m do not commute, whereas coordinate derivatives $\partial/\partial x^\mu$ do commute.

The tetrad formalism

11.1 Tetrad

A **tetrad** (greek foursome) $\gamma_m(x)$ is a set of axes

$$\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\} \quad (11.1)$$

attached to each point x^μ of spacetime. The common case, illustrated in Figure 11.1, is that of an **orthonormal tetrad**, where the axes form a locally inertial frame at each point, so that the dot products of the axes constitute the Minkowski metric η_{mn}

$$\gamma_m \cdot \gamma_n = \eta_{mn} \quad (11.2)$$

However, other tetrads prove useful in appropriate circumstances. There are spin tetrads, null tetrads (notably the Newman-Penrose double null tetrad), and others (indeed, the basis of coordinate tangent vectors e_μ is

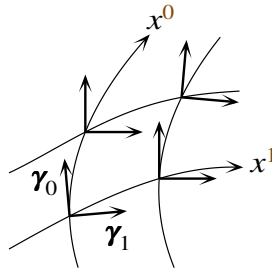


Figure 11.1 Tetrad vectors γ_m form a basis of vectors at each point. A common choice, depicted here, is for the basis vectors γ_m to form an orthonormal set, meaning that their dot products constitute the Minkowski metric, $\gamma_m \cdot \gamma_n = \eta_{mn}$, at each point. The orthonormal frames at neighbouring points need not be aligned with each other by parallel transport, and indeed in curved spacetime it is impossible to choose orthonormal frames that are everywhere aligned.

a tetrad). In general, the tetrad metric is some symmetric matrix γ_{mn}

$$\boxed{\boldsymbol{\gamma}_m \cdot \boldsymbol{\gamma}_n \equiv \gamma_{mn}} . \quad (11.3)$$

The convention in this book is that latin (black) indices label tetrad frames, while greek (brown) indices label coordinate frames.

Why introduce tetrads?

1. The physics is more transparent when expressed in a locally inertial frame (or some other frame adapted to the physics), as opposed to the coordinate frame, where Salvador Dali rules.
2. If you want to consider spin- $\frac{1}{2}$ particles and quantum physics, you better work with tetrads.
3. For good reason, much of the general relativistic literature works with tetrads, so it's useful to understand them.

11.2 Vierbein

The **vierbein** (German four-legs, or colloquially, critter) $e^m{}_\mu$ is defined to be the matrix that transforms between the tetrad frame and the coordinate frame (note the placement of indices: the tetrad index m comes first, then the coordinate index μ)

$$\boxed{\boldsymbol{e}_\mu = e^m{}_\mu \boldsymbol{\gamma}_m} . \quad (11.4)$$

The letter e stems from the German word *einheit* for unity. The vierbein is a 4×4 matrix, with 16 independent components. The inverse vierbein $e_m{}^\mu$ is defined to be the matrix inverse of the vierbein $e^m{}_\mu$, so that

$$e_m{}^\mu e^m{}_\nu = \delta_\nu^\mu , \quad e_m{}^\mu e^n{}_\mu = \delta_m^n . \quad (11.5)$$

Thus equation (11.4) inverts to

$$\boxed{\boldsymbol{\gamma}_m = e_m{}^\mu \boldsymbol{e}_\mu} . \quad (11.6)$$

11.3 The line-element encodes the vierbein

The scalar spacetime distance is

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu = \boldsymbol{e}_\mu \cdot \boldsymbol{e}_\nu dx^\mu dx^\nu = \gamma_{mn} e^m{}_\mu e^n{}_\nu dx^\mu dx^\nu \quad (11.7)$$

from which it follows that the coordinate metric $g_{\mu\nu}$ is

$$\boxed{g_{\mu\nu} = \gamma_{mn} e^m{}_\mu e^n{}_\nu} . \quad (11.8)$$

The shorthand way in which line-elements are commonly written encodes not only a metric but also a vierbein, hence a tetrad. For example, the Schwarzschild line-element

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (11.9)$$

takes the form (11.7) with an orthonormal (Minkowski) tetrad metric $\gamma_{mn} = \eta_{mn}$, and a vierbein encoded in the differentials (one-forms, §15.6)

$$e^0_{\mu} dx^{\mu} = \left(1 - \frac{2M}{r}\right)^{1/2} dt, \quad (11.10a)$$

$$e^1_{\mu} dx^{\mu} = \left(1 - \frac{2M}{r}\right)^{-1/2} dr, \quad (11.10b)$$

$$e^2_{\mu} dx^{\mu} = r d\theta, \quad (11.10c)$$

$$e^3_{\mu} dx^{\mu} = r \sin \theta d\phi, \quad (11.10d)$$

Explicitly, the vierbein of the Schwarzschild line-element is the diagonal matrix

$$e^m_{\mu} = \begin{pmatrix} (1 - 2M/r)^{1/2} & 0 & 0 & 0 \\ 0 & (1 - 2M/r)^{-1/2} & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & 0 & 0 & r \sin \theta \end{pmatrix}, \quad (11.11)$$

and the corresponding inverse vierbein is (note that, because the tetrad index is always in the first place and the coordinate index is always in the second place, the matrices as written are actually inverse transposes of each other, not just inverses)

$$e_m^{\mu} = \begin{pmatrix} (1 - 2M/r)^{-1/2} & 0 & 0 & 0 \\ 0 & (1 - 2M/r)^{1/2} & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1/(r \sin \theta) \end{pmatrix}. \quad (11.12)$$

Concept question 11.1 Schwarzschild vierbein. The components e^0_t and e^1_r of the Schwarzschild vierbein (11.11) are imaginary inside the horizon. What does this mean? Is the vierbein still valid inside the horizon?

11.4 Tetrad transformations

Tetrad transformations are transformations that preserve the fundamental property of interest, for example the orthonormality, of the tetrad. For most tetrads considered in this book, which includes not only orthonormal tetrads, but also spin tetrads and null tetrads (but not coordinate-based tetrads), tetrad transformations are Lorentz transformations. The Lorentz transformation may be, and usually is, a different transformation at each point. Tetrad transformations rotate the tetrad axes γ_k at each point by a Lorentz transformation L_k^m , while keeping the background coordinates x^{μ} unchanged:

$$\boxed{\gamma_k \rightarrow \gamma'_k = L_k^m \gamma_m}. \quad (11.13)$$

In the case that the tetrad axes γ_k are orthonormal, with a Minkowski metric, the Lorentz transformation matrices L_k^m in equation (11.13) take the familiar special relativistic form, but the linear matrices L_k^m in equation (11.13) signify a Lorentz transformation in any case.

For orthonormal, spin, and null tetrads, the tetrad metric γ_{mn} is constant. Lorentz transformations are precisely those transformations that leave the tetrad metric unchanged

$$\gamma'_{kl} = \gamma'_k \cdot \gamma'_l = L_k^m L_l^n \gamma_m \cdot \gamma_n = L_k^m L_l^n \gamma_{mn} = \gamma_{kl} . \quad (11.14)$$

Exercise 11.2 Generators of Lorentz transformations are antisymmetric. From the condition that the tetrad metric γ_{kl} is unchanged by a Lorentz transformation, show that the generator of an infinitesimal Lorentz transformation is an antisymmetric matrix. Is this true only for an orthonormal tetrad, or is it true more generally?

Solution. An infinitesimal Lorentz transformation is the sum of the unit matrix and an infinitesimal piece ΔL_k^m , the generator of the infinitesimal Lorentz transformation,

$$L_k^m = \delta_k^m + \Delta L_k^m . \quad (11.15)$$

Under such an infinitesimal Lorentz transformation, the tetrad metric transforms to

$$\gamma'_{kl} = (\delta_k^m + \Delta L_k^m)(\delta_l^n + \Delta L_l^n) \gamma_{mn} \approx \gamma_{kl} + \Delta L_{kl} + \Delta L_{lk} , \quad (11.16)$$

which by proposition equals the original tetrad metric γ_{kl} , equation (11.14). It follows that

$$\Delta L_{kl} + \Delta L_{lk} = 0 , \quad (11.17)$$

that is, the generator ΔL_{kl} is antisymmetric, as claimed. The result is true whenever the tetrad metric is invariant under Lorentz transformations.

11.5 Tetrad vectors and tensors

Just as coordinate vectors (and tensors) were defined in §2.8 as objects that transformed like (tensor products of) coordinate intervals under coordinate transformations, so also tetrad vectors (and tensors) are defined as objects that transform like (tensor products of) tetrad vectors under tetrad (Lorentz) transformations.

11.5.1 Covariant tetrad 4-vector

A tetrad (Lorentz) transformation transforms the tetrad axes γ_k in accordance with equation (11.13). A covariant **tetrad 4-vector** is defined to be a quantity $A_k = \{A_0, A_1, A_2, A_3\}$ that transforms under a tetrad transformation like the tetrad axes,

$$A_k \rightarrow A'_k = L_k^m A_m . \quad (11.18)$$

11.5.2 Lowering and raising tetrad indices

Just as the indices on a coordinate vector or tensor were lowered and raised with the coordinate metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$, §2.8.3, so also indices on a tetrad vector or tensor are lowered and raised with the tetrad metric γ_{mn} and its inverse γ^{mn} , defined to satisfy

$$\gamma_{km}\gamma^{mn} = \delta_k^n . \quad (11.19)$$

In the tetrads considered in this book (Minkowski, spin, or Newman-Penrose tetrad), the components of the tetrad metric and its inverse are numerically equal, $\gamma_{mn} = \gamma^{mn}$, but this need not be the case in general.

The contravariant (raised index) components A^m and covariant (lowered index) components A_m of a tetrad vector are related by

$$A^m = \gamma^{mn}A_n , \quad A_m = \gamma_{mn}A^n . \quad (11.20)$$

The dual tetrad basis vectors $\boldsymbol{\gamma}^m$ are defined by

$$\boldsymbol{\gamma}^m \equiv \gamma^{mn}\boldsymbol{\gamma}_n . \quad (11.21)$$

By construction, dot products of the dual and tetrad basis vectors equal the unit matrix,

$$\boldsymbol{\gamma}^m \cdot \boldsymbol{\gamma}_n = \delta_n^m , \quad (11.22)$$

while dot products of the dual basis vectors with each other equal the inverse tetrad metric,

$$\boldsymbol{\gamma}^m \cdot \boldsymbol{\gamma}^n = \gamma^{mn} . \quad (11.23)$$

11.5.3 Contravariant tetrad vector

A contravariant tetrad 4-vector A^k transforms under a tetrad transformation as, analogously to equation (11.18),

$$A^k \rightarrow A'^k = L^k_m A^m , \quad (11.24)$$

where L^k_m is the Lorentz transformation inverse to L_k^m . Equation (11.14) implies that Lorentz transformation matrices with indices variously lowered and raised satisfy

$$L_k^m L_m^l = L_{km} L^{lm} = L_{mk} L^{ml} = L^m_k L_m^l = \delta_k^l . \quad (11.25)$$

11.5.4 Abstract vector

A 4-vector can be written in a coordinate- and tetrad- independent fashion as an abstract 4-vector $\mathbf{A}$,

$$\mathbf{A} = \boldsymbol{\gamma}_m A^m = \mathbf{e}_\mu A^\mu . \quad (11.26)$$

Although $\mathbf{A}$ is a 4-vector, it is by construction unchanged by either a coordinate transformation or a tetrad transformation, and is therefore, according to the naming convention adopted in this book, §11.6, both a

coordinate scalar and a tetrad scalar. The coordinate and tetrad components of the 4-vector $\mathbf{A}$ are related by the vierbein,

$$A_\mu = e^m{}_\mu A_m, \quad A_m = e_m{}^\mu A_\mu. \quad (11.27)$$

11.5.5 Scalar product

The scalar product of two 4-vectors may be $\mathbf{A}$ and $\mathbf{B}$ may be written variously

$$\mathbf{A} \cdot \mathbf{B} = A_m B^m = A_\mu B^\mu. \quad (11.28)$$

The scalar product is a scalar, unchanged by either a coordinate or tetrad transformation.

11.5.6 Tetrad tensor

In general, a **tetrad-frame tensor** $A^{kl\dots}_{mn\dots}$ is an object that transforms under tetrad (Lorentz) transformations (11.13) as

$$A'^{kl\dots}_{mn\dots} = L^k{}_a L^l{}_b \dots L_m{}^c L_n{}^d \dots A^{ab\dots}_{cd\dots}. \quad (11.29)$$

11.6 Index and naming conventions for vectors and tensors

In the tetrad formalism tensors can be coordinate tensors, or tetrad tensors, or mixed coordinate-tetrad tensors. For example, the vierbein $e^m{}_\mu$ is itself a mixed coordinate-tetrad tensor.

The convention in this book is to distinguish the various kinds of vector and tensor with an adjective, and by its index:

1. A **coordinate vector** A^μ , with a **brown** greek index, is one that changes in a prescribed way under coordinate transformations. A coordinate transformation is one that changes the coordinates x^μ of the spacetime without actually changing the spacetime or whatever lies in it. A coordinate vector A^μ does not change under a tetrad transformation, and is therefore a tetrad scalar.
2. A **tetrad vector** A^m with a black latin index, is one that changes in a prescribed way under tetrad transformations. A tetrad transformation Lorentz transforms the tetrad axes γ_m at each point of the spacetime without actually changing the spacetime or whatever lies in it. A tetrad vector A^m does not change under a coordinate transformation, and is therefore a coordinate scalar.
3. An **abstract vector** $\mathbf{A}$, identified by boldface, is the thing itself, and is unchanged by either the choice of coordinates or the choice of tetrad. Since the abstract vector is unchanged by either a coordinate transformation or a tetrad transformation, it is a coordinate and tetrad scalar, and has no indices.

All the types of vector have the properties of linearity (additivity, multiplication by scalars) that identify them mathematically as belonging to vector spaces. The important distinction between the types of vector is how they behave under transformations.

Just because something has a coordinate or tetrad index does not make it a coordinate or tetrad tensor. If however an object is a coordinate and/or tetrad tensor, then its indices are lowered and raised as follows:

1. Lower and raise coordinate indices with the coordinate metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$;
2. Lower and raise tetrad indices with the tetrad metric γ_{mn} and its inverse γ^{mn} ;
3. Switch between coordinate and tetrad frames with the vierbein $e^m{}_\mu$ and its inverse $e_m{}^\mu$.

11.7 Gauge transformations

Gauge transformations are transformations of the coordinates or tetrad. Such transformations do not change the underlying spacetime.

Quantities that are unchanged by a coordinate transformation are **coordinate gauge-invariant** (coordinate scalars). Quantities that are unchanged under a tetrad transformation are **tetrad gauge-invariant** (tetrad scalars). For example, tetrad tensors are coordinate gauge-invariant, while coordinate tensors are tetrad gauge-invariant.

Tetrad transformations have the 6 degrees of freedom of Lorentz transformations, with 3 degrees of freedom in spatial rotations, and 3 more in Lorentz boosts. General coordinate transformations have 4 degrees of freedom. Thus there are 10 degrees of freedom in the choice of tetrad and coordinate system. The 16 degrees of freedom of the vierbein, minus the 10 degrees of freedom from the transformations of the tetrad and coordinates, leave 6 physical degrees of freedom in spacetime, the same as in the coordinate approach to general relativity, which is as it should be.

11.8 Directed derivatives

Directed derivatives ∂_m are defined to be the directional derivatives along the axes γ_m

$$\partial_m \equiv \gamma_m \cdot \partial = \gamma_m \cdot e^\mu \frac{\partial}{\partial x^\mu} = e_m{}^\mu \frac{\partial}{\partial x^\mu} \quad \text{a tetrad 4-vector .} \quad (11.30)$$

The directed derivative ∂_m is independent of the choice of coordinates, as signalled by the fact that it has only a tetrad index, no coordinate index.

Unlike coordinate derivatives $\partial/\partial x^\mu$, directed derivatives ∂_m do not commute. Their commutator is

$$\begin{aligned} [\partial_m, \partial_n] &= \left[e_m{}^\mu \frac{\partial}{\partial x^\mu}, e_n{}^\nu \frac{\partial}{\partial x^\nu} \right] \\ &= e_m{}^\mu \frac{\partial e_n{}^\nu}{\partial x^\mu} \frac{\partial}{\partial x^\nu} - e_n{}^\nu \frac{\partial e_m{}^\mu}{\partial x^\nu} \frac{\partial}{\partial x^\mu} \\ &= (-d_{nm}^k + d_{mn}^k) \partial_k \quad \text{not a tetrad tensor} \end{aligned} \quad (11.31)$$

where $d_{lmn} \equiv \gamma_{lk} d_{mn}^k$ is the inverse vierbein derivative

$$d_{lmn} \equiv -\gamma_{lk} e^k{}_{\kappa} e_n{}^{\nu} \frac{\partial e_m{}^{\kappa}}{\partial x^{\nu}} \quad \text{not a tetrad tensor} . \quad (11.32)$$

Since the vierbein and inverse vierbein are inverse to each other, an equivalent definition of d_{lmn} in terms of the vierbein is

$$\boxed{d_{lmn} \equiv \gamma_{lk} e_m{}^{\mu} e_n{}^{\nu} \frac{\partial e^k{}_{\mu}}{\partial x^{\nu}}} \quad \text{not a tetrad tensor} . \quad (11.33)$$

The vierbein derivatives d_{lmn} are also known as Ricci rotation coefficients (or, in the context of Newman-Penrose tetrads, spin coefficients).

11.9 Tetrad covariant derivative

The derivation of tetrad covariant derivatives D_m follows precisely the analogous derivation of coordinate covariant derivatives D_{μ} . The tetrad-frame formulae look entirely similar to the coordinate-frame formulae, with the replacement of coordinate partial derivatives by directed derivatives, $\partial/\partial x^{\mu} \rightarrow \partial_m$, and the replacement of coordinate-frame connections by tetrad-frame connections $\Gamma_{\mu\nu}^{\kappa} \rightarrow \Gamma_{mn}^k$. There are two things to be careful about: first, unlike coordinate partial derivatives, directed derivatives ∂_m do not commute; and second, neither tetrad-frame nor coordinate-frame connections are tensors, and therefore it should be no surprise that the tetrad-frame connections Γ_{lmn} are *not* related to the coordinate-frame connections $\Gamma_{\lambda\mu\nu}$ by the ‘usual’ vierbein transformations. Rather, the tetrad and coordinate connections are related by equation (11.44).

If Φ is a scalar, then $\partial_m \Phi$ is a tetrad 4-vector. The tetrad covariant derivative of a scalar is just the directed derivative

$$\boxed{D_m \Phi = \partial_m \Phi} \quad \text{a tetrad 4-vector} . \quad (11.34)$$

If A^m is a tetrad 4-vector, then $\partial_n A^m$ is *not* a tetrad tensor, and $\partial_n A_m$ is *not* a tetrad tensor. But the abstract 4-vector $\mathbf{A} = \gamma_m A^m$, being by construction invariant under both tetrad and coordinate transformations, is a scalar, and its directed derivative is therefore a 4-vector,

$$\begin{aligned} \partial_n \mathbf{A} &= \partial_n (\gamma_m A^m) \quad \text{a tetrad 4-vector} \\ &= \gamma_m \partial_n A^m + (\partial_n \gamma_m) A^m . \end{aligned} \quad (11.35)$$

For equation (11.35) to make sense, the derivatives $\partial_n \gamma_m$ must be defined, something that is made possible, as in the coordinate approach in §2.9.2, by the postulate of the existence of locally inertial frames. The coordinate partial derivative of γ_m are defined in the usual way by

$$\frac{\partial \gamma_m}{\partial x^{\nu}} \equiv \lim_{\delta x^{\nu} \rightarrow 0} \frac{\gamma_m(x^0, \dots, x^{\nu} + \delta x^{\nu}, \dots, x^3) - \gamma_m(x^0, \dots, x^{\nu}, \dots, x^3)}{\delta x^{\nu}} . \quad (11.36)$$

The right hand of equation (11.36) involves the difference between γ_m at two different points x and $x + \delta x$.

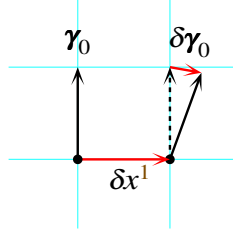


Figure 11.2 The change $\delta\gamma_0$ in the tetrad vector γ_0 over a small coordinate interval δx^1 of spacetime is defined to be the difference between the tetrad vector $\gamma_0(x^1 + \delta x^1)$ at the shifted position $x^1 + \delta x^1$ and the tetrad vector $\gamma_0(x^1)$ at the original position x^1 , parallel-transported to the shifted position. The parallel-transported vector is shown as a dashed arrowed line. The parallel transport is defined with respect to a locally inertial frame, shown as a background square grid aligned with the tetrad at the unshifted position.

The difference is to be interpreted as $\gamma_m(x+\delta x)$ at the shifted point, minus the value of $\gamma_m(x^\nu)$ parallel-transported from position x to the shifted point $x+\delta x$ along the small distance δx between them, as illustrated in Figure 11.2. Parallel transport means, go to a locally inertial frame, then move along the prescribed direction without boosting or precessing. With the coordinate partial derivatives of the tetrad basis vectors so defined, the directed derivatives follow as $\partial_n \gamma_m = e_n{}^\nu \partial \gamma_m / \partial x^\nu$.

The directed derivatives of the tetrad basis vectors define the **tetrad-frame connection coefficients**, Γ_{mn}^k ,

$$\boxed{\partial_n \gamma_m \equiv \Gamma_{mn}^k \gamma_k} \quad \text{not a tetrad tensor .} \quad (11.37)$$

In the usual case where the tetrad metric is Lorentz invariant and the tetrad connections Γ_{kmn} are therefore generators of Lorentz transformations, antisymmetric in their first two indices, Exercise 11.2, I like to call the tetrad connection coefficients **Lorentz connections**. With equation (11.37), equation (11.35) then shows that

$$\partial_n \mathbf{A} = \gamma_k (D_n A^k) \quad \text{a tetrad tensor ,} \quad (11.38)$$

where $D_n A^k$ is the covariant derivative of the contravariant 4-vector A^k

$$\boxed{D_n A^k \equiv \partial_n A^k + \Gamma_{mn}^k A^m} \quad \text{a tetrad tensor .} \quad (11.39)$$

The covariant derivative of a covariant tetrad 4-vector A_k follows similarly from

$$\partial_n \mathbf{A} = \gamma^k (D_n A_k) \quad \text{a tetrad tensor ,} \quad (11.40)$$

where $D_n A_k$ is the covariant derivative of the covariant 4-vector A_k

$$\boxed{D_n A_k \equiv \partial_n A_k - \Gamma_{kn}^m A_m} \quad \text{a tetrad tensor .} \quad (11.41)$$

In general, the covariant derivative of a tetrad-frame tensor is

$$\boxed{D_p A_{mn\dots}^{kl\dots} = \partial_p A_{mn\dots}^{kl\dots} + \Gamma_{qp}^k A_{mn\dots}^{ql\dots} + \Gamma_{qp}^l A_{mn\dots}^{kq\dots} + \dots - \Gamma_{mp}^q A_{qn\dots}^{kl\dots} - \Gamma_{np}^q A_{mq\dots}^{kl\dots} - \dots} \quad (11.42)$$

with a positive Γ term for each contravariant index, and a negative Γ term for each covariant index.

Concept question 11.3 Does a locally inertial frame remain locally inertial under parallel transport? If a locally inertial frame is parallel-transported from one spacetime point to another, is the parallel-transported frame still locally inertial? **Answer.** Yes. This is essentially what the postulate of the existence of locally inertial frames means: at any point there exists a frame, a set of tetrad axes, that are orthonormal at that point, and that remain orthonormal in an infinitesimal neighbourhood of the point. In particular, there exists a frame that remains orthonormal when shifted an infinitesimal distance along any one direction, and that frame is said to be parallel-transported along that direction. In standard general relativity, the postulate of the existence of locally internal frames includes the postulate that torsion vanishes, which may or may not be true in reality, Concept question 2.5. Parallel transport remains well defined even if torsion does not vanish, Concept question 2.4.

11.10 Relation between tetrad and coordinate connections

The relation between the tetrad connections Γ_{mn}^k and their coordinate counterparts $\Gamma_{\mu\nu}^\kappa$ follows from

$$\begin{aligned} \frac{\partial e_\mu}{\partial x^\nu} &= \Gamma_{\mu\nu}^\kappa e_\kappa = \frac{\partial e^m_\mu \gamma_m}{\partial x^\nu} \quad \text{not a tetrad tensor} \\ &= \frac{\partial e^m_\mu}{\partial x^\nu} \gamma_m + e^m_\mu \frac{\partial \gamma_m}{\partial x^\nu} \\ &= e^m_\mu e^n_\nu (d_{mn}^k + \Gamma_{mn}^k) \gamma_k . \end{aligned} \quad (11.43)$$

Thus the relation is

$$\boxed{d_{lmn} + \Gamma_{lmn} = e_l^\lambda e_m^\mu e_n^\nu \Gamma_{\lambda\mu\nu}} \quad \text{not a tetrad tensor} \quad (11.44)$$

where

$$\Gamma_{lmn} \equiv \gamma_{lk} \Gamma_{mn}^k . \quad (11.45)$$

11.11 Antisymmetry of the tetrad connections

The directed derivative of the tetrad metric is

$$\begin{aligned} \partial_n \gamma_{lm} &= \partial_n (\gamma_l \cdot \gamma_m) \\ &= \gamma_l \cdot \partial_n \gamma_m + \gamma_m \cdot \partial_n \gamma_l \\ &= \Gamma_{lmn} + \Gamma_{mln} . \end{aligned} \quad (11.46)$$

In most cases of interest, including orthonormal, spin, and null tetrads, the tetrad metric is chosen to be a constant. For example, if the tetrad is orthonormal, then the tetrad metric is the Minkowski metric, which

is constant, the same everywhere. If the tetrad metric is constant, then all derivatives of the tetrad metric vanish, and then equation (11.46) shows that the tetrad connections are antisymmetric in their first two indices

$$\boxed{\Gamma_{lmn} = -\Gamma_{mln}} . \quad (11.47)$$

This antisymmetry reflects the fact that Γ_{lmn} is the generator of a Lorentz transformation for each n , Exercise 11.2.

11.12 Torsion tensor

The **torsion tensor** S_{kl}^m , which general relativity assumes to vanish, is defined in the usual way, equation (2.57), by the commutator of the covariant derivative acting on a scalar Φ

$$\boxed{[D_k, D_l] \Phi = S_{kl}^m \partial_m \Phi} \quad \text{a tetrad tensor} . \quad (11.48)$$

The expression (11.41) for the covariant derivatives coupled with the commutator (11.31) of directed derivatives shows that the torsion tensor is

$$\boxed{S_{kl}^m = d_{kl}^m + \Gamma_{kl}^m - d_{lk}^m - \Gamma_{lk}^m} \quad \text{a tetrad tensor} , \quad (11.49)$$

which is equivalent to the coordinate expression (2.58) for the torsion in view of the relation (11.44) between tetrad and coordinate connections. The torsion tensor S_{kl}^m is antisymmetric in $k \leftrightarrow l$, as is evident from its definition (11.48).

11.13 No-torsion condition

General relativity assumes vanishing torsion

$$\boxed{S_{kl}^m = 0} . \quad (11.50)$$

For vanishing torsion, equation (11.49) implies

$$d_{mkl} + \Gamma_{mkl} = d_{mlk} + \Gamma_{mlk} \quad \text{not a tetrad tensor} , \quad (11.51)$$

which is equivalent to the usual symmetry condition $\Gamma_{\lambda\kappa\mu} = \Gamma_{\lambda\mu\kappa}$ on the coordinate frame connections in view of the relation (11.44) between tetrad and coordinate connections.

11.14 Tetrad connections in terms of the vierbein

In the general case of non-constant tetrad metric, and non-vanishing torsion, the following manipulation, from equations (11.46) and (11.49), analogous to the corresponding manipulation (2.61) in the coordinate

frame,

$$\begin{aligned}\partial_n \gamma_{lm} + \partial_m \gamma_{ln} - \partial_l \gamma_{mn} &= \Gamma_{lmn} + \Gamma_{mln} + \Gamma_{lnm} + \Gamma_{nlm} - \Gamma_{mnl} - \Gamma_{nml} \\ &= 2\Gamma_{lmn} + S_{lmn} + S_{mln} + S_{nlm} - d_{lnm} + d_{lmn} - d_{mln} + d_{mnl} - d_{nlm} + d_{nml}\end{aligned}\quad (11.52)$$

implies that the tetrad connections Γ_{lmn} are given in terms of the derivatives $\partial_n \gamma_{lm}$ of the tetrad metric, the torsion S_{lmn} , and the vierbein derivatives d_{lmn} by

$$\begin{aligned}\Gamma_{lmn} &= \frac{1}{2} (\partial_n \gamma_{lm} + \partial_m \gamma_{ln} - \partial_l \gamma_{mn} + S_{lmn} + S_{mnl} + S_{nml} \\ &\quad + d_{lnm} - d_{lmn} + d_{mln} - d_{mnl} + d_{nlm} - d_{nml}) \quad \text{not a tetrad tensor} .\end{aligned}\quad (11.53)$$

If torsion vanishes, as general relativity assumes, and if furthermore the tetrad metric is constant, then equation (11.53) simplifies to the following expression for the tetrad connections in terms of the vierbein derivatives d_{lmn} defined by (11.33), analogous to the expression (2.63) for coordinate-frame connections in terms of coordinate derivatives of the metric,

$$\boxed{\Gamma_{lmn} = \frac{1}{2} (d_{lnm} - d_{lmn} + d_{mln} - d_{mnl} + d_{nlm} - d_{nml})} \quad \text{not a tetrad tensor} . \quad (11.54)$$

This is the formula that allows tetrad connections to be calculated from the vierbein.

11.15 Torsion-free covariant derivative

As in §2.12, the torsion-free part of the covariant derivative is a covariant derivative even when torsion is present. When torsion is present and it is desirable to make the torsion part explicit, it is convenient to distinguish torsion-free quantities with a $\circ$ superscript. The torsion-full tetrad connection Γ_{lmn} is a sum of the torsion-free (Levi-Civita) connection $\mathring{\Gamma}_{lmn}$ and the contortion tensor K_{lmn} ,

$$\Gamma_{lmn} = \mathring{\Gamma}_{lmn} + K_{lmn} , \quad (11.55)$$

where from equation (11.53) the contortion tensor K_{lmn} and the torsion tensor S_{lmn} are related by (see equations (16.131) for a decomposition of contortion and torsion into irreducible parts)

$$K_{lmn} = \frac{1}{2} (S_{lmn} + S_{mnl} + S_{nml}) = -S_{nlm} + \frac{3}{2} S_{[lmn]} \quad \text{a tetrad tensor} , \quad (11.56a)$$

$$S_{lmn} = K_{lmn} - K_{lnm} = -K_{mnl} + 3K_{[lmn]} \quad \text{a tetrad tensor} . \quad (11.56b)$$

Like the tetrad connection Γ_{lmn} , the contortion K_{lmn} is antisymmetric in its first two indices. The torsion-full covariant derivative D_n differs from the torsion-free covariant derivative $\mathring{D}_n$ by the contortion,

$$D_n A^k \equiv \mathring{D}_n A^k + K_{mn}^k A^m \quad \text{a tetrad tensor} . \quad (11.57)$$

In this book the symbol D_n by default denotes the torsion-full covariant derivative. In some places however, such as in the theory of differential forms, the symbol D_n is used for brevity to denote the torsion-free covariant derivative, even in the presence of torsion. When D_n denotes the torsion-free covariant derivative, it will be stated so explicitly.

11.16 Riemann curvature tensor

The **Riemann curvature tensor** R_{klmn} is defined in the usual way, equation (2.110), by the commutator of the covariant derivative acting on a 4-vector. In the presence of torsion,

$$[D_k, D_l] A_m \equiv S_{kl}^n D_n A_m + R_{klmn} A^n \quad \text{a tetrad tensor} . \quad (11.58)$$

If torsion vanishes, as general relativity assumes, then the definition (11.58) reduces to

$$\boxed{[D_k, D_l] A_m \equiv R_{klmn} A^n} \quad \text{a tetrad tensor} . \quad (11.59)$$

The expression (11.41) for the covariant derivative coupled with the torsion equation (11.48) yields the following formula for the tetrad-frame Riemann tensor in terms of tetrad connection, for the general case of non-vanishing torsion:

$$R_{klmn} = \partial_k \Gamma_{mnl} - \partial_l \Gamma_{mnk} + \Gamma_{ml}^p \Gamma_{pnk} - \Gamma_{mk}^p \Gamma_{pnl} + (\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p) \Gamma_{mnp} \quad \text{a tetrad tensor} . \quad (11.60)$$

The formula has extra terms $(\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p) \Gamma_{mnp}$ compared to the formula (2.112) for the coordinate-frame Riemann tensor $R_{\kappa\lambda\mu\nu}$. The expression (11.56b) for torsion in terms of the contortion shows that the factor $\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p$ can be written in terms of torsion-free connections $\overset{\circ}{\Gamma}_{kl}^p$, or in terms of the vierbein derivatives d_{kl}^p , as

$$\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p = \overset{\circ}{\Gamma}_{kl}^p - \overset{\circ}{\Gamma}_{lk}^p = d_{kl}^p - d_{lk}^p . \quad (11.61)$$

If torsion vanishes, as general relativity assumes, then

$$\boxed{R_{klmn} = \partial_k \Gamma_{mnl} - \partial_l \Gamma_{mnk} + \Gamma_{ml}^p \Gamma_{pnk} - \Gamma_{mk}^p \Gamma_{pnl} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \Gamma_{mnp}} \quad \text{a tetrad tensor} . \quad (11.62)$$

The symmetries of the tetrad-frame Riemann tensor are the same as those of the coordinate-frame Riemann tensor, §2.19.4. For vanishing torsion, these are

$$R_{klmn} = R_{([kl][mn])} , \quad (11.63a)$$

$$R_{k[lmn]} = 0 . \quad (11.63b)$$

Exercise 11.4 Riemann tensor. From the definition (11.58), derive the expression (11.60) for the Riemann tensor. [Hint: Start by expanding out the definition (11.58) using the definition (11.42) of the covariant derivative. You will find it easier to derive an expression for the Riemann tensor with one index raised, such as $R_{klm}{}^n$, but you should resist the temptation to leave it there, because the symmetries of the Riemann tensor are obscured when one index is raised. To switch to all lowered indices, you will need to convert terms such as $\partial_k \Gamma_{ml}^n$ by

$$\partial_k \Gamma_{ml}^n = \partial_k (\gamma^{np} \Gamma_{pml}) = \gamma^{np} \partial_k \Gamma_{pml} + \Gamma_{pml} \partial_k \gamma^{np} . \quad (11.64)$$

You should show that the directed derivative $\partial_k \gamma^{np}$ in this expression is related to tetrad connections through a formula similar to equation (11.46),

$$\partial_k \gamma^{np} = -\Gamma^{np}{}_k - \Gamma^{pn}{}_k , \quad (11.65)$$

which you should recognize as equivalent to $D_k \gamma^{np} = 0$.]

Solution.

$$\begin{aligned}
[D_k, D_l] A_m &= (\partial_k (D_l A_m) - \Gamma_{lk}^n D_n A_m - \Gamma_{mk}^n D_l A_n) - (k \leftrightarrow l) \\
&= (\partial_k (\partial_l A_m - \Gamma_{ml}^n A_n) + \Gamma_{kl}^n D_n A_m + \Gamma_{ml}^n D_k A_n) - (k \leftrightarrow l) \\
&= (\partial_k \partial_l A_m - (\partial_k \Gamma_{ml}^n) A_n - \Gamma_{ml}^n \partial_k A_n + \Gamma_{kl}^n D_n A_m + \Gamma_{ml}^n D_k A_n) - (k \leftrightarrow l) \\
&= (d_{kl}^n - d_{lk}^n) \partial_n A_m + (\Gamma_{kl}^n - \Gamma_{lk}^n) D_n A_m - (\partial_k \Gamma_{pml} - \Gamma_{ml}^q (\Gamma_{pqk} + \Gamma_{qpk}) - (k \leftrightarrow l)) A^p \\
&\quad - (\Gamma_{ml}^n \Gamma_{pnk} - (k \leftrightarrow l)) A^p \\
&= (d_{kl}^n - d_{lk}^n + \Gamma_{kl}^n - \Gamma_{lk}^n) D_n A_m + (d_{kl}^n - d_{lk}^n) \Gamma_{pmn} A^p - (\partial_k \Gamma_{pml} - \Gamma_{ml}^q \Gamma_{qpk} - (k \leftrightarrow l)) A^p \\
&= S_{kl}^n D_n A_m + (\partial_k \Gamma_{nml} - \Gamma_{ml}^p \Gamma_{pnk} + d_{kl}^p \Gamma_{nmp} - (k \leftrightarrow l)) A^n \\
&= S_{kl}^n D_n A_m + R_{klmn} A^n .
\end{aligned} \tag{11.66}$$

Exercise 11.5 Antisymmetry of the Riemann tensor. Argue that the antisymmetry of R_{klmn} in mn , with or without torsion, can be deduced from

$$0 = [D_k, D_l] \gamma_{mn} = S_{kl}^p D_p \gamma_{mn} + R_{klmp} \delta_n^p + R_{klnp} \delta_m^p = R_{klmn} + R_{klnm} . \tag{11.67}$$

Exercise 11.6 Cyclic symmetry of the Riemann tensor. Show that the cyclic symmetry (11.63b) is a consequence of the assumption of vanishing torsion.

Solution. Use the Jacobi identity applied to a scalar, $[D_k, [D_m, D_l]] \Phi = 0$. Show that if Φ is a scalar, then

$$\begin{aligned}
2D_{[k} D_l D_m] \Phi &= [D_{[k}, D_l] D_m] \Phi = (R_{[klm]}^n - S_{[kl}^p S_{m]p}^n) D_n \Phi + S_{[kl}^n D_m] D_n \Phi \\
&= D_{[k} [D_l, D_m] \Phi = (D_{[k} S_{lm]}^n) D_n \Phi + S_{[kl}^n D_m] D_n \Phi .
\end{aligned} \tag{11.68}$$

Consequently

$$R_{[klm]}^n = D_{[k} S_{lm]}^n + S_{[kl}^p S_{m]p}^n . \tag{11.69}$$

An equivalent expression in terms of the torsion-free covariant derivative $\overset{\circ}{D}_k$ and the contortion K_{mnl} is

$$R_{[klm]}^n = \overset{\circ}{D}_{[k} S_{lm]}^n + K_{p[kl}^n S_{m]}^p . \tag{11.70}$$

Exercise 11.7 Symmetry of the Riemann tensor. Show that the cyclic symmetry (11.63b) implies the symmetry $kl \leftrightarrow mn$, given the antisymmetries $k \leftrightarrow l$ and $m \leftrightarrow n$. Given Exercise 11.6, this shows that the symmetry $kl \leftrightarrow mn$ is, like the cyclic symmetry, a consequence of vanishing torsion.

Solution. Show that

$$2(R_{klmn} - R_{mnkl}) = 3(R_{k[lmn]} - R_{l[kmn]} - R_{m[nkl]} + R_{n[mkl]}) , \tag{11.71}$$

or alternatively,

$$2(R_{klmn} - R_{mnkl}) = 3(R_{[klm]n} - R_{[kln]m} - R_{[mnk]l} + R_{[mnl]k}) . \tag{11.72}$$

Exercise 11.8 Number of components of the Riemann tensor. How many independent components does the Riemann tensor have, in 4-dimensional spacetime?

Solution. If torsion vanishes, 20. If torsion does not vanish, 36. The extra 16 components come from $R_{[klm]n}$, which is related to torsion by equation (11.69), and which has $4 \times 4 = 16$ components if torsion does not vanish.

Concept question 11.9 Must connections vanish if Riemann vanishes? Must the tetrad connections Γ_{lmn} vanish if the Riemann tensor vanishes identically, $R_{klmn} = 0$? **Answer.** No. For a counterexample, take flat (Minkowski) space expressed in spherical polar coordinates $\{t, r, \theta, \phi\}$. The non-vanishing tetrad-frame connections are $\Gamma_{212} = \Gamma_{313} = 1/r$ and $\Gamma_{323} = \cot \theta / r$ (compare equations (20.23)).

11.16.1 Riemann tensor in a mixed coordinate-tetrad frame

In Chapter 16, Einstein's equations will be obtained from an action principle, as first done by Hilbert (1915). The Hilbert Lagrangian takes a particularly insightful form if the Riemann tensor is expressed in a mixed coordinate-tetrad basis.

The coordinate-frame covariant derivative D_{κ} of a tetrad-frame vector a_n is

$$D_{\kappa} a_n = e^k_{\kappa} D_k a_n = \frac{\partial a_n}{\partial x^{\kappa}} - \Gamma_{n\kappa}^m a_m \quad \text{a coordinate-tetrad tensor} , \quad (11.73)$$

where $\Gamma_{n\kappa}^m$ is the tetrad-frame connection with its last index converted into the coordinate frame with the vierbein,

$$\Gamma_{n\kappa}^m \equiv e^k_{\kappa} \Gamma_{nk}^m \quad \text{a coordinate vector, but not a tetrad tensor} . \quad (11.74)$$

As usual, the connection with all indices lowered is defined by $\Gamma_{ln\kappa} \equiv \gamma_{lm} \Gamma_{n\kappa}^m$. The connections $\Gamma_{mn\kappa}$ should not be confused with the coordinate-frame connections (Christoffel symbols) $\Gamma_{\mu\nu\kappa}$. The relation between the two is, from equation (11.44),

$$\Gamma_{mn\kappa} = -e^k_{\kappa} d_{mnk} + e_m^{\mu} e_n^{\nu} \Gamma_{\mu\nu\kappa} . \quad (11.75)$$

In 4 dimensions there are $6 \times 4 = 24$ distinct connections $\Gamma_{mn\kappa}$ (with or without torsion), whereas there are $4 \times 10 = 40$ distinct coordinate-frame connections $\Gamma_{\mu\nu\kappa}$ (without torsion, or $4 \times 4 \times 4 = 64$ with torsion).

The last term on the right hand side of equation (11.60) for the Riemann tensor can be written, in view of equations (11.49) and (11.32),

$$(\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p) \Gamma_{mnp} = (\partial_l e_k^{\kappa} - \partial_k e_l^{\kappa}) \Gamma_{mn\kappa} . \quad (11.76)$$

The Riemann tensor $R_{\kappa\lambda mn}$ in the mixed coordinate-tetrad basis is then

$$\boxed{R_{\kappa\lambda mn} = \frac{\partial \Gamma_{mn\lambda}}{\partial x^{\kappa}} - \frac{\partial \Gamma_{mn\kappa}}{\partial x^{\lambda}} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} - \Gamma_{m\kappa}^p \Gamma_{pn\lambda}} \quad \text{a coordinate-tetrad tensor} , \quad (11.77)$$

which is valid with or without torsion. Equation (11.77) resembles superficially the coordinate-frame expression (2.112) for the Riemann tensor, but it is more economical in that there are only 24 connections $\Gamma_{mn\kappa}$ instead of the 40 (or 64, with torsion) coordinate-frame connections $\Gamma_{\mu\nu\kappa}$.

The torsion $S_{\kappa\lambda}^m$ in the mixed coordinate-tetrad basis is

$$S_{\kappa\lambda}^m = -\frac{\partial e^m_{\lambda}}{\partial x^{\kappa}} + \frac{\partial e^m_{\kappa}}{\partial x^{\lambda}} - \Gamma_{l\kappa}^m e^l_{\lambda} + \Gamma_{k\lambda}^m e^k_{\kappa} \quad \text{a coordinate-tetrad tensor .} \quad (11.78)$$

Equations (11.77) and (11.78) constitute Cartan's equations of structure (Cartan, 1904) (see §16.14.2).

11.17 Ricci, Einstein, Bianchi

The usual suite of formulae leading to Einstein's equations apply. Since all the quantities are tensors, and all the equations are tensor equations, their form follows immediately from their coordinate counterparts.

Ricci tensor:

$$R_{km} \equiv \gamma^{ln} R_{klmn} . \quad (11.79)$$

Ricci scalar:

$$R \equiv \gamma^{km} R_{km} . \quad (11.80)$$

Einstein tensor:

$$G_{km} \equiv R_{km} - \frac{1}{2} \gamma_{km} R . \quad (11.81)$$

Einstein's equations:

$$G_{km} = 8\pi G T_{km} . \quad (11.82)$$

The trace of the Einstein equations implies that $R = -8\pi G T$, so the Einstein equations (11.82) can equally well be written with the trace terms transferred from the left to the right hand side,

$$R_{km} = 8\pi G \left(T_{km} - \frac{1}{2} \gamma_{km} T \right) . \quad (11.83)$$

Bianchi identities in the absence of torsion:

$$D_k R_{lmnp} + D_l R_{mknp} + D_m R_{klnp} = 0 , \quad (11.84)$$

which most importantly imply covariant conservation of the Einstein tensor, hence conservation of energy-momentum

$$D^k T_{km} = 0 . \quad (11.85)$$

11.18 Expressions with torsion

If torsion does not vanish, then the Riemann tensor, and consequently also the Ricci and Einstein tensors, can be split into torsion-free (distinguished by a $^{\circ}$ overscript) and torsion parts (e.g. Hehl, Heyde, and Kerlick 1976). A similar split occurs in the ADM formalism where a certain gauge choice (fixing the time component

γ_0 of the tetrad to be orthogonal to hypersurfaces of constant time) splits the tetrad connection into a tensor part, the extrinsic curvature, and a remainder, equation (17.27).

The contortion tensor K_{lmn} was defined previously as the torsion part of the connection Γ_{lmn} , equation (11.55). The unique non-vanishing contraction of the contortion tensor defines the contortion vector K_m ,

$$K_m \equiv K_{mn}^n = S_{mn}^n . \quad (11.86)$$

The torsion-full Riemann tensor R_{klmn} is a sum of the torsion-free Riemann tensor $\mathring{R}_{klmn}$ and a torsion part (note that $K_{pkl} - K_{plk} - S_{pkl} = 0$, so the “extra” term in R_{klmn} , equation (11.60), vanishes when K_{pkl} is the contortion),

$$R_{klmn} = \mathring{R}_{klmn} + \mathring{D}_k K_{mnl} - \mathring{D}_l K_{mnk} + K_{ml}^p K_{pnk} - K_{mk}^p K_{pnl} . \quad (11.87)$$

The Ricci tensor is

$$R_{km} = \mathring{R}_{km} - \mathring{D}_k K_m - \mathring{D}^n K_{mnk} + K_{mpn} K^{np}_k + K_{pmk} K^p , \quad (11.88)$$

and the Ricci scalar is

$$R = \mathring{R} - 2\mathring{D}_n K^n + K_{mpn} K^{npm} - K_n K^n . \quad (11.89)$$

The antisymmetric part of the Einstein tensor is, from contracting equation (11.69),

$$G_{[km]} = R_{[km]} = \frac{3}{2} R_{[klm]}^l = \frac{3}{2} \left(D_{[k} S_{l]m}^l + S_{[kl}^p S_{m]p}^l \right) , \quad (11.90)$$

which vanishes for vanishing torsion.

The Jacobi identity (2.126) implies, in addition to the 16 conditions (11.69), the 24 Bianchi identities

$$D_{[k} R_{lm]np} + S_{[kl}^q R_{m]qnp} = 0 . \quad (11.91)$$

The doubly-contracted Bianchi identities are

$$-\frac{3}{2} \gamma^{kn} \gamma^{lp} \left(D_{[k} R_{lm]np} + S_{[kl}^q R_{m]qnp} \right) = D^k G_{mk} - \frac{1}{2} S_{kl}^q R_{mq}{}^{kl} - S_{km}^q R_q{}^k = 0 . \quad (11.92)$$

11.19 General relativity in two spacetime dimensions

General relativity in 2 spacetime dimensions is weird. There are zero Bianchi identities (2.128) in 2 spacetime dimensions, so the Bianchi identities do not identify any covariantly conserved tensor. The Einstein tensor itself vanishes identically in 2 spacetime dimensions.

There are consistent extensions of general relativity in 2 spacetime dimensions, such as string-inspired dilaton gravity (Grumiller, Kummer, and Vassilevich, 2002). However, those will not be considered here.

Historically, the main application of 2-dimensional relativity has been to explore quantum field theory in curved spacetime, since in 2 spacetime dimensions the quantum energy-momentum tensor induced by any prescribed geometry can be calculated exactly (even though the classical energy-momentum tensor is indeterminate).

The closest thing to a consistent realisation of classical general relativity in 2 spacetime dimensions is as follows.

In 2 spacetime dimensions, the Riemann tensor has just one distinct component, R_{0101} , and that component is determined entirely by the Ricci scalar R . The tetrad-frame Riemann and Ricci tensors are related to the Ricci scalar R by

$$R_{klmn} = \frac{1}{2} (\gamma_{km}\gamma_{ln} - \gamma_{kn}\gamma_{lm}) R, \quad R_{km} = \frac{1}{2} \gamma_{km} R. \quad (11.93)$$

In an arbitrary number of N spacetime dimensions, contracting the Einstein equations implies that the Ricci scalar R is proportional to the trace T of the energy-momentum tensor,

$$(1 - \frac{1}{2}N)R = \kappa_N T, \quad (11.94)$$

where κ_N is Newton's gravitational constant in N spacetime dimensions, suitably normalized. For $N = 2$, the factor on the left of equation (11.94) vanishes; but one can imagine absorbing the zero factor into a redefinition of the gravitational constant κ_N , so that

$$R = \kappa'_2 T \quad (11.95)$$

for some κ'_2 . Now impose that the energy-momentum tensor T_{km} is covariantly conserved,

$$D^k T_{km} = 0. \quad (11.96)$$

In $N = 2$ spacetime dimensions, the trace relation (11.95) together with the covariant conservation condition (11.96) imply almost uniquely the form of the energy-momentum tensor.

The conserved energy-momentum tensor T_{km} takes its simplest expression when the metric is expressed in conformally flat form. The metric in $N = 2$ spacetime dimensions is a symmetric 2×2 matrix. By a suitable coordinate transformation of the 2 coordinates, the metric can be brought to the conformally flat form

$$ds^2 = e^{2\xi} (-dt^2 + dx^2) = -e^{2\xi} dv du, \quad (11.97)$$

where $v \equiv t + x$ and $u \equiv t - x$ are null coordinates, and ξ is a function of the two coordinates. The Newman-Penrose tetrad-frame components of the conserved energy-momentum tensor T_{km} are then

$$-\frac{R}{2} = -4e^{-2\xi} \frac{\partial^2 \xi}{\partial v \partial u} = -\kappa'_2 \frac{T}{2} = \kappa'_2 T_{vu}, \quad (11.98a)$$

$$4e^{-2\xi} \left[\frac{\partial^2 \xi}{\partial v^2} - \left(\frac{\partial \xi}{\partial v} \right)^2 + f^+(v) \right] = \kappa'_2 T_{vv}, \quad (11.98b)$$

$$4e^{-2\xi} \left[\frac{\partial^2 \xi}{\partial u^2} - \left(\frac{\partial \xi}{\partial u} \right)^2 + f^-(u) \right] = \kappa'_2 T_{uu}, \quad (11.98c)$$

where $f^+(v)$ and $f^-(u)$ are arbitrary functions of respectively v and u . There is a residual gauge freedom $v \rightarrow V(v)$ and $u \rightarrow U(u)$ in the choice of null coordinates that allows the conformal function to be adjusted $\xi \rightarrow \xi + \xi^+(v) + \xi^-(u)$ by arbitrary additive functions of v and u . This residual gauge freedom allows the

functions $f^+(v)$ and $f^-(u)$ in equations (11.98b) and (11.98c) to be adjusted arbitrarily. If desired, $f^+(v)$ and $f^-(u)$ can be set to zero.

The classical 2-dimensional general relativity described by equations (11.98) is not very interesting; for example there is no 2-dimensional analogue of the Schwarzschild black hole, Exercise 11.10.

Where equations (11.98) prove more interesting is that they also describe the expectation value $\langle T_{kl} \rangle$ of the renormalized quantum energy-momentum induced by a given geometry in 2 spacetime dimensions (Birrell and Davies, 1982). That is, the expectation value $\langle T \rangle$ of the quantum trace in 2 spacetime dimensions is proportional to the Ricci scalar (Birrell and Davies, 1982, eq. (6.121)), and the quantum energy-momentum tensor $\langle T_{kl} \rangle$ is covariantly conserved, therefore equations (11.98) are satisfied by $\langle T_{kl} \rangle$. In 4 spacetime dimensions the quantum energy-momentum tensor $\langle T_{kl} \rangle$ is extremely difficult to calculate in a general spacetime, so clues from 2 spacetime dimensions can be illuminating.

Exercise 11.10 Black holes in 2 spacetime dimensions? Does the analogue of a Schwarzschild black hole exist in 2 spacetime dimensions?

Solution. No. Require the spacetime to be empty outside some radius. The vanishing of the Ricci scalar (11.98a) implies that

$$\xi = \xi^+(v) + \xi^-(u) \quad (11.99)$$

for some functions ξ^+ and ξ^- of the null coordinates v and u . But then the coordinate transformations of the null coordinates

$$dV = e^{2\xi^+(v)} dv, \quad dU = e^{2\xi^-(u)} du \quad (11.100)$$

bring the line-element to

$$ds^2 = -dVdU, \quad (11.101)$$

which is just flat (Minkowski) space in $N = 2$ dimensions.

Exercise 11.11 Tidal forces falling into a Schwarzschild black hole. In the Schwarzschild or Gullstrand-Painlevé orthonormal tetrad, or indeed in any orthonormal tetrad of the Schwarzschild geometry where t and r represent the time and radial directions and θ and ϕ represent the transverse (angular) directions, the non-zero components of the tetrad-frame Riemann tensor are

$$\frac{1}{2}R_{trtr} = -R_{t\theta t\theta} = -R_{t\phi t\phi} = R_{r\theta r\theta} = R_{r\phi r\phi} = -\frac{1}{2}R_{\theta\phi\theta\phi} = C, \quad (11.102)$$

where

$$C = -M/r^3 \quad (11.103)$$

is the Weyl scalar (the spin-0 component of the Weyl tensor).

1. **Tidal forces.** A person at rest in the tetrad has, by definition, tetrad-frame 4-velocity $u^m = \{1, 0, 0, 0\}$.

From the equation of geodesic deviation, equation (3.20),

$$\frac{D^2 \delta \xi_m}{D\tau^2} + R_{klmn} \delta \xi^k u^l u^n = 0, \quad (11.104)$$

deduce the tidal acceleration on the person in the radial and transverse directions. Does the tidal acceleration stretch or compress? [Hint: The equation of geodesic deviation, §3.3, gives the proper acceleration between two points a small distance $\delta\xi^m$ apart, where ξ^m are the locally inertial coordinates of the tetrad frame. Notice that this problem is much easier to solve with tetrads than with the traditional coordinate approach. Note also that since the Weyl tensor takes the same form (11.102) independent of the radial boost, the tidal acceleration is the same regardless of the radial velocity of the infaller.]

2. **Choose a black hole to fall into.** What is the mass of the black hole for which the tidal acceleration M/r^3 is 1 gee per metre at the horizon? If you wanted to fall through the horizon of a black hole without first being torn apart, what mass of black hole would you choose? [Hint: 1 gee is the gravitational acceleration at the surface of the Earth.]
3. **Time to die.** In a previous problem you showed that the proper time to free-fall radially from radius r to the singularity of a Schwarzschild black hole, for a faller who starts at zero velocity at infinity (so $E = 1$), is

$$\tau = \frac{2}{3} \sqrt{\frac{r^3}{2M}} . \quad (11.105)$$

How long, in seconds, does it take to fall to the singularity from the place where the tidal acceleration is 1 gee per metre? Comment?

4. **Tear-apart radius.** At what radius r , in km, do you start to get torn apart, if that happens when the tidal acceleration is 1 gee per metre? Express your answer in terms of the black hole mass M in units of a solar mass $M_\odot$, that is, in the form $r = ? (M/M_\odot)^?$.
5. **Spaghettified?** In Exercise 7.6 you showed that the infall velocity of a person who free-falls radially from zero velocity at infinity (so $E = 1$) is

$$\frac{dr}{d\tau} = -\sqrt{\frac{2M}{r}} . \quad (11.106)$$

Show that radial component ($\delta\xi^r$) of the equation of geodesic deviation (11.104) for such a person solves to

$$\delta\xi^r = \frac{A}{\sqrt{r}} + Br^2 , \quad (11.107)$$

where A and B are constants. If a person tears apart when the tidal acceleration is 1 gee per metre, and the parts of the person free-fall thereafter, is the person actually spaghettified? [Hint: If the frame is in free-fall, then the covariant derivatives $D/D\tau$ in the equation of geodesic deviation may be replaced by ordinary derivatives $d/d\tau$ in that frame. The last part of the question — Is the person actually spaghettified? — is a concept question: given the solution (11.107), can you interpret what it means?]

Exercise 11.12 Totally antisymmetric tensor.

1. In an orthonormal tetrad γ_m where γ_0 points to the future and $\gamma_1, \gamma_2, \gamma_3$ are right-handed, the contravariant totally antisymmetric tensor ε^{klmn} is defined by (this is the opposite sign from the Misner, Thorne, and Wheeler (1973) notation)

$$\varepsilon^{klmn} \equiv [klmn] , \quad (11.108)$$

where $[klmn]$ is the totally antisymmetric symbol

$$[klmn] \equiv \begin{cases} +1 & \text{if } klmn \text{ is an even permutation of } 0123, \\ -1 & \text{if } klmn \text{ is an odd permutation of } 0123, \\ 0 & \text{if } klmn \text{ are not all different.} \end{cases} \quad (11.109)$$

The choice of $+$ sign in the definition (11.108) of ε^{klmn} is determined by the definition (13.19) of the pseudoscalar I_N of the geometric algebra in N dimensions as a product of all N basis vectors, equation (15.74); see Concept question 15.3. The corresponding covariant totally antisymmetric tensor ε_{klmn} is

$$\varepsilon_{klmn} = -[klmn], \quad (11.110)$$

in which the $-$ sign is the determinant of the tetrad (Minkowski) metric. Argue that in a general basis e_μ the contravariant totally antisymmetric tensor $\varepsilon^{\kappa\lambda\mu\nu}$ is

$$\varepsilon^{\kappa\lambda\mu\nu} = e_k^\kappa e_l^\lambda e_m^\mu e_n^\nu \varepsilon^{klmn} = e^{-1} [\kappa\lambda\mu\nu], \quad (11.111)$$

while its covariant counterpart is

$$\varepsilon_{\kappa\lambda\mu\nu} = -e [\kappa\lambda\mu\nu], \quad (11.112)$$

where $e \equiv |e^m_\mu|$ is the determinant of the vierbein.

2. Show that in 4 dimensions

$$\varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -4! e^{[k}_\kappa e^l_\lambda e^m_\mu e^{n]}_\nu. \quad (11.113)$$

Conclude that

$$e_k^\kappa \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -6 e^{[l}_\lambda e^m_\mu e^{n]}_\nu, \quad (11.114a)$$

$$e_k^\kappa e_l^\lambda \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -4 e^{[m}_\mu e^{n]}_\nu, \quad (11.114b)$$

$$e_k^\kappa e_l^\lambda e_m^\mu \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -6 e^n_\nu, \quad (11.114c)$$

$$e_k^\kappa e_l^\lambda e_m^\mu e_n^\nu \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -24. \quad (11.114d)$$

The coefficient of the p th contraction is $-p!(4-p)!$.

Spin and Newman-Penrose tetrads

THIS CHAPTER NEEDS REWRITING.

This Chapter discusses spin tetrads (§??) and Newman-Penrose tetrads (§12.2). The Chapter goes on to show how the fields that describe electromagnetic (§??) and gravitational (§12.3) waves have a natural and insightful complex structure that is brought out in a Newman-Penrose tetrad. The Newman-Penrose formalism provides a natural context for the Petrov classification of the Weyl tensor (§12.4).

12.1 Spin tetrad formalism

In quantum mechanics, fundamental particles have spin. The 3 generations of leptons (electrons, muons, taus, and their respective neutrino partners) and quarks (up, charm, top, and their down, strange, and bottom partners) have spin $\frac{1}{2}$ (in units $\hbar = 1$). The carrier particles of the electromagnetic force (photons), the weak force (the $W^\pm$ and Z bosons), and the colour force (the 8 gluons), have spin 1. The carrier of the gravitational force, the graviton, is expected to have spin 2, though as of 2010 no gravitational wave, let alone its quantum, the graviton, has been detected.

General relativity is a classical, not quantum, theory. Nevertheless the spin properties of classical waves, such as electromagnetic or gravitational waves, are already apparent classically.

12.1.1 Spin tetrad

A systematic way to project objects into spin components is to work in a spin tetrad. As will become apparent below, equation (12.5), spin describes how an object transforms under rotation about some preferred axis. In the case of an electromagnetic or gravitational wave, the natural preferred axis is the direction of propagation of the wave. With respect to the direction of propagation, electromagnetic waves prove to have two possible spins, or helicities, ± 1 , while gravitational waves have two possible spins, or helicities, ± 2 . A preferred axis might also be set by an experimenter who chooses to measure spin along some particular direction. The following treatment takes the preferred direction to lie along the z -axis γ_z , but there is no loss of generality in making this choice.

Start with an orthonormal tetrad $\{\gamma_t, \gamma_x, \gamma_y, \gamma_z\}$. If the preferred tetrad axis is the z -axis γ_z , then the spin tetrad axes $\{\gamma_+, \gamma_-\}$ are defined to be complex combinations of the transverse axes $\{\gamma_x, \gamma_y\}$,

$$\gamma_+ \equiv \frac{1}{\sqrt{2}}(\gamma_x + i\gamma_y) , \quad (12.1a)$$

$$\gamma_- \equiv \frac{1}{\sqrt{2}}(\gamma_x - i\gamma_y) . \quad (12.1b)$$

The tetrad metric of the spin tetrad $\{\gamma_t, \gamma_z, \gamma_+, \gamma_-\}$ is

$$\gamma_{mn} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} . \quad (12.2)$$

Notice that the spin axes $\{\gamma_+, \gamma_-\}$ are themselves null, $\gamma_+ \cdot \gamma_+ = \gamma_- \cdot \gamma_- = 0$, whereas their scalar product with each other is non-zero $\gamma_+ \cdot \gamma_- = 1$. The null character of the spin axes is what makes spin especially well-suited to describing fields, such as electromagnetism and gravity, that propagate at the speed of light. An even better trick in dealing with fields that propagate at the speed of light is to work in a Newman-Penrose tetrad, §12.2, in which all 4 tetrad axes are taken to be null.

12.1.2 Transformation of spin under rotation about the preferred axis

Under a right-handed rotation by angle χ about the preferred axis γ_z , the transverse axes γ_x and γ_y transform as

$$\begin{aligned} \gamma_x &\rightarrow \cos \chi \gamma_x + \sin \chi \gamma_y , \\ \gamma_y &\rightarrow \sin \chi \gamma_x - \cos \chi \gamma_y . \end{aligned} \quad (12.3)$$

It follows that the spin axes γ_+ and γ_- transform under a right-handed rotation by angle χ about γ_z as

$$\gamma_{\pm} \rightarrow e^{\mp i\chi} \gamma_{\pm} . \quad (12.4)$$

The transformation (12.4) identifies the spin axes γ_+ and γ_- as having spin $+1$ and -1 respectively.

12.1.3 Spin

More generally, an object can be defined as having spin s if it varies by

$$e^{-si\chi} \quad (12.5)$$

under a right-handed rotation by angle χ about the preferred axis γ_z . Thus an object of spin s is unchanged by a rotation of $2\pi/s$ about the preferred axis. A spin-0 object is symmetric about the γ_z axis, unchanged by a rotation of any angle about the axis. The γ_z axis itself is spin-0, as is the time axis γ_t .

The components of a tensor in a spin tetrad inherit spin properties from that of the spin basis. The general

rule is that the spin s of any tensor component is equal to the number of + covariant indices minus the number of – covariant indices:

$$\boxed{\text{spin } s = \text{number of } + \text{ minus } - \text{ covariant indices}} . \quad (12.6)$$

12.1.4 Spin flip

Under a reflection through the y -axis, the spin axes swap:

$$\gamma_+ \leftrightarrow \gamma_- , \quad (12.7)$$

which may also be accomplished by complex conjugation. Reflection through the y -axis, or equivalently complex conjugation, changes the sign of all spin indices of a tensor component

$$+ \leftrightarrow - . \quad (12.8)$$

In short, complex conjugation flips spin, a pretty feature of the spin formalism.

12.1.5 Spin versus spherical harmonics

In physical problems, such as in cosmological perturbations, or in perturbations of spherical black holes, or in the hydrogen atom, spin often appears in conjunction with an expansion in spherical harmonics. Spin should not be confused with spherical harmonics.

Spin and spherical harmonics appear together whenever the problem at hand has a symmetry under the 3D special orthogonal group $SO(3)$ of spatial rotations (special means of unit determinant; the full orthogonal group $O(3)$ contains in addition the discrete transformation corresponding to reflection of one of the axes, which flips the sign of the determinant). Rotations in $SO(3)$ are described by 3 Euler angles $\{\theta, \phi, \chi\}$. Spin is associated with the Euler angle χ . The usual spherical harmonics $Y_{\ell m}(\theta, \phi)$ are the spin-0 eigenfunctions of $SO(3)$. The eigenfunctions of the full $SO(3)$ group are the spin harmonics SIGN?

$${}_s Y_{\ell m}(\theta, \phi, \chi) = \Theta_{\ell m s}(\theta, \phi, \chi) e^{im\phi} e^{is\chi} . \quad (12.9)$$

12.1.6 Spin components of the Einstein tensor

With respect to a spin tetrad, the components of the Einstein tensor G_{mn} are

$$G_{mn} = \begin{pmatrix} G_{tt} & G_{tz} & G_{t+} & G_{t-} \\ G_{tz} & G_{zz} & G_{z+} & G_{z-} \\ G_{t+} & G_{z+} & G_{++} & G_{+-} \\ G_{t-} & G_{z-} & G_{+-} & G_{--} \end{pmatrix} . \quad (12.10)$$

From this it is apparent that the 10 components of the Einstein tensor decompose into 4 spin-0 components, 4 spin- ± 1 components, and 2 spin- ± 2 components:

$$\begin{aligned}
 -2: & \quad G_{--} , \\
 -1: & \quad G_{t-} , G_{z-} , \\
 0: & \quad G_{tt} , G_{tz} , G_{zz} , G_{+-} , \\
 +1: & \quad G_{t+} , G_{z+} , \\
 +2: & \quad G_{++} .
 \end{aligned} \tag{12.11}$$

The 4 spin-0 components are all real; in particular G_{+-} is real since $G_{+-}^* = G_{-+} = G_{+-}$. The 4 spin- ± 1 and 2 spin- ± 2 components comprise 3 complex components

$$G_{++}^* = G_{--} , \quad G_{t+}^* = G_{t-} , \quad G_{z+}^* = G_{z-} . \tag{12.12}$$

In some contexts, for example in cosmological perturbation theory, REALLY? the various spin components are commonly referred to as scalar (spin-0), vector (spin- ± 1), and tensor (spin- ± 2).

12.2 Newman-Penrose tetrad formalism

The Newman-Penrose formalism (Newman and Penrose, 1962; Newman and Penrose, 2009) provides a particularly powerful way to deal with fields that propagate at the speed of light. The Newman-Penrose formalism adopts a tetrad in which the two axes γ_v (outgoing) and γ_u (ingoing) along the direction of propagation are chosen to be lightlike, while the two axes γ_+ and γ_- transverse to the direction of propagation are chosen to be spin axes.

Sadly, the literature on the Newman-Penrose formalism is characterized by an arcane and random notation whose principal purpose seems to be to perpetuate exclusivity for an old-boys club of people who understand it. This is unfortunate given the intrinsic power of the formalism. Held (1974) comments that the Newman-Penrose formalism presents “a formidable notational barrier to the uninitiate.” For example, the tetrad connections Γ_{kmn} are called “spin coefficients,” and assigned individual greek letters that obscure their transformation properties. Do not be fooled: all the standard tetrad formalism presented in Chapter 11 carries through unaltered. One ill-born child of the notation that persists in widespread use is ψ_{2-s} for the spin s component of the Weyl tensor, equations (12.30).

Gravitational waves are commonly characterized by the Newman-Penrose (NP) components of the Weyl tensor. The NP components of the Weyl tensor are sometimes referred to as the NP scalars. The designation as NP scalars is potentially misleading, because the NP components of the Weyl tensor form a tetrad-frame tensor, not a set of scalars (though of course the tetrad-frame Weyl tensor is, like any tetrad-frame quantity, a coordinate scalar). The NP components do become proper quantities, and in that sense scalars, when referred to the frame of a particular observer, such as a gravitational wave telescope, observing along a particular direction. However, the use of the word scalar to describe the components of a tensor is unfortunate.

12.2.1 Newman-Penrose tetrad

A Newman-Penrose tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$ is defined in terms of an orthonormal tetrad $\{\gamma_t, \gamma_x, \gamma_y, \gamma_z\}$ by

$$\boxed{\gamma_v \equiv \frac{1}{\sqrt{2}}(\gamma_t + \gamma_z)} , \quad (12.13a)$$

$$\boxed{\gamma_u \equiv \frac{1}{\sqrt{2}}(\gamma_t - \gamma_z)} , \quad (12.13b)$$

$$\boxed{\gamma_+ \equiv \frac{1}{\sqrt{2}}(\gamma_x + i\gamma_y)} , \quad (12.13c)$$

$$\boxed{\gamma_- \equiv \frac{1}{\sqrt{2}}(\gamma_x - i\gamma_y)} , \quad (12.13d)$$

or in matrix form

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & i & 0 \\ 0 & 1 & -i & 0 \end{pmatrix} \begin{pmatrix} \gamma_t \\ \gamma_x \\ \gamma_y \\ \gamma_z \end{pmatrix} . \quad (12.14)$$

All four tetrad axes are null

$$\gamma_v \cdot \gamma_v = \gamma_u \cdot \gamma_u = \gamma_+ \cdot \gamma_+ = \gamma_- \cdot \gamma_- = 0 . \quad (12.15)$$

In a profound sense, the null, or lightlike, character of each the four NP axes explains why the NP formalism is well adapted to treating fields that propagate at the speed of light. The tetrad metric of the Newman-Penrose tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$ is

$$\gamma_{mn} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} . \quad (12.16)$$

12.2.2 Boost weight

A boost by rapidity θ along the γ_z axis multiplies the outgoing and ingoing axes γ_v and γ_u by a blueshift factor e^θ and its reciprocal,

$$\begin{aligned} \gamma_v &\rightarrow e^\theta \gamma_v , \\ \gamma_u &\rightarrow e^{-\theta} \gamma_u . \end{aligned} \quad (12.17)$$

In terms of the velocity $v = \tanh \theta$, the blueshift factor is the special relativistic Doppler shift factor

$$e^\theta = \left(\frac{1+v}{1-v} \right)^{1/2} . \quad (12.18)$$

More generally, object is said to have boost weight n if it varies by

$$e^{n\theta} \quad (12.19)$$

under a boost by rapidity θ along the preferred direction γ_z . Thus γ_v has boost weight $+1$, and γ_u has boost weight -1 . The spin axes $\gamma_{\pm}$ both have boost weight 0 . The NP components of a tensor inherit their boost weight properties from those of the NP basis. The general rule is that the boost weight n of any tensor component is equal to the number of v covariant indices minus the number of u covariant indices:

$$\boxed{\text{boost weight } n = \text{number of } v \text{ minus } u \text{ covariant indices}}. \quad (12.20)$$

12.2.3 Lorentz transformations

Under a Lorentz transformation consisting of a combination of a Lorentz boost by rapidity ξ about t - x and a rotation by angle ζ about y - z , an orthonormal tetrad $\gamma_m \equiv \{\gamma_t, \gamma_x, \gamma_y, \gamma_z\}$ transforms as **FIX SIGNS**

$$\gamma_m \rightarrow \gamma'_m = \begin{pmatrix} \cosh(\xi) & -\sinh(\xi) & 0 & 0 \\ -\sinh(\xi) & \cosh(\xi) & 0 & 0 \\ 0 & 0 & \cos(\zeta) & \sin(\zeta) \\ 0 & 0 & -\sin(\zeta) & \cos(\zeta) \end{pmatrix} \begin{pmatrix} \gamma_t \\ \gamma_x \\ \gamma_y \\ \gamma_z \end{pmatrix}. \quad (12.21)$$

Under the same Lorentz transformation, the bivector axes $\gamma_{tm} \equiv \{\gamma_{tx}, \gamma_{ty}, \gamma_{tz}\}$ transform as

$$\gamma_{tm} \rightarrow \gamma'_{tm} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\zeta + i\xi) & \sin(\zeta + i\xi) \\ 0 & -\sin(\zeta + i\xi) & \cos(\zeta + i\xi) \end{pmatrix} \begin{pmatrix} \gamma_{tx} \\ \gamma_{ty} \\ \gamma_{tz} \end{pmatrix}. \quad (12.22)$$

12.3 Weyl tensor

The Weyl tensor is the trace-free part of the Riemann tensor,

$$\boxed{C_{klmn} \equiv R_{klmn} - \frac{1}{2}(\gamma_{km}R_{ln} - \gamma_{kn}R_{lm} + \gamma_{ln}R_{km} - \gamma_{lm}R_{kn}) + \frac{1}{6}(\gamma_{km}\gamma_{ln} - \gamma_{kn}\gamma_{lm})R}. \quad (12.23)$$

By construction, the Weyl tensor vanishes when contracted on any pair of indices. Whereas the Ricci and Einstein tensors vanish identically in any region of spacetime containing no energy-momentum, $T_{mn} = 0$, the Weyl tensor can be non-vanishing. Physically, the Weyl tensor describes tidal forces and gravitational waves.

12.3.1 Complexified Weyl tensor

The Weyl tensor is, like the Riemann tensor, a symmetric matrix of bivectors. Just as the electromagnetic bivector F_{kl} has a natural complex structure, so also the Weyl tensor C_{klmn} has a natural complex structure. The properties of the Weyl tensor emerge most plainly when that complex structure is made manifest.

In an orthonormal tetrad $\{\gamma_t, \gamma_x, \gamma_y, \gamma_z\}$, the Weyl tensor C_{klmn} can be written as a 6×6 symmetric bivector matrix, organized as a 2×2 matrix of 3×3 blocks, with the structure

$$C = \begin{pmatrix} C_{EE} & C_{EB} \\ C_{BE} & C_{BB} \end{pmatrix} = \left(\begin{array}{ccc|ccc} C_{txtx} & C_{txty} & C_{txtz} & C_{txzy} & C_{txxz} & C_{txyx} \\ C_{tytx} & \dots & \dots & \dots & \dots & \dots \\ C_{tztx} & \dots & \dots & \dots & \dots & \dots \\ \hline C_{zytx} & \dots & \dots & \dots & \dots & \dots \\ C_{xztx} & \dots & \dots & \dots & \dots & \dots \\ C_{yxtx} & \dots & \dots & \dots & \dots & \dots \end{array} \right), \quad (12.24)$$

where E denotes electric indices, B magnetic indices, per the designation (??). The condition of being symmetric implies that the 3×3 blocks C_{EE} and C_{BB} are symmetric, while $C_{BE} = C_{EB}^\top$. The cyclic symmetry (11.63b) of the Riemann, hence Weyl, tensor implies that the off-diagonal 3×3 block C_{EB} (and likewise C_{BE}) is traceless.

The natural complex structure motivates defining a complexified Weyl tensor $\tilde{C}_{klmn}$ by

$$\tilde{C}_{klmn} \equiv \frac{1}{4} \left(\delta_k^p \delta_l^q + \frac{i}{2} \varepsilon_{kl}{}^{pq} \right) \left(\delta_m^r \delta_n^s + \frac{i}{2} \varepsilon_{mn}{}^{rs} \right) C_{pqrs} \quad \text{a tetrad tensor} \quad (12.25)$$

analogously to the definition (??) of the complexified electromagnetic field. The definition (12.25) of the complexified Weyl tensor $\tilde{C}_{klmn}$ is valid in any frame, not just an orthonormal frame. In an orthonormal frame, if the Weyl tensor C_{klmn} is organized according to the structure (12.24), then the complexified Weyl tensor $\tilde{C}_{klmn}$ defined by equation (12.25) has the structure

$$\tilde{C} = \frac{1}{4} \begin{pmatrix} 1 & -i \\ -i & -1 \end{pmatrix} (C_{EE} - C_{BB} + iC_{EB} + iC_{BE}). \quad (12.26)$$

Thus the independent components of the complexified Weyl tensor $\tilde{C}_{klmn}$ constitute a 3×3 complex symmetric traceless matrix $C_{EE} - C_{BB} + i(C_{EB} + C_{BE})$, with 5 complex degrees of freedom. Although the complexified Weyl tensor $\tilde{C}_{klmn}$ is defined, equation (12.25), as a projection of the Weyl tensor, it nevertheless retains all the 10 degrees of freedom of the original Weyl tensor C_{klmn} .

The same complexification projection operator applied to the trace (Ricci) parts of the Riemann tensor yields only the Ricci scalar multiplied by that unique combination of the tetrad metric that has the symmetries of the Riemann tensor. Thus complexifying the trace parts of the Riemann tensor produces nothing useful.

12.3.2 Newman-Penrose components of the Weyl tensor

With respect to a NP null tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$, equation (41.1), the Weyl tensor C_{klmn} has 5 distinct complex components, here denoted ψ_s , of spins respectively $s = -2, -1, 0, +1$, and $+2$:

$$\begin{aligned}
 -2: \quad \psi_{-2} & \equiv C_{u-u-} , \\
 -1: \quad \psi_{-1} & \equiv C_{uvu-} = C_{+-u-} , \\
 0: \quad \psi_0 & \equiv \frac{1}{2} (C_{uvuv} + C_{uv+-}) = \frac{1}{2} (C_{+-+-} + C_{uv+-}) = C_{v+-u} , \\
 +1: \quad \psi_1 & \equiv C_{vuv+} = C_{-+v+} , \\
 +2: \quad \psi_2 & \equiv C_{v+v+} .
 \end{aligned} \tag{12.27}$$

The complex conjugates ψ_s^* of the 5 NP components of the Weyl tensor are:

$$\begin{aligned}
 \psi_{-2}^* & = C_{u+u+} , \\
 \psi_{-1}^* & = C_{uvu+} = C_{-+u+} , \\
 \psi_0^* & = \frac{1}{2} (C_{uvuv} + C_{uv-+}) = \frac{1}{2} (C_{-+-+} + C_{uv-+}) = C_{v-+u} , \\
 \psi_1^* & = C_{vuv-} = C_{+-v-} , \\
 \psi_2^* & = C_{v-v-} .
 \end{aligned} \tag{12.28}$$

whose spins have the opposite sign, in accordance with the rule (12.8) that complex conjugation flips spin. The above expressions (12.27) and (12.28) account for all the NP components C_{klmn} of the Weyl tensor but four, which vanish identically:

$$C_{v+v-} = C_{u+u-} = C_{v+u+} = C_{v-u-} = 0 . \tag{12.29}$$

The above convention that the index s on the NP component ψ_s labels its spin differs from the standard convention, where the spin s component of the Weyl tensor is impenetrably denoted ψ_{2-s} (e.g. Chandrasekhar (1983)):

$$\begin{aligned}
 -2: \quad \psi_4 , \\
 -1: \quad \psi_3 , \\
 0: \quad \psi_2 , \quad (\text{standard convention, not followed here}) \\
 +1: \quad \psi_1 , \\
 +2: \quad \psi_0 .
 \end{aligned} \tag{12.30}$$

12.3.3 Newman-Penrose components of the complexified Weyl tensor

The non-vanishing NP components of the complexified Weyl tensor $\tilde{C}_{klmn}$ defined by equation (12.25) are

$$\begin{aligned}
 \tilde{C}_{u-u-} & = \psi_{-2} , \\
 \tilde{C}_{uvu-} & = \tilde{C}_{+-u-} = \psi_{-1} , \\
 \tilde{C}_{uvuv} = \tilde{C}_{+-+-} & = \tilde{C}_{uv+-} = \tilde{C}_{v+-u} = \psi_0 , \\
 \tilde{C}_{vuv+} & = \tilde{C}_{-+v+} = \psi_1 , \\
 \tilde{C}_{v+v+} & = \psi_2 .
 \end{aligned} \tag{12.31}$$

whereas any component with either of its two bivector indices equal to $v-$ or $u+$ vanishes. As with the complexified electromagnetic field, the rule that complex conjugation flips spin fails here because the complexification operator breaks the rule. Equations (12.31) show that the complexified Weyl tensor in an NP tetrad contains just 5 distinct non-vanishing complex components, and those components are precisely equal to the complex spin components ψ_s .

With respect to a triple of bivector indices ordered as $\{u-, uv, +v\}$, the NP components of the complexified Weyl tensor constitute the 3×3 complex symmetric matrix

$$\tilde{C}_{klmn} = \begin{pmatrix} \psi_{-2} & \psi_{-1} & \psi_0 \\ \psi_{-1} & \psi_0 & \psi_1 \\ \psi_0 & \psi_1 & \psi_2 \end{pmatrix}. \quad (12.32)$$

12.3.4 Components of the complexified Weyl tensor in an orthonormal tetrad

The complexified Weyl tensor forms a 3×3 complex symmetric traceless matrix in any frame, not just an NP frame. In an orthonormal frame, with respect to a triple of bivector indices $\{tx, ty, tz\}$, the complexified Weyl tensor $\tilde{C}_{klmn}$ can be expressed in terms of the NP spin components ψ_s as

$$\tilde{C}_{klmn} = \begin{pmatrix} \psi_0 & \frac{1}{2}(\psi_1 - \psi_{-1}) & -\frac{i}{2}(\psi_1 + \psi_{-1}) \\ \frac{1}{2}(\psi_1 - \psi_{-1}) & -\frac{1}{2}\psi_0 + \frac{1}{4}(\psi_2 + \psi_{-2}) & -\frac{i}{4}(\psi_2 - \psi_{-2}) \\ -\frac{i}{2}(\psi_1 + \psi_{-1}) & -\frac{i}{4}(\psi_2 - \psi_{-2}) & -\frac{1}{2}\psi_0 - \frac{1}{4}(\psi_2 + \psi_{-2}) \end{pmatrix}. \quad (12.33)$$

12.3.5 Propagating components of gravitational waves

For outgoing gravitational waves, only the spin -2 component ψ_{-2} (the one conventionally called ψ_4) propagates, carrying gravitational waves from a source to infinity:

$$\psi_{-2} : \text{propagating, outgoing}. \quad (12.34)$$

This propagating, outgoing -2 component has spin -2 , but its complex conjugate has spin $+2$, so effectively both spin components, or helicities, or circular polarizations, of an outgoing gravitational wave are embodied in the single complex component. The remaining 4 complex NP components (spins -1 to 2) of an outgoing gravitational wave are short range, describing the gravitational field near the source.

Similarly, only the spin $+2$ component ψ_2 of an ingoing gravitational wave propagates, carrying energy from infinity:

$$\psi_2 : \text{propagating, ingoing}. \quad (12.35)$$

12.4 Petrov classification of the Weyl tensor

As seen above, the complexified Weyl tensor is a complex symmetric traceless 3×3 matrix. If the matrix were real symmetric (or complex Hermitian), then standard mathematical theorems would guarantee that

Table 12.1: Petrov classification of the Weyl tensor

Petrov type	Distinct eigenvalues	Distinct eigenvectors	Normal form of the complexified Weyl tensor
I	3	3	$\begin{pmatrix} \psi_0 & 0 & 0 \\ 0 & -\frac{1}{2}\psi_0 + \frac{1}{2}\psi_2 & 0 \\ 0 & 0 & -\frac{1}{2}\psi_0 - \frac{1}{2}\psi_2 \end{pmatrix}$
D	2	3	$\begin{pmatrix} \psi_0 & 0 & 0 \\ 0 & -\frac{1}{2}\psi_0 & 0 \\ 0 & 0 & -\frac{1}{2}\psi_0 \end{pmatrix}$
II	2	2	$\begin{pmatrix} \psi_0 & 0 & 0 \\ 0 & -\frac{1}{2}\psi_0 + \frac{1}{4}\psi_2 & -\frac{i}{4}\psi_2 \\ 0 & -\frac{i}{4}\psi_2 & -\frac{1}{2}\psi_0 - \frac{1}{4}\psi_2 \end{pmatrix}$
O	1	3	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
N	1	2	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{4}\psi_2 & -\frac{i}{4}\psi_2 \\ 0 & -\frac{i}{4}\psi_2 & -\frac{1}{4}\psi_2 \end{pmatrix}$
III	1	1	$\begin{pmatrix} 0 & \frac{1}{2}\psi_1 & -\frac{i}{2}\psi_1 \\ \frac{1}{2}\psi_1 & 0 & 0 \\ -\frac{i}{2}\psi_1 & 0 & 0 \end{pmatrix}$

it would be diagonalizable, with a complete set of eigenvalues and eigenvectors. But the Weyl matrix is complex symmetric, and there is no such theorem.

The mathematical theorems state that a matrix is diagonalizable if and only if it has a complete set of linearly independent eigenvectors. Since there is always at least one distinct linearly independent eigenvector associated with each distinct eigenvalue, if all eigenvalues are distinct, then necessarily there is a complete set of eigenvectors, and the Weyl tensor is diagonalizable. However, if some of the eigenvalues coincide, then there may not be a complete set of linearly independent eigenvectors, in which case the Weyl tensor is not diagonalizable.

The Petrov classification, tabulated in Table 12.1, classifies the Weyl tensor in accordance with the number of distinct eigenvalues and eigenvectors. The normal form is with respect to an orthonormal frame aligned with the eigenvectors to the extent possible. The tetrad with respect to which the complexified Weyl tensor takes its normal form is called the **Weyl principal tetrad**. The Weyl principal tetrad is unique except in cases D, O, and N. For Types D and N, the Weyl principal tetrad is unique up to Lorentz transformations

that leave the eigen-bivector γ_{tz} unchanged, which is to say, transformations generated by the Lorentz rotor $\exp(\zeta\gamma_{tz})$ where ζ is complex.

The Kerr-Newman geometry is Type D. General spherically symmetric geometries are Type D. The Friedmann-Lemaître-Robertson-Walker geometry is Type O. Plane gravitational waves are Type N.

The geometric algebra

The geometric algebra is a conceptually appealing and mathematically powerful formalism. If you want to understand rotations, Lorentz transformations, spin- $\frac{1}{2}$ particles, and supersymmetry, and you want to do actual calculations elegantly and (relatively) easily, then the geometric algebra is the thing to learn.

The extension of the geometric algebra to Minkowski space is called the spacetime algebra, which is the subject of Chapter 14. The natural extensions of the geometric and spacetime algebras to spinors are called the supergeometric algebra, covered in Chapters 40 and 41. All these algebras may be referred to collectively as geometric algebras. I am generally unenthusiastic about mathematical formalism for its own sake. The geometric algebras are a mathematical language that Nature appears to speak.

The geometric algebra builds on a broad mathematical heritage beginning with the work of Grassmann (1862; 1877) and Clifford (1878). The exposition in this book owes much to the conceptual rethinking of the subject by David Hestenes (Hestenes, 1966; Hestenes and Sobczyk, 1987).

This Chapter starts by setting up the geometric algebra in N -dimensional Euclidean space $\mathbb{R}^N$, then specializes to the cases of 2 and 3 dimensions. The generalization to 4-dimensional Minkowski space, where the geometric algebra is called the spacetime algebra, is deferred to Chapter 14. The 4-dimensional spacetime algebra proves to be identical to the Clifford algebra of the Dirac γ -matrices, which explains the adoption of the symbol $\boldsymbol{\gamma}_m$ to denote the basis vectors of a tetrad. Although the formalism is presented initially in Euclidean or Minkowski space, everything generalizes immediately to general relativity, where the basis vectors $\boldsymbol{\gamma}_m$ form the basis of an orthonormal tetrad at each point of spacetime.

This book follows the standard physics convention that a rotor R rotates a multivector $\boldsymbol{a}$ as $\boldsymbol{a} \rightarrow R\boldsymbol{a}\tilde{R}$ and a spinor φ as $\varphi \rightarrow R\varphi$. This, along with the standard definition (13.19) for the pseudoscalar, has the consequence that a right-handed rotation corresponds to $R = e^{-i\theta/2}$ with θ increasing, and that rotations accumulate to the left, that is, a rotation R followed by a rotation S is the product SR . The physics convention is opposite to that adopted in OpenGL and by the computer graphics industry, where a right-handed rotation corresponds to $R = e^{i\theta/2}$, and rotations accumulate to the right, that is, R followed by S is RS .

In this book, a multivector is written in boldface. A rotor is written in normal (not bold) face as a reminder that, even though a rotor is an even member of the geometric algebra, it can also be regarded as a spin- $\frac{1}{2}$ object with a transformation law (13.75) different from that (13.56) of multivectors. Earlier latin indices

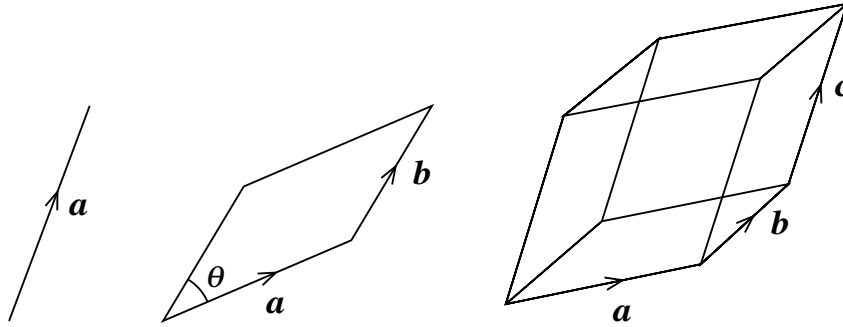


Figure 13.1 Multivectors of grade 1, 2, and 3: a vector $\mathbf{a}$ (left), a bivector $\mathbf{a} \wedge \mathbf{b}$ (middle), and a trivector $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$ (right).

$a, b, \dots$ run over spatial indices $1, 2, \dots$ only, while mid latin indices $m, n, \dots$ run over both time and space indices $0, 1, 2, \dots$.

13.1 Products of vectors

In 3-dimensional Euclidean space $\mathbb{R}^3$, there are two familiar ways of taking the product of two vectors, the scalar product and the vector product.

1. The **scalar product** $\mathbf{a} \cdot \mathbf{b}$, also known as the dot product or inner product, of two vectors $\mathbf{a}$ and $\mathbf{b}$ is a scalar of magnitude $|\mathbf{a}| |\mathbf{b}| \cos \theta$, where $|\mathbf{a}|$ and $|\mathbf{b}|$ are the lengths of the two vectors, and θ the angle between them. The scalar product is commutative, $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$.
2. The **vector product**, $\mathbf{a} \times \mathbf{b}$, also known as the cross product, is a vector of magnitude $|\mathbf{a}| |\mathbf{b}| \sin \theta$, directed perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, such that $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a} \times \mathbf{b}$ form a right-handed set. The vector product is anticommutative, $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$.

The definition of the scalar product continues to work fine in a Euclidean space of any dimension, but the definition of the vector product works only in three dimensions, because in two dimensions there is no vector perpendicular to two vectors, and in four or more dimensions there are many vectors perpendicular to two vectors. It is therefore useful to define a more general version, the outer product (Grassmann, 1862) that works in Euclidean space $\mathbb{R}^N$ of any dimension.

3. The **outer product** $\mathbf{a} \wedge \mathbf{b}$, also known as the wedge product or exterior product, of two vectors $\mathbf{a}$ and $\mathbf{b}$ is a **bivector**, a multivector of dimension 2, or **grade 2**. The bivector $\mathbf{a} \wedge \mathbf{b}$ is the directed 2-dimensional area, of magnitude $|\mathbf{a}| |\mathbf{b}| \sin \theta$, of the parallelogram formed by the vectors $\mathbf{a}$ and $\mathbf{b}$, as illustrated in Figure 13.1. The bivector has an orientation, or handedness, defined by circulating the parallelogram first along $\mathbf{a}$, then along $\mathbf{b}$. The outer product is anticommutative, $\mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a}$, like its forebear the vector product.

The outer product can be repeated, so that $(\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c}$ is a **trivector**, a directed volume, a multivector

of grade 3. The magnitude of the trivector is the volume of the parallelepiped defined by the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, illustrated in Figure 13.1. The outer product is by construction associative, $(\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c} = \mathbf{a} \wedge (\mathbf{b} \wedge \mathbf{c})$. Associativity, together with anticommutativity of bivectors, implies that the trivector $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$ is totally antisymmetric under permutations of the three vectors, that is, it is unchanged under even permutations, and changes sign under odd permutations. The ordering of an outer product thus defines one of two handednesses.

It is a familiar concept that a vector $\mathbf{a}$ can be regarded as a geometric object, a directed length, independent of the coordinates used to describe it. The components of a vector change when the reference frame changes, but the vector itself remains the same physical thing. In the same way, a bivector $\mathbf{a} \wedge \mathbf{b}$ is a directed area, and a trivector $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$ is a directed volume, both geometric objects with a physical meaning independent of the coordinate system.

In two dimensions the triple outer product of any three vectors is zero, $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = 0$, because the volume of a parallelepiped confined to a plane is zero. More generally, in N -dimensional space $\mathbb{R}^N$, the outer product of $N + 1$ vectors is zero

$$\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \cdots \wedge \mathbf{a}_{N+1} = 0 \quad (N \text{ dimensions}) . \quad (13.1)$$

13.2 Geometric product

The inner and outer products offer two different ways of multiplying vectors. However, by itself neither product conforms to the usual desideratum of multiplication, that the product of two elements of a set be an element of the set. Taking the inner product of a vector with another vector lowers the dimension by one, while taking the outer product raises the dimension by one.

Grassmann (1877) and Clifford (1878) resolved the problem by defining a **multivector** as any linear combination of scalars, vectors, bivectors, and objects of higher grade. Let $\gamma_1, \gamma_2, \dots, \gamma_n$ form an orthonormal basis for N -dimensional Euclidean space $\mathbb{R}^N$. A multivector in $N = 2$ dimensions is then a linear combination of

$$\begin{array}{lll} 1, & \gamma_1, \gamma_2, & \gamma_1 \wedge \gamma_2, \\ 1 \text{ scalar} & 2 \text{ vectors} & 1 \text{ bivector} \end{array} \quad (13.2)$$

forming a linear space of dimension $1 + 2 + 1 = 4 = 2^2$. Similarly, a multivector in $N = 3$ dimensions is a linear combination of

$$\begin{array}{llll} 1, & \gamma_1, \gamma_2, \gamma_3, & \gamma_1 \wedge \gamma_2, \gamma_2 \wedge \gamma_3, \gamma_3 \wedge \gamma_1, & \gamma_1 \wedge \gamma_2 \wedge \gamma_3, \\ 1 \text{ scalar} & 3 \text{ vectors} & 3 \text{ bivectors} & 1 \text{ trivector} \end{array} \quad (13.3)$$

forming a linear space of dimension $1 + 3 + 3 + 1 = 8 = 2^3$. In general, multivectors in N dimensions form a linear space of dimension 2^N , with $N!/[n!(N-n)!]$ distinct basis elements of grade n .

A multivector $\mathbf{a}$ in N -dimensional Euclidean space $\mathbb{R}^N$ can thus be written as a linear combination of basis elements

$$\mathbf{a} = \sum_{\text{distinct } \{a,b,\dots,d\} \subseteq \{1,2,\dots,N\}} a^{ab\dots d} \gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_d \quad (13.4)$$

the sum being over all 2^N distinct subsets of $\{1, 2, \dots, N\}$. The index on each component $a^{ab\dots d}$ is a totally antisymmetric quantity, reflecting the total antisymmetry of $\gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_d$.

The point of introducing multivectors is to allow multiplication to be defined so that the product of two multivectors is a multivector. The key trick is to define the **geometric product** $\mathbf{ab}$ of two vectors $\mathbf{a}$ and $\mathbf{b}$ to be the sum of their inner and outer products:

$$\boxed{\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b}} . \quad (13.5)$$

That is a seriously big trick, and if you buy a ticket to it, you are in for a seriously big ride. As a particular example of (13.5), the geometric product of any element γ_a of the orthonormal basis with itself is a scalar, and with any other element of the basis is a bivector:

$$\gamma_a \gamma_b = \begin{cases} 1 & (a = b) \\ \gamma_a \wedge \gamma_b & (a \neq b) \end{cases} . \quad (13.6)$$

Conversely, the rules (39.70), plus distributivity, imply the multiplication rule (13.5). A generalization of the rule (39.70) completes the definition of the geometric product:

$$\gamma_a \gamma_b \dots \gamma_d = \gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_d \quad (a, b, \dots, d \text{ all distinct}) . \quad (13.7)$$

The rules (39.70) and (13.7), along with the usual requirements of associativity and distributivity, combined with commutativity of scalars and anticommutativity of pairs of γ_a , uniquely define multiplication over the space of multivectors. For example, the product of the bivector $\gamma_1 \wedge \gamma_2$ with the vector γ_1 is

$$(\gamma_1 \wedge \gamma_2) \gamma_1 = \gamma_1 \gamma_2 \gamma_1 = -\gamma_2 \gamma_1 \gamma_1 = -\gamma_2 . \quad (13.8)$$

Sometimes it is convenient to denote the outer product (13.7) of distinct basis elements by the abbreviated symbol γ_A or $\gamma_{ab\dots d}$,

$$\gamma_A = \gamma_{ab\dots d} \equiv \gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_d \quad (a, b, \dots, d \text{ all distinct}) . \quad (13.9)$$

By construction, γ_A with $A = ab\dots d$ is antisymmetric in its indices $a, b, \dots, d$. The product of two general multivectors $\mathbf{a} = a^A \gamma_A$ and $\mathbf{b} = b^B \gamma_B$ is

$$\mathbf{ab} = a^A b^B \gamma_A \gamma_B , \quad (13.10)$$

with paired indices A and B implicitly summed over distinct subsets of $\{1, \dots, N\}$. By construction, the geometric algebra is associative,

$$(\mathbf{ab})\mathbf{c} = \mathbf{a}(\mathbf{bc}) . \quad (13.11)$$

Does the geometric algebra form a group under multiplication? No. One of the defining properties of a group is that every element should have an inverse. But, for example,

$$(1 + \gamma_1)(1 - \gamma_1) = 0 \quad (13.12)$$

shows that neither $1 + \gamma_1$ nor $1 - \gamma_1$ has an inverse.

13.3 Reverse

The **reverse** of any basis element is defined to be the reversed product

$$\gamma_a \wedge \widetilde{\gamma_b \wedge \dots \wedge \gamma_d} \equiv \gamma_d \wedge \dots \wedge \gamma_b \wedge \gamma_a . \quad (13.13)$$

The product of a basis multivector γ_A and its reverse is 1,

$$\gamma_A \widetilde{\gamma_A} = \widetilde{\gamma_A} \gamma_A = 1 . \quad (13.14)$$

The reverse $\widetilde{\mathbf{a}}$ of any multivector $\mathbf{a}$ is the multivector obtained by reversing each of its components. Reversion leaves unchanged all multivectors whose grade mod 4 is 0 or 1, and changes the sign of all multivectors whose grade mod 4 is 2 or 3. Thus the reverse of a multivector $\mathbf{a}$ of pure grade p is

$$\widetilde{\mathbf{a}} = (-)^{[p/2]} \mathbf{a} , \quad (13.15)$$

where $[p/2]$ signifies the largest integer less than or equal to $p/2$. For example, scalars and vectors are unchanged by reversion, but bivectors and trivectors change sign. Reversion satisfies

$$\widetilde{\mathbf{a} + \mathbf{b}} = \widetilde{\mathbf{a}} + \widetilde{\mathbf{b}} , \quad (13.16)$$

$$\widetilde{\mathbf{ab}} = \widetilde{\mathbf{b}} \widetilde{\mathbf{a}} . \quad (13.17)$$

Among other things, it follows that the reverse of any product of multivectors is the reversed product, as you would hope:

$$\widetilde{\mathbf{ab} \dots \mathbf{c}} = \widetilde{\mathbf{c}} \dots \widetilde{\mathbf{ba}} . \quad (13.18)$$

13.4 The pseudoscalar and the Hodge dual

Orthogonal to any n -dimensional subspace of N -dimensional space is an $(N-n)$ -dimensional space, called the Hodge dual space. For example, the Hodge dual of a bivector in 2 dimensions is a 0-dimensional object, a pseudoscalar. Similarly, the Hodge dual of a bivector in 3 dimensions is a 1-dimensional object, a pseudovector.

13.4.1 Pseudoscalar

Define the **pseudoscalar** I_N in N dimensions to be

$$I_N \equiv \gamma_1 \wedge \gamma_2 \wedge \dots \wedge \gamma_N \quad (13.19)$$

with reverse

$$\widetilde{I_N} = (-)^{[N/2]} I_N , \quad (13.20)$$

where $[N/2]$ signifies the largest integer less than or equal to $N/2$. The square of the pseudoscalar is

$$I_N^2 = (-1)^{[N/2]} = \begin{cases} 1 & \text{if } N \pmod{4} = 0 \text{ or } 1 \\ -1 & \text{if } N \pmod{4} = 2 \text{ or } 3 . \end{cases} \quad (13.21)$$

The pseudoscalar anticommutes (commutes) with vectors $\mathbf{a}$, that is, with multivectors of grade 1, if N is even (odd):

$$\begin{aligned} I_N \mathbf{a} &= -\mathbf{a} I_N & \text{if } N \text{ is even} \\ I_N \mathbf{a} &= \mathbf{a} I_N & \text{if } N \text{ is odd} . \end{aligned} \quad (13.22)$$

This implies that the pseudoscalar I_N commutes with all even grade elements of the geometric algebra, and that it anticommutes (commutes) with all odd elements of the algebra if N is even (odd). Concisely, if $\mathbf{a}$ has grade p , then

$$I_N \mathbf{a} = (-1)^{p(N-p)} \mathbf{a} I_N . \quad (13.23)$$

Exercise 13.1 Schur's lemma. Prove that the only multivectors that commute with all elements of the algebra are linear combinations of the scalar 1 and, if N is odd, the pseudoscalar I_N .

Solution. Suppose that $\mathbf{a}$ is a multivector that commutes with all elements of the algebra. Then in particular $\mathbf{a}$ commutes with every basis element $\boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b \wedge \dots \wedge \boldsymbol{\gamma}_d$. Since multiplication by a basis element permutes the basis elements amongst each other (and multiplies each by ± 1), it follows that $\mathbf{a}$ commutes with a basis element only if each of the components of $\mathbf{a}$ commutes separately with that basis element. Thus each component of $\mathbf{a}$ must commute separately with all basis elements of the algebra. Amongst the basis elements of the algebra, only the scalar 1, and, if the dimension N is odd, the pseudoscalar I_N , equation (13.22), commute with all other basis elements. Thus $\mathbf{a}$ must be some linear combination of 1 and, if N is odd, the pseudoscalar I_N .

13.4.2 Hodge dual

The **Hodge dual** $^*\mathbf{a}$ of a multivector $\mathbf{a}$ in N dimensions is defined by pre-multiplication by the pseudoscalar I_N ,

$$^*\mathbf{a} \equiv I_N \mathbf{a} . \quad (13.24)$$

In 3 dimensions, the Hodge duals of the basis vectors $\boldsymbol{\gamma}_a$ are the bivectors

$$I_3 \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 \wedge \boldsymbol{\gamma}_3 , \quad I_3 \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 \wedge \boldsymbol{\gamma}_1 , \quad I_3 \boldsymbol{\gamma}_3 = \boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2 . \quad (13.25)$$

Thus in 3 dimensions the bivector $\mathbf{a} \wedge \mathbf{b}$ is seen to be the pseudovector Hodge dual to the familiar vector product $\mathbf{a} \times \mathbf{b}$:

$$\mathbf{a} \wedge \mathbf{b} = I_3 \mathbf{a} \times \mathbf{b} . \quad (13.26)$$

13.5 General products of multivectors

13.5.1 Pure grade components of products of multivectors

It is useful to be able to project out a particular grade component of a multivector. The grade p component of a multivector $\mathbf{a}$ is denoted

$$\langle \mathbf{a} \rangle_p , \quad (13.27)$$

so that for example $\langle \mathbf{a} \rangle_0$, $\langle \mathbf{a} \rangle_1$, and $\langle \mathbf{a} \rangle_2$ represent respectively the scalar, vector, and bivector components of $\mathbf{a}$. By construction, a multivector is the sum of its pure grade components, $\mathbf{a} = \langle \mathbf{a} \rangle_0 + \langle \mathbf{a} \rangle_1 + \dots + \langle \mathbf{a} \rangle_N$.

The geometric product of a multivector $\mathbf{a}$ of pure grade p with a multivector $\mathbf{b}$ of pure grade q is in general a sum of multivectors of grades $|p-q|$ to $\min(p+q, N)$. The product $\mathbf{ab}$ is in general neither commutative nor anticommutative, but the pure grade components of the product commute or anticommute according to

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{pq-n^2} \langle \mathbf{ba} \rangle_{p+q-2n} \quad (13.28)$$

for $n = [(p+q-N)/2]$ to $\min(p, q)$. Written out in components, the grade $p+q-2n$ component of the geometric product of $\mathbf{a} = a^A \boldsymbol{\gamma}_A$ and $\mathbf{b} = b^B \boldsymbol{\gamma}_B$ is

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{[n/2]} a^{AC} b_C^B \boldsymbol{\gamma}_A \wedge \boldsymbol{\gamma}_B , \quad (13.29)$$

implicitly summed over distinct sequences A , B , and C of respectively $p-n$, $q-n$, and n indices. The factor $(-)^{[n/2]}$ comes from the square of a grade- n orthonormal multivector, $\boldsymbol{\gamma}_C \boldsymbol{\gamma}_C = (-)^{[n/2]}$. Only components with the $p+q+n$ indices of ABC all distinct contribute.

Equation (13.29) can also be written

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{[n/2]} \frac{(p+q-2n)!}{(p-n)!(q-n)!} a^{[AC} b_C^{B]} \boldsymbol{\gamma}_{AB} , \quad (13.30)$$

implicitly summed over distinct sequences AB and C of respectively $p+q-2n$ and n indices. The binomial factor is the number of ways of picking the $p-n$ distinct indices of A and the $q-n$ distinct indices of B from each distinct antisymmetric sequence AB of $p+q-2n$ indices.

13.5.2 Wedge product

The wedge product of multivectors of arbitrary grade is defined, consistent with the convention of differential forms, §15.8, to be the highest possible grade component of the geometric product. The wedge product of a multivector $\mathbf{a}$ of grade p with a multivector $\mathbf{b}$ of grade q is thus defined to be

$$\boxed{\mathbf{a} \wedge \mathbf{b} \equiv \langle \mathbf{ab} \rangle_{p+q}} . \quad (13.31)$$

The definition (13.31) is consistent with the definition of the wedge product of vectors (multivectors of grade 1) in §13.1. The wedge product is commutative or anticommutative as pq is even or odd,

$$\mathbf{a} \wedge \mathbf{b} = (-)^{pq} \mathbf{b} \wedge \mathbf{a} , \quad (13.32)$$

which is a special case of equation (13.28). The wedge product is associative,

$$(\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c} = \mathbf{a} \wedge (\mathbf{b} \wedge \mathbf{c}) . \quad (13.33)$$

In accordance with the definition (13.31), the wedge product of a scalar a (a multivector of grade 0) with a multivector $\mathbf{b}$ equals the usual product of the scalar and the multivector,

$$a \wedge \mathbf{b} = a\mathbf{b} \quad \text{if } a \text{ is a scalar} , \quad (13.34)$$

again consistent with the convention of differential forms.

13.5.3 Dot product

The dot product of multivectors of arbitrary grade is defined to be the lowest grade component of their geometric product,

$$\boxed{\mathbf{a} \cdot \mathbf{b} \equiv \langle \mathbf{a}\mathbf{b} \rangle_{|p-q|}} , \quad (13.35)$$

except that the dot product of a scalar, a zero grade multivector, with any multivector is conveniently defined to be zero,

$$a \cdot \mathbf{b} = 0 \quad \text{if } a \text{ is a scalar} . \quad (13.36)$$

The convention (13.36) is adopted to ensure that, if $\mathbf{b}$ is a vector, then $\mathbf{a}\mathbf{b} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b}$ for any multivector $\mathbf{a}$, including the case where $\mathbf{a}$ is a scalar. The dot product is symmetric or antisymmetric,

$$\mathbf{a} \cdot \mathbf{b} = (-1)^{(p-q)q} \mathbf{b} \cdot \mathbf{a} \quad \text{for } p \geq q . \quad (13.37)$$

The dot product is *not* associative.

13.5.4 Scalar product

The dot product of two multivectors of the same grade is a scalar, and in this case the dot product can be called the scalar product. The scalar product of two multivectors of the same grade p is

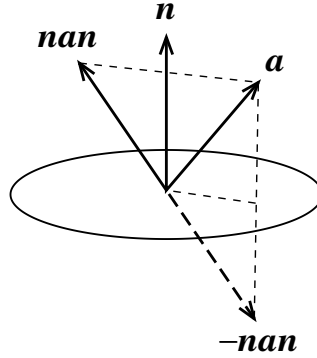
$$\mathbf{a} \cdot \mathbf{b} = a^A \boldsymbol{\gamma}_A \cdot b^B \boldsymbol{\gamma}_B = (-1)^{[p/2]} a^A b_A , \quad (13.38)$$

implicitly summed over distinct sequences A of p indices. Equation (13.38) is a special case of equation (13.29).

13.5.5 Triple products of multivectors

The associativity of the geometric product implies that the grade 0 component of a triple product of multivectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ of grades respectively p , q , r satisfies an associative law

$$\langle \mathbf{a}\mathbf{b}\mathbf{c} \rangle_0 = \langle \langle \mathbf{a}\mathbf{b} \rangle_r \mathbf{c} \rangle_0 = \langle \mathbf{a} \langle \mathbf{b}\mathbf{c} \rangle_p \rangle_0 . \quad (13.39)$$

Figure 13.2 Reflection of a vector $\mathbf{a}$ through axis $\mathbf{n}$.

More generally, the grade s component of a triple product of multivectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ of non-zero grades respectively p , q , r (any grade 0 multivector, i.e. scalar, can be taken outside the product) satisfies

$$\langle \mathbf{abc} \rangle_s = \sum_{n=|r-s|}^{r+s} \langle \langle \mathbf{ab} \rangle_n \mathbf{c} \rangle_s = \sum_{n=|p-s|}^{p+s} \langle \mathbf{a} \langle \mathbf{bc} \rangle_n \rangle_s . \quad (13.40)$$

Often some terms vanish, simplifying the relation. As an example of the triple-product relation (13.40), if $\mathbf{a}$ and $\mathbf{b}$ are multivectors of grades p and q respectively, and neither are scalars, and their wedge product does not vanish (that is, $p + q \leq N$), then the wedge and dot products of $\mathbf{a}$ and $\mathbf{b}$ are related by Hodge duality relations

$$I_N(\mathbf{a} \wedge \mathbf{b}) = (I_N \mathbf{a}) \cdot \mathbf{b} , \quad (\mathbf{a} \wedge \mathbf{b}) I_N = \mathbf{a} \cdot (\mathbf{b} I_N) , \quad (13.41)$$

where I_N is the pseudoscalar (13.19).

13.6 Reflection

Multiplying a vector (a multivector of grade 1) by a vector shifts the grade (dimension) of the vector by ± 1 . Thus, if one wants to transform a vector into another vector (with the same grade, one), at least two multiplications by a vector are required.

The simplest non-trivial transformation of a vector $\mathbf{a}$ is

$$\mathbf{n} : \mathbf{a} \rightarrow \mathbf{n} \mathbf{a} \mathbf{n} , \quad (13.42)$$

in which the vector $\mathbf{a}$ is multiplied on both left and right with a unit vector $\mathbf{n}$. If $\mathbf{a}$ is resolved into components $\mathbf{a}_{\parallel}$ and $\mathbf{a}_{\perp}$ respectively parallel and perpendicular to $\mathbf{n}$, then the transformation (13.42) is

$$\mathbf{n} : \mathbf{a}_{\parallel} + \mathbf{a}_{\perp} \rightarrow \mathbf{a}_{\parallel} - \mathbf{a}_{\perp} , \quad (13.43)$$

which represents a **reflection** of the vector $\mathbf{a}$ through the axis $\mathbf{n}$, a reversal of all components of the vector perpendicular to $\mathbf{n}$, as illustrated by Figure 13.2. Note that $-\mathbf{n}\mathbf{a}\mathbf{n}$ is the reflection of $\mathbf{a}$ through the hypersurface normal to $\mathbf{n}$, a reversal of the component of the vector parallel to $\mathbf{n}$.

The operation of left- and right-multiplying by a unit vector $\mathbf{n}$ reflects not only vectors, but multivectors $\mathbf{a}$ in general:

$$\mathbf{n} : \mathbf{a} \rightarrow \mathbf{n}\mathbf{a}\mathbf{n} . \quad (13.44)$$

For example, the product $\mathbf{a}\mathbf{b}$ of two vectors transforms as

$$\mathbf{n} : \mathbf{a}\mathbf{b} \rightarrow \mathbf{n}(\mathbf{a}\mathbf{b})\mathbf{n} = (\mathbf{n}\mathbf{a}\mathbf{n})(\mathbf{n}\mathbf{b}\mathbf{n}) \quad (13.45)$$

which works because $\mathbf{n}^2 = 1$.

A reflection leaves any scalar λ unchanged, $\mathbf{n} : \lambda \rightarrow \mathbf{n}\lambda\mathbf{n} = \lambda\mathbf{n}^2 = \lambda$. Geometrically, a reflection preserves the lengths of, and angles between, all vectors.

13.7 Rotation

Two successive reflections yield a rotation. Consider reflecting a vector $\mathbf{a}$ (a multivector of grade 1) first through the unit vector $\mathbf{m}$, then through the unit vector $\mathbf{n}$:

$$\mathbf{nm} : \mathbf{a} \rightarrow \mathbf{n}\mathbf{m}\mathbf{a}\mathbf{m}\mathbf{n} . \quad (13.46)$$

Any component $\mathbf{a}_\perp$ of $\mathbf{a}$ simultaneously orthogonal to both $\mathbf{m}$ and $\mathbf{n}$ (i.e. $\mathbf{m} \cdot \mathbf{a}_\perp = \mathbf{n} \cdot \mathbf{a}_\perp = 0$) is unchanged by the transformation (13.46), since each reflection flips the sign of $\mathbf{a}_\perp$:

$$\mathbf{nm} : \mathbf{a}_\perp \rightarrow \mathbf{n}\mathbf{m}\mathbf{a}_\perp\mathbf{m}\mathbf{n} = -\mathbf{n}\mathbf{a}_\perp\mathbf{n} = \mathbf{a}_\perp . \quad (13.47)$$

Rotations inherit from reflections the property of preserving the lengths of, and angles between, all vectors. Thus the transformation (13.46) must represent a **rotation** of those components $\mathbf{a}_\parallel$ of $\mathbf{a}$ lying in the 2-dimensional plane spanned by $\mathbf{m}$ and $\mathbf{n}$, as illustrated by Figure 13.3. To determine the angle by which the plane is rotated, it suffices to consider the case where the vector $\mathbf{a}_\parallel$ is equal to $\mathbf{m}$ (or $\mathbf{n}$, as a check). It is not too hard to figure out that, if the angle from $\mathbf{m}$ to $\mathbf{n}$ is $\theta/2$, then the rotation angle is θ in the same sense, from $\mathbf{m}$ to $\mathbf{n}$.

For example, if $\mathbf{m}$ and $\mathbf{n}$ are parallel, so that $\mathbf{m} = \pm\mathbf{n}$, then the angle between $\mathbf{m}$ and $\mathbf{n}$ is $\theta/2 = 0$ or π , and the transformation (13.46) rotates the vector $\mathbf{a}_\parallel$ by $\theta = 0$ or 2π , that is, it leaves $\mathbf{a}_\parallel$ unchanged. This makes sense: two reflections through the same plane leave everything unchanged. If on the other hand $\mathbf{m}$ and $\mathbf{n}$ are orthogonal, then the angle between them is $\theta/2 = \pm\pi/2$, and the transformation (13.46) rotates $\mathbf{a}_\parallel$ by $\theta = \pm\pi$, that is, it maps $\mathbf{a}_\parallel$ to $-\mathbf{a}_\parallel$.

The rotation (13.46) can be abbreviated

$$\boxed{R : \mathbf{a} \rightarrow R\mathbf{a}\tilde{R}} \quad (13.48)$$

where $R = \mathbf{nm}$ is called a **rotor**, and $\tilde{R} = \mathbf{mn}$ is its reverse. Rotors are **unimodular**, satisfying $R\tilde{R} =$

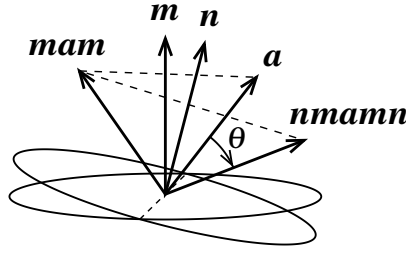


Figure 13.3 Two successive reflections of a vector $\mathbf{a}$, first through $\mathbf{m}$, then through $\mathbf{n}$, yield a rotation of a vector $\mathbf{a}$ by the bivector $\mathbf{mn}$. Baffled? Hey, draw your own picture.

$\tilde{R}R = 1$. According to the discussion above, the transformation (13.48) corresponds to a rotation by angle θ in the $\mathbf{mn}$ -plane if the angle from $\mathbf{m}$ to $\mathbf{n}$ is $\theta/2$. Then $\mathbf{m} \cdot \mathbf{n} = \cos \theta/2$ and $\mathbf{m} \wedge \mathbf{n} = (\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2) \sin \theta/2$, where $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are two orthonormal vectors spanning the $\mathbf{m}$ - $\mathbf{n}$ plane, oriented so that the angle from $\boldsymbol{\gamma}_1$ to $\boldsymbol{\gamma}_2$ is positive $\pi/2$ (i.e. $\boldsymbol{\gamma}_1$ is the x -axis and $\boldsymbol{\gamma}_2$ the y -axis). Note that the outer product $\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$ is invariant under rotations in the $\mathbf{mn}$ -plane, hence independent of the choice of orthonormal basis vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. It follows that the rotor $R = \mathbf{nm} = \mathbf{n} \cdot \mathbf{m} + \mathbf{n} \wedge \mathbf{m}$ corresponding to a right-handed rotation by θ in the $\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2$ -plane is given by

$$R = \cos \frac{\theta}{2} - (\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2) \sin \frac{\theta}{2} . \quad (13.49)$$

The rotor (13.49) can also be written as an exponential of the bivector $\boldsymbol{\theta} = \theta \boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$,

$$R = e^{-\boldsymbol{\theta}/2} . \quad (13.50)$$

It is straightforward to check that the action of the rotor (13.49) on the basis vectors $\boldsymbol{\gamma}_a$ is

$$R : \boldsymbol{\gamma}_1 \rightarrow R\boldsymbol{\gamma}_1\tilde{R} = \boldsymbol{\gamma}_1 \cos \theta + \boldsymbol{\gamma}_2 \sin \theta , \quad (13.51a)$$

$$R : \boldsymbol{\gamma}_2 \rightarrow R\boldsymbol{\gamma}_2\tilde{R} = \boldsymbol{\gamma}_2 \cos \theta - \boldsymbol{\gamma}_1 \sin \theta , \quad (13.51b)$$

$$R : \boldsymbol{\gamma}_a \rightarrow R\boldsymbol{\gamma}_a\tilde{R} = \boldsymbol{\gamma}_a \quad (a \neq 1, 2) , \quad (13.51c)$$

which corresponds to a right-handed rotation of the basis vectors $\boldsymbol{\gamma}_a$ by angle θ in the $\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2$ -plane. The inverse rotation is

$$\tilde{R} : \mathbf{a} \rightarrow \tilde{R}\mathbf{a}R \quad (13.52)$$

with

$$\tilde{R} = \cos \frac{\theta}{2} + (\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2) \sin \frac{\theta}{2} . \quad (13.53)$$

A rotation of the form (13.49), a rotation in a single plane, is called a **simple rotation**.

In the geometric algebra, a rotation is considered to rotate the axes $\boldsymbol{\gamma}_a \rightarrow \boldsymbol{\gamma}'_a$ while leaving the components

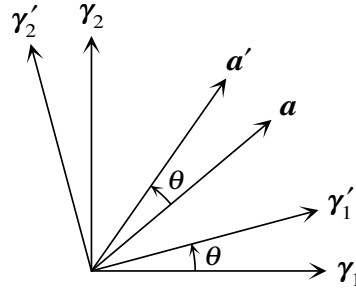


Figure 13.4 Right-handed rotation of a vector $\mathbf{a}$ by angle θ in the $\gamma_1\gamma_2$ -plane. A rotation in the geometric algebra is an active rotation, which rotates the axes $\gamma_a \rightarrow \gamma'_a$ while leaving the components a^a of a multivector unchanged, equation (13.54). In other words, multivectors $\mathbf{a}$ are considered to be attached to the frame, and a rotation bodily rotates the frame and everything attached to it.

a^a of a multivector unchanged. Thus a rotation transforms a vector $\mathbf{a}$ as

$$R : \mathbf{a} = a^a \gamma_a \rightarrow \mathbf{a}' = a^a \gamma'_a . \quad (13.54)$$

Figure 13.4 illustrates a right-handed rotation by angle θ of a vector $\mathbf{a}$ in the $\gamma_1\gamma_2$ -plane.

A rotation first by R and then by S transforms a vector $\mathbf{a}$ as

$$SR : \mathbf{a} \rightarrow SR\mathbf{a}\widetilde{R}\widetilde{S} = SR\mathbf{a}\widetilde{SR} . \quad (13.55)$$

Thus the composition of two rotations, first R and then S , is given by their geometric product SR . This is the physics convention, where rotations accumulate to the left (in contrast to the computer graphics convention, where rotations accumulate to the right). In three dimensions or less, all rotations are simple, but in four dimensions or higher, compositions of simple rotations can yield rotations that are not simple. For example, a rotation in the $\gamma_1\gamma_2$ -plane followed by a rotation in the $\gamma_3\gamma_4$ -plane is not equivalent to any simple rotation. However, it will be seen in §14.3 that bivectors in the 4D spacetime algebra have a natural complex structure, which allows 4D spacetime rotations to take a simple form similar to (13.49), but with complex angle θ and two orthogonal planes of rotation combined into a complex pair of planes.

A rotor R rotates not only vectors, but multivectors $\mathbf{a}$ in general:

$$\boxed{R : \mathbf{a} \rightarrow R\mathbf{a}\widetilde{R}} . \quad (13.56)$$

For example, the product $\mathbf{ab}$ of two vectors transforms as

$$R : \mathbf{ab} \rightarrow R(\mathbf{ab})\widetilde{R} = (R\mathbf{a}\widetilde{R})(R\mathbf{b}\widetilde{R}) \quad (13.57)$$

which works because $\widetilde{R}R = 1$.

To summarize, the characterization of rotations by rotors has considerable advantages. Firstly, the transformation (13.56) applies to multivectors $\mathbf{a}$ of arbitrary grade in arbitrarily many dimensions. Secondly, the composition law is particularly simple, the composition of two rotations being given by their geometric

product. A third advantage is that rotors rotate not only vectors and multivectors, but also spin- $\frac{1}{2}$ objects—indeed rotors are themselves spin- $\frac{1}{2}$ objects—as might be suspected from the intriguing factor of $\frac{1}{2}$ in front of the angle θ in equation (13.49).

Concept question 13.2 How fast do bivectors rotate? Rotors rotate half as fast as vectors. How fast do bivectors rotate?

1. Bivectors don't rotate.
2. Half as fast as vectors.
3. The same as vectors.
4. Twice as fast as vectors.
5. None of the above.

13.8 Rotor group

The **rotor group** is the group generated by the bivectors of the geometric algebra. The rotor group in N dimensions is also called $\text{Spin}(N)$, and is the covering group of the special orthogonal group $\text{SO}(N)$ of proper rotations in N dimensions (the S in $\text{SO}(N)$ signifies special, that is, matrices of unit determinant, which removes improper rotations with determinant -1 that occur when a spatial axis is reflected).

The rotor, or rotation, group is an example of a continuous group called a **Lie group**. A right-handed rotation $\exp(-\frac{1}{2}\theta\gamma_a\wedge\gamma_b)$ by finite angle θ in the $\gamma_a\gamma_b$ -plane can be thought of as being built up from an infinite number of infinitesimal rotations $\exp(-\frac{1}{2}\delta\theta\gamma_a\wedge\gamma_b)$ by angles $\delta\theta$. To linear order, an infinitesimal rotation by angle $\delta\theta$ in the $\gamma_a\gamma_b$ -plane is

$$\exp(-\frac{1}{2}\delta\theta\gamma_a\wedge\gamma_b) = 1 - \frac{1}{2}\delta\theta\gamma_a\wedge\gamma_b. \quad (13.58)$$

The bivector $-\gamma_a\wedge\gamma_b$ is said to be the **generator** of a right-handed rotation in the $\gamma_a\gamma_b$ -plane.

The Baker-Campbell-Hausdorff formula states that the product of exponentials of not-necessarily-commuting elements θ and ϕ is

$$\exp(\theta)\exp(\phi) = \exp\left(\theta + \phi + \frac{1}{2}[\theta, \phi] + \frac{1}{12}[[\theta, \phi], \phi] - \frac{1}{12}[[\theta, \phi], \theta] + \dots\right), \quad (13.59)$$

where $[\theta, \phi] \equiv \theta\phi - \phi\theta$ is the commutator of θ and ϕ , also called their **Lie bracket**. Thus finite rotations are built from exponentials of linear combinations of generators and their commutators. A set of linearly independent generators that close under commutation provides a basis for the **Lie algebra** of a Lie group. The commutator of two bivectors is a bivector, so the Lie algebra of rotations is the set of bivectors. The rotor group is the Lie group generated by the bivectors.

Concept question 13.3 What is the dimension of the rotor group in N dimensions? **Answer.** The dimension of the rotor group is the number of its generators, its bivectors, which is $N(N-1)/2$.

Concept question 13.4 Is the rotor group the same as the group of even, unimodular elements of the geometric algebra? All rotors are even, unimodular elements of the geometric algebra. The properties of being even and unimodular are preserved under composition, so the set of even, unimodular elements forms a group. Is the rotor group the same as the group of even, unimodular elements? **Answer.** In low dimensions $N \leq 5$ yes, but in general no. See part 4 of Exercise 13.6.

Exercise 13.5 The even geometric algebra in $N+1$ dimensions is isomorphic to the full geometric algebra in N dimensions. Show that the even geometric algebra in $N+1$ dimensions is isomorphic to the full geometric algebra in N dimensions. Conclude that the dimension of the even geometric algebra in $N+1$ dimensions is 2^N .

Solution. Decompose a multivector $\mathbf{a}$ in N dimensions into its even and odd parts, $\mathbf{a} = \mathbf{a}_{\text{even}} + \mathbf{a}_{\text{odd}}$. The mapping

$$\mathbf{a}_{\text{even}} + \mathbf{a}_{\text{odd}} \leftrightarrow \mathbf{a}_{\text{even}} + \mathbf{a}_{\text{odd}} \gamma_{N+1} \quad (13.60)$$

is an isomorphism between the N -dimensional geometric algebra and the $(N+1)$ -dimensional even algebra ($\mathbf{a}_{\text{even}}$ and $\mathbf{a}_{\text{odd}} \gamma_{N+1}$ are both elements of the even algebra in $N+1$ dimensions). The mapping is an isomorphism because it respects addition and multiplication, and it respects rotations that leave γ_{N+1} invariant, that is, rotations in the N -dimensional geometric algebra.

Exercise 13.6 Lie groups generated by multivectors. An element R of a Lie group generated by a set of multivectors γ_A takes the form $R = \exp(-\frac{1}{2} \sum_A \theta_A \gamma_A)$. The element R acts on elements $\mathbf{a}$ of the geometric algebra by $R : \mathbf{a} \rightarrow R\mathbf{a}R^{-1}$, where the inverse of R is $R^{-1} = \exp(\frac{1}{2} \sum_A \theta_A \gamma_A)$. A set of multivectors γ_A generates a Lie group provided that the set is closed under commutation, in accordance with the Baker-Campbell-Hausdorff formula (13.59). Show that the non-zero commutators of two orthonormal multivectors of grades respectively p and q in N dimensions have grades $p+q-2n$ where

$$n \in [\max(0, p+q-N), \min(p, q)] \quad (13.61)$$

is an even integer if both p and q are odd, or an odd integer if either of p or q is even. In particular, show that the non-zero commutators of two orthonormal multivectors of the same grade p have grades $2+4j$ where $j \in [0, [(p-1)/2]]$ is an integer. Conclude that, if $\hat{p}$ denotes a multivector of grade $p \pmod{4}$, then

$$[\hat{p}, \hat{p}] = \hat{2}, \quad [\hat{2}, \hat{p}] = \hat{p}, \quad [\hat{0}, \hat{1}] = \hat{3}, \quad [\hat{0}, \hat{3}] = \hat{1}, \quad [\hat{1}, \hat{3}] = \hat{0}. \quad (13.62)$$

Conclude that the following are Lie groups generated by multivectors in the geometric algebra. All groups preserve the scalar product of two multivectors. All groups have the rotor group as a subgroup. The notation $G^A(N)$ for the group generated by multivectors with grades modulo 4 in the set A follows Shirokov (2018).

1. The rotor group, generated by bivectors. The rotor group acting on a multivector $\mathbf{a}$ preserves the grade of $\mathbf{a}$. The dimension (number of generators) of the group is $N(N-1)/2$.

2. The group generated by vectors and bivectors (multivectors of grades 1 and 2). The dimension of the group is $N(N+1)/2$.
3. Pseudo versions of the above groups, namely:
 - a. The group generated by bivectors and pseudobivectors, dimension $N(N-1)$ for $N \geq 5$.
 - b. The group generated by pseudovectors and bivectors, dimension $N(N+1)/2$ for $N \geq 4$.
 - c. The group generated by vectors, pseudovectors, bivectors and pseudobivectors, dimension $N(N+1)$ for $N \geq 5$.
4. The group $G^2(N)$ generated by multivectors of grade 2 (mod 4) (thus grades 2, 6, 10, ...). The group may be called the even unimodular group since it is the largest group whose elements R are all even and unimodular, satisfying $R^{-1} = \tilde{R}$. In dimensions $N \leq 5$, the even unimodular group coincides with the rotor group. The group preserves the grade $p \bmod 4$ of a multivector. The dimension of the group is

$$\dim G^2(N) = 2^{[(N-2)/2]} (2^{[(N-1)/2]} + s) , \quad s = \begin{cases} -1 \\ 0 \\ 1 \end{cases} \text{ as } (N+2) \bmod 8 = \begin{cases} 1, 2, 3, \\ 0, 4, \\ 5, 6, 7. \end{cases} \quad (13.63)$$

5. The group $G^{12}(N)$ generated by multivectors of grade (1 or 2) (mod 4) (thus grades 1, 2, 5, 6, 9, 10, ...). The group is the largest group whose elements R all have inverses equal to their parity-flipped reverses, parity-flipped meaning all odd-grade components change sign. The dimension of the group is

$$\dim G^{12}(N) = \dim G^2(N+1) . \quad (13.64)$$

6. The group $G^{23}(N)$ generated by multivectors of grade (2 or 3) (mod 4) (thus grades 2, 3, 6, 7, 10, 11, ...). The group may be called the unimodular group, since it is the largest group whose elements R are all unimodular, satisfying $R^{-1} = \tilde{R}$. The dimension of the group is

$$\begin{aligned} \dim G^{23}(N) &= 2^{N-1} - \dim G^2(N+1) + 2 \dim G^2(N) \\ &= 2^{[(N-1)/2]} (2^{[N/2]} + s) , \quad s = \begin{cases} -1 \\ 0 \\ 1 \end{cases} \text{ as } (N+1) \bmod 8 = \begin{cases} 1, 2, 3, \\ 0, 4, \\ 5, 6, 7. \end{cases} \end{aligned} \quad (13.65)$$

7. The even group $G^{02}(N)$ generated by multivectors of grade 0 (mod 2) (thus grades 0, 2, 4, 6, ...). The even group preserves the grade $p \bmod 2$ of a multivector (that is, whether the multivector is even or odd). The dimension of the group is 2^{N-1} . The special even group $SG^{02}(N)$ is generated by even multivectors excluding the unit element (thus grades 2, 4, 6, ...). The dimension of the special even group is $2^{N-1} - 1$.
8. The full group $G^{0123}(N)$ generated by multivectors of all grades (thus grades 0, 1, 2, 3, ...). The dimension of the group is 2^N . The special even group $SG^{0123}(N)$ is generated by multivectors excluding the unit element (thus grades 1, 2, 3, ...). The dimension of the special group is $2^N - 1$.
9. There are also complex Lie groups in which some generators are permitted to be imaginary or complex. The complex Lie groups are:
 - a. The complex rotor group generated by complex bivectors.

- b. The group generated by imaginary vectors and real bivectors.
- c. The group generated by complex vectors and complex bivectors.
- d. Pseudo versions of the above.
- e. The remaining groups can be denoted G^{AiB} following Shirokov (2018), with real generators of grades $A \pmod{4}$ and imaginary generators of grades $B \pmod{4}$:

$$G^{2i2}, \quad G^{2ip}, \quad G^{2pi2p}, \quad G^{2pi\overline{2p}}, \quad G^{0123i0123}, \quad (13.66)$$

where p runs over $0, 1, 3$, and $\overline{2p}$ denotes the opposite of $2p$ (for example $\overline{20} = 13$).

Solution. The dimension of each Lie group $G^A(N)$, the number of its generators, is established as follows. Let m_k denote the number of multivectors of grade $k \pmod{4}$,

$$m_k \equiv \sum_{p=k \pmod{4}} \binom{N}{p}. \quad (13.67)$$

The binomial theorem implies (i is the imaginary)

$$(1 + i^j)^N = \sum_{k=0}^3 i^{jk} m_k \quad \text{for } j = 0 \text{ to } 3, \quad (13.68)$$

or explicitly

$$\begin{pmatrix} 2^N \\ (1+i)^N \\ 0 \\ (1-i)^N \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix} \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}. \quad (13.69)$$

Equation (13.68) inverts to

$$m_k = \frac{1}{4} \sum_{j=0}^3 (-i)^{kj} (1 + i^j)^N, \quad (13.70)$$

or explicitly

$$\begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix} \begin{pmatrix} 2^N \\ (1+i)^N \\ 0 \\ (1-i)^N \end{pmatrix}. \quad (13.71)$$

The dimensions of the Lie groups are

$$\dim G^2(N) = m_2, \quad (13.72a)$$

$$\dim G^{12}(N) = m_1 + m_2, \quad (13.72b)$$

$$\dim G^{23}(N) = m_2 + m_3, \quad (13.72c)$$

$$\dim G^{02}(N) = m_0 + m_2, \quad (13.72d)$$

$$\dim G^{0123}(N) = m_0 + m_1 + m_2 + m_3. \quad (13.72e)$$

13.9 Active and passive rotations

So far in this book, indices have indicated how an object transforms, so that the notation

$$a^m \gamma_m \quad (13.73)$$

indicates a scalar, an object that is unchanged by a transformation, because the transformation of the contravariant vector a^m cancels against the corresponding transformation of the covariant vector γ_m .

However, the transformation (13.56) of a multivector is an example of an **active** transformation that rotates the basis vectors γ_A while keeping the coefficients a^A fixed, as opposed to a **passive** transformation that rotates the tetrad while keeping the thing itself unchanged. An active rotation bodily rotates a multivector $\mathbf{a}$, whereas a passive rotation rotates the frame without changing the multivector. Figure 13.4 illustrates the example of an active right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane, equations (13.51).

Under an active rotation, a multivector $\mathbf{a} \equiv a^A \gamma_A$ (implicit summation over distinct antisymmetrized subsets A of $\{1, \dots, N\}$) is not a scalar under the transformation (13.56), but rather transforms to the multivector $\mathbf{a}' \equiv a^A \gamma'_A$ given by

$$R : a^A \gamma_A \rightarrow a^A \tilde{R} \gamma_A R = a^A \gamma'_A . \quad (13.74)$$

13.10 A rotor is a spin- $\frac{1}{2}$ object

A rotor is an even, unimodular element of the geometric algebra, §13.7. As a multivector, a rotor R would transform under a rotation by the rotor S as $R \rightarrow SR\tilde{S}$. As a rotor, however, the rotor R transforms under a rotation by the rotor S as

$$S : R \rightarrow SR , \quad (13.75)$$

according to the transformation law (13.55). That is, composition in the rotor group is defined by the transformation (13.75): R rotated by S is SR .

The expression (13.49) for a simple rotation in the $\gamma_1\gamma_2$ -plane shows that the rotor corresponding to a rotation by 2π is -1 . Thus under a rotation (13.75) by 2π , a rotor R changes sign:

$$2\pi : R \rightarrow -R . \quad (13.76)$$

A rotation by 4π is necessary to bring the rotor R back to its original value:

$$4\pi : R \rightarrow R . \quad (13.77)$$

Thus a rotor R behaves like a spin- $\frac{1}{2}$ object, requiring 2 full rotations to restore it to its original state.

The two different transformation laws for a rotor — as a multivector, and as a rotor — describe two different physical situations. The transformation of a rotor as a multivector answers the question, what is

the form of a rotor R rotated into another, primed, frame? In the unprimed frame, the rotor R transforms a multivector $\mathbf{a}$ to $R\mathbf{a}\tilde{R}$. In the primed frame rotated by rotor S from the unprimed frame, $\mathbf{a}' = S\mathbf{a}\tilde{S}$, the transformed rotor is $SR\tilde{S}$, since

$$\mathbf{a}' = S\mathbf{a}\tilde{S} \rightarrow SR\mathbf{a}\tilde{R}\tilde{S} = SR\tilde{S}\mathbf{a}'S\tilde{S} = SR\tilde{S}\mathbf{a}'\widetilde{SR\tilde{S}}. \quad (13.78)$$

By contrast, the transformation (13.75) of a rotor as a rotor answers the question, what is the rotor corresponding to a rotation R followed by a rotation S ?

13.11 2D rotations and complex numbers

In $N \leq 5$ dimensions, the rotor group consists of even, unimodular multivectors of the geometric subalgebra, part 4 of Exercise 13.6. In two dimensions, the even grade multivectors are linear combinations of the basis set

$$\begin{array}{ll} 1, & I_2, \\ 1 \text{ scalar} & 1 \text{ bivector (pseudoscalar)} \end{array} \quad (13.79)$$

forming a linear space of dimension 2. The sole bivector is the pseudoscalar $I_2 \equiv \boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$, equation (13.19), the highest grade element in 2 dimensions. The rotor R that produces a right-handed rotation by angle θ is, according to equation (13.49),

$$R = e^{-\boldsymbol{\theta}/2} = e^{-I_2 \theta/2} = \cos \frac{\theta}{2} - I_2 \sin \frac{\theta}{2}, \quad (13.80)$$

where $\boldsymbol{\theta} = I_2 \theta$ is the bivector whose magnitude is $(\tilde{\boldsymbol{\theta}}\boldsymbol{\theta})^{1/2} = \theta$.

Since the square of the pseudoscalar I_2 is minus one, the pseudoscalar resembles the pure imaginary i , the square root of -1 . Sure enough, the mapping

$$I_2 \leftrightarrow i \quad (13.81)$$

defines an isomorphism between the algebra of even grade multivectors in 2 dimensions and the field of complex numbers

$$a + I_2 b \leftrightarrow a + i b. \quad (13.82)$$

With the isomorphism (13.82), the rotor R that produces a right-handed rotation by angle θ is equivalent to the complex number

$$\boxed{R = e^{-i\theta/2}}. \quad (13.83)$$

The associated reverse rotor $\tilde{R}$ is

$$\tilde{R} = e^{i\theta/2}, \quad (13.84)$$

the complex conjugate of R . The group of 2D rotors is isomorphic to the group of complex numbers of unit magnitude, the unitary group $U(1)$,

$$2\text{D rotors} \cong U(1) . \quad (13.85)$$

Let $\mathbf{z}$ denote an even multivector, equivalent to some complex number by the isomorphism (13.82). According to the transformation formula (13.56), under the rotation $R = e^{-i\theta/2}$, the even multivector, or complex number, $\mathbf{z}$ transforms as

$$R : \mathbf{z} \rightarrow e^{-i\theta/2} \mathbf{z} e^{i\theta/2} = e^{-i\theta/2} e^{i\theta/2} \mathbf{z} = \mathbf{z} , \quad (13.86)$$

which is true because even multivectors in 2 dimensions commute, as complex numbers should. Equation (13.86) shows that the even multivector, or complex number, $\mathbf{z}$ is unchanged by a rotation. This might seem strange: shouldn't the rotation rotate the complex number $\mathbf{z}$ by θ in the Argand plane? The answer is that the rotation $R : \mathbf{a} \rightarrow R\mathbf{a}\tilde{R}$ rotates *vectors* $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ (Exercise 13.7), as already seen in the transformation (13.51). The same rotation leaves the scalar 1 and the bivector $I_2 \equiv \boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$ unchanged. If temporarily you permit yourself to think in 3 dimensions, you see that the bivector $\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$ is Hodge dual to the pseudovector $\boldsymbol{\gamma}_1 \times \boldsymbol{\gamma}_2$, which is the axis of rotation and is itself unchanged by the rotation, even though the individual vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are rotated.

Exercise 13.7 Rotation of a vector. Confirm that a right-handed rotation by angle θ rotates the axes $\boldsymbol{\gamma}_a$ by

$$R : \boldsymbol{\gamma}_1 \rightarrow e^{-i\theta/2} \boldsymbol{\gamma}_1 e^{i\theta/2} = \boldsymbol{\gamma}_1 \cos \theta + \boldsymbol{\gamma}_2 \sin \theta , \quad (13.87a)$$

$$R : \boldsymbol{\gamma}_2 \rightarrow e^{-i\theta/2} \boldsymbol{\gamma}_2 e^{i\theta/2} = \boldsymbol{\gamma}_2 \cos \theta - \boldsymbol{\gamma}_1 \sin \theta , \quad (13.87b)$$

in agreement with (13.51). The important thing to notice is that the pseudoscalar I_2 , hence i , anticommutes with the vectors $\boldsymbol{\gamma}_a$.

13.12 Quaternions

A **quaternion** can be regarded as a kind of souped-up complex number,

$$q = a + \imath b_1 + j b_2 + k b_3 , \quad (13.88)$$

where a and b_a ($a = 1, 2, 3$) are real numbers, and the three imaginary numbers $\imath, j, k$, are defined to satisfy¹

$$\imath^2 = j^2 = k^2 = -\imath j k = -1 . \quad (13.89)$$

¹ The choice $\imath j k = 1$ in the definition (13.89) is the *opposite* of the conventional definition $\imath j k = -1$ famously carved by William Rowan Hamilton in the stone of Brougham Bridge while walking with his wife along the Royal Canal to Dublin on 16 October 1843 (O'Donnell, 1983). To map to Hamilton's definition, you can take $\imath = -i$, $j = -j$, $k = -k$, or alternatively $\imath = i$, $j = -j$, $k = k$, or $\imath = k$, $j = j$, $k = i$. The adopted choice $\imath j k = 1$ has the merit that it avoids a treacherous minus sign in the isomorphism (13.105) between 3-dimensional pseudovectors and quaternions. The present choice also conforms to the convention used by OpenGL and other computer graphics programs.

Remark the dotless i (and j), to distinguish these quaternionic imaginaries from other possible imaginaries. A consequence of equations (13.89) is that each pair of imaginary numbers anticommutes:

$$ij = -ji = -k, \quad jk = -kj = -i, \quad ki = -ik = -j. \quad (13.90)$$

It is convenient to abbreviate the three imaginaries by i_a with $a = 1, 2, 3$,

$$\{i, j, k\} \equiv \{i_1, i_2, i_3\}. \quad (13.91)$$

The quaternion (13.88) can then be expressed compactly as a sum of its scalar a and vector (actually pseudovector, as will become apparent below from the isomorphism (13.105)) $\mathbf{b} = i_a b_a$ parts

$$q = a + \mathbf{b} = a + i_a b_a, \quad (13.92)$$

implicitly summed over $a = 1, 2, 3$. A fundamentally useful formula, which follows from the defining equations (13.89), is

$$\mathbf{ab} = (i_a a_a)(i_b b_b) = -\mathbf{a} \cdot \mathbf{b} - \mathbf{a} \times \mathbf{b} = -a_a b_a - i_a \varepsilon_{abc} a_b b_c, \quad (13.93)$$

where $\mathbf{a} \cdot \mathbf{b}$ and $\mathbf{a} \times \mathbf{b}$ denote the usual 3D scalar and vector products, and ε_{abc} is the usual totally antisymmetric matrix, with $\varepsilon_{123} = 1$. The product of two quaternions $p \equiv a + \mathbf{b}$ and $q \equiv c + \mathbf{d}$ can thus be written

$$\begin{aligned} pq &= (a + \mathbf{b})(c + \mathbf{d}) = (a + i_a b_a)(c + i_b d_b) \\ &= ac - \mathbf{b} \cdot \mathbf{d} + a\mathbf{d} + c\mathbf{b} - \mathbf{b} \times \mathbf{d} = ac - b_a d_a + i_a (ad_a + cb_a - \varepsilon_{abc} b_b d_c). \end{aligned} \quad (13.94)$$

The **quaternionic conjugate** $\tilde{q}$ of a quaternion $q \equiv a + \mathbf{b}$ is (the overbar symbol $\sim$ for quaternionic conjugation distinguishes it from the asterisk symbol $*$ for complex conjugation)

$$\tilde{q} = a - \mathbf{b} = a - i_a b_a. \quad (13.95)$$

The quaternionic conjugate of a product is the reversed product of quaternionic conjugates

$$\widetilde{pq} = \tilde{q}\tilde{p} \quad (13.96)$$

just like reversion in the geometric algebra, equation (13.17). The choice of the same symbol, an overbar, to represent both reversion and quaternionic conjugation is not coincidental. The **magnitude** $|q|$ of the quaternion $q \equiv a + \mathbf{b}$ is

$$|q| = (\tilde{q}q)^{1/2} = (q\tilde{q})^{1/2} = (a^2 + \mathbf{b} \cdot \mathbf{b})^{1/2} = (a^2 + b_a b_a)^{1/2}. \quad (13.97)$$

The magnitude of a quaternion is also called its **modulus**. A quaternion that has unit modulus, $\tilde{q}q = 1$, is called **unimodular**. The **inverse** q^{-1} of the quaternion, satisfying $qq^{-1} = q^{-1}q = 1$, is

$$q^{-1} = \tilde{q}/(\tilde{q}q) = (a - \mathbf{b})/(a^2 + \mathbf{b} \cdot \mathbf{b}) = (a - i_a b_a)/(a^2 + b_a b_a). \quad (13.98)$$

13.13 3D rotations and quaternions

As before, Concept question 13.4, in $N \leq 5$ dimensions, the rotor group consists of even, unimodular multivectors of the geometric subalgebra. In three dimensions, the even grade multivectors are linear combinations of the basis set

$$\begin{array}{ll} 1, & I_3\gamma_1, I_3\gamma_2, I_3\gamma_3, \\ 1 \text{ scalar} & 3 \text{ bivectors (pseudovectors)} \end{array} \quad (13.99)$$

forming a linear space of dimension 4. The three bivectors are pseudovectors, equation (13.25). The squares of the pseudovector basis elements are all minus one,

$$(I_3\gamma_1)^2 = (I_3\gamma_2)^2 = (I_3\gamma_3)^2 = -1, \quad (13.100)$$

and they anticommute with each other,

$$\begin{aligned} (I_3\gamma_1)(I_3\gamma_2) &= -(I_3\gamma_2)(I_3\gamma_1) = -I_3\gamma_3, \\ (I_3\gamma_2)(I_3\gamma_3) &= -(I_3\gamma_3)(I_3\gamma_2) = -I_3\gamma_1, \\ (I_3\gamma_3)(I_3\gamma_1) &= -(I_3\gamma_1)(I_3\gamma_3) = -I_3\gamma_2. \end{aligned} \quad (13.101)$$

The rotor R that produces a rotation by angle θ right-handedly about unit direction $n_a = \{n_1, n_2, n_3\}$, satisfying $n_a n_a = 1$, is, according to equation (13.49),

$$R = e^{-\theta/2} = e^{-\mathbf{n}\theta/2} = \cos \frac{\theta}{2} - \mathbf{n} \sin \frac{\theta}{2}. \quad (13.102)$$

where θ is the bivector

$$\theta \equiv \mathbf{n}\theta = I_3\gamma_a n_a \theta. \quad (13.103)$$

of magnitude $(\tilde{\theta}\theta)^{1/2} = \theta$ and unit direction $\mathbf{n} \equiv I_3\gamma_a n_a$ (satisfying $\tilde{\mathbf{n}}\mathbf{n} = 1$). The pseudovector I_3 is a commuting imaginary, commuting with all members of the 3D geometric algebra, both odd and even, and satisfying

$$I_3^2 = -1. \quad (13.104)$$

Comparison of equations (13.100) and (13.101) to equations (13.89) and (13.90), shows that the mapping

$$I_3\gamma_a \leftrightarrow \iota_a \quad (a = 1, 2, 3) \quad (13.105)$$

defines an isomorphism between the space of even multivectors in 3 dimensions and the non-commutative division algebra of quaternions

$$a + I_3\gamma_a b_a \leftrightarrow a + \iota_a b_a. \quad (13.106)$$

With the equivalence (13.106), the rotor R given by equation (13.102) can be interpreted as a quaternion, with θ the quaternion

$$\theta \equiv \mathbf{n}\theta = \iota_a n_a \theta. \quad (13.107)$$

The associated reverse rotor $\tilde{R}$ is

$$\tilde{R} = e^{\theta/2} = e^{\mathbf{n}\theta/2} = \cos \frac{\theta}{2} + \mathbf{n} \sin \frac{\theta}{2}, \quad (13.108)$$

the quaternionic conjugate of R .

The group of rotors is isomorphic to the group of unimodular quaternions, quaternions $q = a + \iota_1 b_1 + \iota_2 b_2 + \iota_3 b_3$ satisfying $q\tilde{q} = a^2 + b_1^2 + b_2^2 + b_3^2 = 1$. Unimodular quaternions evidently define a unit 3-sphere in the 4-dimensional space of coordinates $\{a, b_1, b_2, b_3\}$. From this it is apparent that the rotor group in 3 dimensions has the geometry of a 3-sphere S^3 .

Exercise 13.8 3D rotation matrices. This exercise is a precursor to Exercise 14.9. The principal message of the exercise is that rotating using matrices is more complicated than rotating using quaternions. Let $\mathbf{b} \equiv \boldsymbol{\gamma}_a b_a$ be a 3D vector, a multivector of grade 1 in the 3D geometric algebra. Use the quaternionic composition rule (13.93) to show that the vector $\mathbf{b}$ transforms under a right-handed rotation by angle θ about unit direction $\mathbf{n} = \boldsymbol{\gamma}_a n_a$ as

$$R: \mathbf{b} \rightarrow R\mathbf{b}\tilde{R} = \mathbf{b} + 2 \sin \frac{\theta}{2} \mathbf{n} \times \left(\cos \frac{\theta}{2} \mathbf{b} + \sin \frac{\theta}{2} \mathbf{n} \times \mathbf{b} \right). \quad (13.109)$$

Here the cross-product $\mathbf{n} \times \mathbf{b}$ denotes the usual vector product, which is dual to the bivector product $\mathbf{n} \wedge \mathbf{b}$, equation (13.26). Suppose that the quaternionic components of the rotor R are $\{w, x, y, z\}$, that is, $R = e^{-\iota_a n_a \theta/2} = w + \iota_1 x + \iota_2 y + \iota_3 z$. Show that the transformation (13.109) is (note that the 3×3 rotation matrix is written to the left of the vector, in accordance with the physics convention that rotations accumulate to the left):

$$R: \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} \rightarrow \begin{pmatrix} w^2 + x^2 - y^2 - z^2 & 2(xy - wz) & 2(zx + wy) \\ 2(xy + wz) & w^2 - x^2 + y^2 - z^2 & 2(yz - wx) \\ 2(zx - wy) & 2(yz + wx) & w^2 - x^2 - y^2 + z^2 \end{pmatrix} \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix}. \quad (13.110)$$

Confirm that the 3×3 rotation matrix on the right hand side of the transformation (13.110) is an orthogonal matrix (its inverse is its transpose) provided that the rotor is unimodular, $R\tilde{R} = 1$, so that $w^2 + x^2 + y^2 + z^2 = 1$. As a simple example, show that the transformation (13.110) in the case of a right-handed rotation by angle θ about the 3-axis (the 12-plane), where $w = \cos(\theta/2)$ and $z = -\sin(\theta/2)$, is

$$R: \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} \rightarrow \begin{pmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix}. \quad (13.111)$$

13.14 Pauli matrices

The multiplication rules of the basis vectors $\boldsymbol{\gamma}_a$ of the 3D geometric algebra are identical to those of the Pauli matrices σ_a used in the theory of nonrelativistic spin- $\frac{1}{2}$ particles.

The **Pauli matrices** form a vector of 2×2 complex (with respect to a scalar quantum-mechanical imaginary i) matrices whose three components are each traceless ($\text{Tr } \sigma_a = 0$), Hermitian ($\sigma_a^\dagger = \sigma_a$), and unitary ($\sigma_a^{-1} = \sigma_a^\dagger$):

$$\sigma_1 \equiv \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 \equiv \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (13.112)$$

The Pauli matrices anticommute with each other

$$\sigma_1\sigma_2 = -\sigma_2\sigma_1 = i\sigma_3, \quad \sigma_2\sigma_3 = -\sigma_3\sigma_2 = i\sigma_1, \quad \sigma_3\sigma_1 = -\sigma_1\sigma_3 = i\sigma_2. \quad (13.113)$$

The particular choice (13.112) of Pauli matrices is conventional but not unique: any three traceless, Hermitian, unitary, anticommuting 2×2 complex matrices will do. The product of the 3 Pauli matrices is i times the unit matrix,

$$\sigma_1\sigma_2\sigma_3 = i \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (13.114)$$

If the scalar 1 in the geometric algebra is identified with the unit 2×2 matrix, and the pseudoscalar I_3 is identified with the imaginary i times the unit matrix, then the 3D geometric algebra is isomorphic to the algebra generated by the Pauli matrices, the **Pauli algebra**, through the mapping

$$1 \leftrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \gamma_a \leftrightarrow \sigma_a, \quad I_3 \leftrightarrow i \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (13.115)$$

The 3D pseudoscalar I_3 commutes with all elements of the 3D geometric algebra.

Concept question 13.9 Properties of Pauli matrices. The Pauli matrices are traceless, Hermitian, unitary, and anticommuting. What do these properties correspond to in the geometric algebra? Are all these properties necessary for the Pauli algebra to be isomorphic to the 3D geometric algebra? Are the properties sufficient?

In 3 dimensions, the rotation group is the group of even, unimodular multivectors of the geometric algebra. The isomorphism (13.115) establishes that the rotation group is isomorphic to the group of complex 2×2 matrices of the form

$$a + i\sigma_a b_a, \quad (13.116)$$

with a, b_a ($a = 1, 3$) real, and with the unimodular condition requiring that $a^2 + b_a b_a = 1$. It is straightforward to check (Exercise 13.10) that the group of such matrices constitutes the group of unitary complex 2×2 matrices of unit determinant, the special unitary group $\text{SU}(2)$. The isomorphisms

$$a + I_3 \gamma_a b_a \leftrightarrow a + \iota_a b_a \leftrightarrow a + i\sigma_a b_a \quad (13.117)$$

have thus established isomorphisms between the group of 3D rotors, the group of unimodular quaternions, and the special unitary group of complex 2×2 matrices

$$\text{3D rotors} \cong \text{unimodular quaternions} \cong \text{SU}(2). \quad (13.118)$$

An isomorphism that maps a group into a set of matrices, such that group multiplication corresponds to ordinary matrix multiplication, is called a **representation** of the group. The representation of the rotation group as 2×2 complex matrices may be termed the **Pauli representation**. The Pauli representation is the lowest dimensional representation of the 3D rotation group. In the Pauli representation, the rotor (13.102) corresponding to a right-handed rotation by angle θ about unit axis n_a is the matrix

$$R = \cos \frac{\theta}{2} - in_a \sigma_a \sin \frac{\theta}{2} . \quad (13.119)$$

Exercise 13.10 Translate a rotor into an element of SU(2). Show that the rotor $R = e^{-in_a n_a \theta/2}$, equation (13.102), corresponding to a right-handed rotation by angle θ about unit axis n_a is equivalent to the special unitary 2×2 matrix

$$R \leftrightarrow \begin{pmatrix} \cos \frac{\theta}{2} - in_3 \sin \frac{\theta}{2} & -(n_2 + in_1) \sin \frac{\theta}{2} \\ (n_2 - in_1) \sin \frac{\theta}{2} & \cos \frac{\theta}{2} + in_3 \sin \frac{\theta}{2} \end{pmatrix} . \quad (13.120)$$

Show that the reverse rotor $\tilde{R}$ is equivalent to the Hermitian conjugate $R^\dagger$ of the corresponding 2×2 matrix. Show that the determinant of the matrix equals $\tilde{R}R$, which is 1.

13.15 Pauli spinors as quaternions, or scaled rotors

Any Pauli spinor φ can be expressed uniquely in the form of a 2×2 matrix $\mathbf{q}$, the Pauli representation of a quaternion q , acting on the spin-up basis element $\epsilon_\uparrow$ (the precise translation between Pauli spinors and quaternions is left as Exercises 13.11 and 13.12):

$$\varphi = \mathbf{q} \epsilon_\uparrow . \quad (13.121)$$

In this section (and in the Exercises) the 2×2 matrix $\mathbf{q}$ is written in boldface to distinguish it from the quaternion q that it represents, but the distinction is not fundamental. A quaternion can always be decomposed into a product $q = \lambda R$ of a real scalar λ and a rotor, or unimodular quaternion, R . The real scalar λ can be taken without loss of generality to be positive, since any minus sign can be absorbed into a rotation by 2π of the rotor R . Thus a Pauli spinor φ can also be expressed as a scaled rotor λR acting on the spin-up basis element $\epsilon_\uparrow$,

$$\varphi = \lambda R \epsilon_\uparrow . \quad (13.122)$$

One is used to thinking of a Pauli spinor as an intrinsically quantum-mechanical object. The mapping (13.121) or (13.122) between Pauli spinors and quaternions or scaled rotors shows that Pauli spinors also

have a classical interpretation: they encode a real amplitude λ , and a rotation R . This provides a mathematical basis for the idea that, through their spin, fundamental particles “know” about the rotational structure of space.

The isomorphism between the vector spaces of Pauli spinors and quaternions does not extend to multiplication; that is, the product of two Pauli spinors φ_1 and φ_2 equivalent to the complex 2×2 matrices $\mathbf{q}_1$ and $\mathbf{q}_2$ does not equal the Pauli spinor equivalent to the product $\mathbf{q}_1 \mathbf{q}_2$. The problem is that the Pauli representation of a Pauli spinor φ is a column vector, and two column vectors cannot be multiplied. The question of how to multiply spinors is deferred to Chapter 40 on the supergeometric algebra.

Meanwhile, it is possible to multiply a row spinor and a column spinor. The spinor $\tilde{\varphi}$ reverse to the spinor (13.121) is defined to be the row spinor

$$\tilde{\varphi} \equiv \boldsymbol{\epsilon}_\uparrow^\top \tilde{\mathbf{q}}, \quad (13.123)$$

where $\tilde{\mathbf{q}}$ is the matrix representation of the quaternionic conjugate $\tilde{q}$ of the quaternion q , and $\boldsymbol{\epsilon}_\uparrow^\top$ is the transpose of the column spinor $\boldsymbol{\epsilon}_\uparrow$, which is the row spinor

$$\boldsymbol{\epsilon}_\uparrow^\top = \begin{pmatrix} 1 & 0 \end{pmatrix}. \quad (13.124)$$

The scalar product $\tilde{\varphi}\varphi$ is real and positive, equation (13.133). It is legitimate to multiply a row spinor $\tilde{\varphi}$ by a column spinor χ , yielding a complex number. The product $\tilde{\varphi}\chi$ is a scalar under spatial rotations,

$$R: \tilde{\varphi}\chi \rightarrow \tilde{\varphi}\tilde{R}R\chi = \tilde{\varphi}\chi, \quad (13.125)$$

and therefore provides a viable definition of a scalar product of Pauli spinors. The problem of defining a scalar product of Pauli spinors is resumed in §40.6.

Exercise 13.11 Translate a Pauli spinor into a quaternion. Given any Pauli spinor

$$\varphi \equiv \varphi^\uparrow \boldsymbol{\epsilon}_\uparrow + \varphi^\downarrow \boldsymbol{\epsilon}_\downarrow = \begin{pmatrix} \varphi^\uparrow \\ \varphi^\downarrow \end{pmatrix}, \quad (13.126)$$

show that the corresponding real quaternion q , and the equivalent 2×2 complex matrix $\mathbf{q}$ in the Pauli representation (13.112), such that $\varphi = \mathbf{q} \boldsymbol{\epsilon}_\uparrow$, are

$$q = \{\text{Re } \varphi^\uparrow, \text{Im } \varphi^\downarrow, -\text{Re } \varphi^\downarrow, \text{Im } \varphi^\uparrow\} \leftrightarrow \mathbf{q} = \begin{pmatrix} \varphi^\uparrow & -\varphi^{\downarrow*} \\ \varphi^\downarrow & \varphi^{\uparrow*} \end{pmatrix}. \quad (13.127)$$

Show that the quaternion conjugate $\tilde{q}$ and the equivalent 2×2 matrix $\tilde{\mathbf{q}}$ in the Pauli representation are

$$\tilde{q} = \{\text{Re } \varphi^\uparrow, -\text{Im } \varphi^\downarrow, \text{Re } \varphi^\downarrow, -\text{Im } \varphi^\uparrow\} \leftrightarrow \tilde{\mathbf{q}} = \begin{pmatrix} \varphi^{\uparrow*} & \varphi^{\downarrow*} \\ -\varphi^\downarrow & \varphi^\uparrow \end{pmatrix}. \quad (13.128)$$

Conclude that the reverse matrix $\tilde{\mathbf{q}}$ equals its Hermitian conjugate, $\tilde{\mathbf{q}} = \mathbf{q}^\dagger$, and that the reverse Pauli spinor $\tilde{\varphi}$ defined by equation (13.123) is

$$\tilde{\varphi} \equiv \boldsymbol{\epsilon}_\uparrow^\top \tilde{\mathbf{q}} = \boldsymbol{\epsilon}_\uparrow^\top \mathbf{q}^\dagger = \begin{pmatrix} \varphi^{\uparrow*} & \varphi^{\downarrow*} \end{pmatrix} = \varphi^\dagger. \quad (13.129)$$

Exercise 13.12 Translate a quaternion into a Pauli spinor. Show that the quaternion $q \equiv w + ix + jy + kz$ is equivalent in the Pauli representation (13.112) to the 2×2 matrix $\mathbf{q}$

$$q = \{w, x, y, z\} \leftrightarrow \mathbf{q} = \begin{pmatrix} w + iz & ix + y \\ ix - y & w - iz \end{pmatrix}. \quad (13.130)$$

Conclude that the Pauli spinor $\varphi = \mathbf{q} \epsilon_{\uparrow}$ corresponding to the quaternion q is

$$\varphi \equiv \mathbf{q} \epsilon_{\uparrow} = \begin{pmatrix} w + iz \\ ix - y \end{pmatrix}, \quad (13.131)$$

and that the reverse spinor $\tilde{\varphi}$ defined by equation (13.123) is

$$\tilde{\varphi} = \varphi^{\dagger} = \epsilon_{\uparrow}^{\top} \mathbf{q}^{\dagger} = \begin{pmatrix} w - iz & -ix - y \end{pmatrix}. \quad (13.132)$$

Hence conclude that $\tilde{\varphi}\varphi$ is the real positive scalar magnitude squared $\lambda^2 = \tilde{q}q$ of the quaternion q ,

$$\tilde{\varphi}\varphi = \varphi^{\dagger}\varphi = \tilde{q}q = \lambda^2, \quad (13.133)$$

with

$$\lambda^2 = w^2 + x^2 + y^2 + z^2. \quad (13.134)$$

Exercise 13.13 Can a Pauli spinor be rotated into its complex conjugate? Can a Pauli spinor φ be rotated into its complex conjugate φ^* ?

Solution. Yes. The question is, does there exist a rotor R such that $R\varphi = \varphi^*$? If q and q^* are the quaternions equivalent to φ and φ^* , then

$$R = q^* q^{-1}. \quad (13.135)$$

More generally, a Pauli spinor may be rotated into any other Pauli spinor of the same modulus.

13.16 Spin axis

In the Pauli representation, the spinor basis elements ϵ_a are eigenvectors of the Pauli operator σ_3 with eigenvalues ± 1 ,

$$\sigma_3 \epsilon_{\uparrow} = +\epsilon_{\uparrow}, \quad \sigma_3 \epsilon_{\downarrow} = -\epsilon_{\downarrow}. \quad (13.136)$$

The **spin axis** of a Pauli spinor φ is defined to be the direction along which the Pauli spinor is pure up. In the Pauli representation, the spin axis of the spin-up basis spinor $\epsilon_{\uparrow}$ is the positive 3-axis, while the spin axis of the spin-down basis spinor $\epsilon_{\downarrow}$ is the negative 3-axis. The spin axis of a Pauli spinor $\varphi = \lambda R \epsilon_{\uparrow}$ is the unit direction $\{n_1, n_2, n_3\}$ of the rotated 3-axis, given by

$$\sigma_a n_a = R \sigma_3 \tilde{R}. \quad (13.137)$$

Equation (13.137) is confirmed by the fact that $\sigma_a n_a$ has eigenvalue +1 acting on φ :

$$\sigma_a n_a \varphi = (R\sigma_3 \tilde{R}) (\lambda R \epsilon_{\uparrow}) = \lambda R \sigma_3 \epsilon_{\uparrow} = \lambda R \epsilon_{\uparrow} = \varphi . \quad (13.138)$$

Exercise 13.14 Orthonormal eigenvectors of the spin operator. Show that, in the Pauli representation, the orthonormal eigenvectors $\epsilon_{\uparrow n}$ and $\epsilon_{\downarrow n}$ of the spin operator $\sigma_a n_a$ projected along the unit direction $\{n_1, n_2, n_3\}$ are

$$\epsilon_{\uparrow n} = \frac{1}{\sqrt{2(1+n_3)}} \begin{pmatrix} 1+n_3 \\ n_1 + in_2 \end{pmatrix} , \quad \epsilon_{\downarrow n} = \frac{1}{\sqrt{2(1-n_3)}} \begin{pmatrix} -1+n_3 \\ n_1 + in_2 \end{pmatrix} . \quad (13.139)$$

The spacetime algebra

The **spacetime algebra** is the geometric algebra in Minkowski space. This Chapter is restricted to the case of 4-dimensional Minkowski space, but the formalism generalizes to any number of dimensions where some of the dimensions are timelike and the others are spacelike. Happily, the elegant formalism of the geometric algebra carries through to the spacetime algebra. See Exercise 41.5 for the general case of K space dimensions and M time dimensions.

14.1 Spacetime algebra

Let γ_m ($m = 0, 1, 2, 3$) denote an orthonormal basis of spacetime, with γ_0 representing the time axis, and γ_a ($a = 1, 2, 3$) the spatial axes. Geometric multiplication in the spacetime algebra is defined by

$$\boxed{\gamma_m \gamma_n = \gamma_m \cdot \gamma_n + \gamma_m \wedge \gamma_n} , \quad (14.1)$$

just as in the geometric algebra, equation (13.5). The key difference between the spacetime basis γ_m and Euclidean bases is that scalar products of the basis vectors γ_m form the Minkowski metric η_{mn} ,

$$\gamma_m \cdot \gamma_n = \eta_{mn} , \quad (14.2)$$

whereas scalar products of Euclidean basis elements γ_a formed the unit matrix, $\gamma_a \cdot \gamma_b = \delta_{ab}$, equation (39.70). In less abbreviated form, equations (14.1) state that the geometric product of each basis element with itself is

$$-\gamma_0^2 = \gamma_1^2 = \gamma_2^2 = \gamma_3^2 = 1 , \quad (14.3)$$

while geometric products of different basis elements γ_m anticommute

$$\gamma_m \gamma_n = -\gamma_n \gamma_m = \gamma_m \wedge \gamma_n \quad (m \neq n) . \quad (14.4)$$

In the Dirac theory of relativistic spin- $\frac{1}{2}$ particles, §14.7, the Dirac γ -matrices are required to satisfy

$$\{\gamma_m, \gamma_n\} = 2\eta_{mn} \quad (14.5)$$

where $\{\}$ denotes the anticommutator, $\{\gamma_m, \gamma_n\} \equiv \gamma_m \gamma_n + \gamma_n \gamma_m$. The multiplication rules (14.5) for the Dirac γ -matrices are the same as those for geometric multiplication in the spacetime algebra, equations (14.3) and (14.4).

A 4-vector $\mathbf{a}$, a multivector of grade 1 in the geometric algebra of spacetime, is

$$\mathbf{a} = a^m \gamma_m = a^0 \gamma_0 + a^1 \gamma_1 + a^2 \gamma_2 + a^3 \gamma_3 . \quad (14.6)$$

Such a 4-vector $\mathbf{a}$ would be denoted $\not{a}$ in the Feynman slash notation. The product of two 4-vectors $\mathbf{a}$ and $\mathbf{b}$ is

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b} = a^m b^n \gamma_m \cdot \gamma_n + a^m b^n \gamma_m \wedge \gamma_n = a^m b^n \eta_{mn} + \frac{1}{2} a^m b^n [\gamma_m, \gamma_n] . \quad (14.7)$$

It is convenient to denote three of the six bivectors of the spacetime algebra by σ_a ,

$$\sigma_a \equiv \gamma_0 \gamma_a \quad (a = 1, 2, 3) . \quad (14.8)$$

The symbol σ_a is used because the algebra of bivectors σ_a is isomorphic to the algebra of Pauli matrices σ_a . The pseudoscalar, the highest grade basis element of the spacetime algebra, is denoted I

$$\gamma_0 \gamma_1 \gamma_2 \gamma_3 = \sigma_1 \sigma_2 \sigma_3 = I . \quad (14.9)$$

The pseudoscalar I satisfies

$$I^2 = -1 , \quad I \gamma_m = -\gamma_m I , \quad I \sigma_a = \sigma_a I . \quad (14.10)$$

The basis elements of the 4-dimensional spacetime algebra are then

$$\begin{array}{ccccc} 1, & \gamma_m, & \sigma_a, I \sigma_a, & I \gamma_m, & I, \\ 1 \text{ scalar} & 4 \text{ vectors} & 6 \text{ bivectors} & 4 \text{ pseudovectors} & 1 \text{ pseudoscalar} \end{array} \quad (14.11)$$

forming a linear space of dimension $1 + 4 + 6 + 4 + 1 = 16 = 2^4$. The reverse is defined in the usual way, equation (13.13), leaving unchanged multivectors of grade (mod 4) 0 or 1, and changing the sign of multivectors of grade (mod 4) 2 or 3.

$$\tilde{1} = 1 , \quad \tilde{\gamma}_m = \gamma_m , \quad \tilde{\sigma}_a = -\sigma_a , \quad \tilde{I \sigma}_a = -I \sigma_a , \quad \tilde{I \gamma}_m = -I \gamma_m , \quad \tilde{I} = I . \quad (14.12)$$

In the 3D geometric algebra a bivector was also a rotor, satisfying $R \tilde{R} = 1$, but in the 4D spacetime algebra only the spatial bivectors $I \sigma_a$ are rotors, satisfying $I \sigma_a \tilde{I \sigma}_a = 1$. The boost bivectors satisfy $\sigma_a \tilde{\sigma}_a = -1$ not 1, so are not rotors. Nevertheless, if $\theta \sigma_a$ is a boost bivector, then its exponential $R \equiv e^{-\theta \sigma_a / 2}$ is a rotor,

$$R = e^{-\theta \sigma_a / 2} = 1 - (\theta/2) \sigma_a + \frac{(\theta/2)^2}{2!} - \frac{(\theta/2)^3}{3!} \sigma_a + \dots = \cosh(\theta/2) - \sigma_a \sinh(\theta/2) , \quad (14.13)$$

since its inverse is indeed its reverse $\tilde{R} = e^{-\theta \tilde{\sigma}_a / 2} = e^{\theta \sigma_a / 2} = \cosh(\theta/2) + \sigma_a \sinh(\theta/2)$.

The mapping

$$\gamma_a^{(3)} \leftrightarrow \sigma_a \quad (a = 1, 2, 3) \quad (14.14)$$

(the superscript $^{(3)}$ distinguishes the 3D basis vectors from the 4D spacetime basis vectors) defines an isomorphism between the 8-dimensional geometric algebra (13.3) of 3 spatial dimensions and the 8-dimensional

even spacetime subalgebra. Among other things, the isomorphism (14.14) implies the equivalence of the 3D spatial pseudoscalar I_3 and the 4D spacetime pseudoscalar I ,

$$I_3 \leftrightarrow I , \quad (14.15)$$

since $I_3 = \gamma_1^{(3)} \gamma_2^{(3)} \gamma_3^{(3)}$ and $I = \sigma_1 \sigma_2 \sigma_3$.

14.2 Complex quaternions

A **complex quaternion** (also called a **biquaternion** by W. R. Hamilton) is a quaternion

$$q = a + \mathbf{b} = a + \iota_a b_a , \quad (14.16)$$

in which the four coefficients a, b_a ($a = 1, 2, 3$) are each complex numbers

$$a = a_R + Ia_I , \quad b_a = b_{a,R} + Ib_{a,I} . \quad (14.17)$$

The imaginary I is taken to commute with each of the quaternionic imaginaries ι_a . The choice of symbol I is deliberate: in the isomorphism (14.33) between the even spacetime algebra and complex quaternions, the commuting imaginary I is isomorphic to the spacetime pseudoscalar I .

All of the equations in §13.12 on real quaternions remain valid without change, including the multiplication, conjugation, and inversion formulae (13.93)–(13.98). The quaternionic conjugate $\tilde{q}$ of a complex quaternion $q \equiv a + \mathbf{b}$ is conjugated with respect to the quaternionic imaginaries ι_a , but the complex coefficients a and b_a are *not* conjugated with respect to the complex imaginary I ,

$$\tilde{q} = a + \tilde{\mathbf{b}} = a - \mathbf{b} = a - \iota_a b_a . \quad (14.18)$$

The **modulus** $|q|$ of a complex quaternion $q \equiv a + \mathbf{b}$,

$$|q| = (\tilde{q}q)^{1/2} = (q\tilde{q})^{1/2} = (a^2 + \mathbf{b} \cdot \mathbf{b})^{1/2} = (a^2 + b_a b_a)^{1/2} , \quad (14.19)$$

is a complex number, not a real number. The name modulus to denote $|q|$ is preferred over magnitude, to avoid confusion with the magnitude of a complex number. A quaternion is said to be **unimodular** if its modulus is 1,

$$\tilde{q}q = 1 . \quad (14.20)$$

The unimodular condition (14.20) is a complex condition, stating that the real and imaginary (with respect to I) parts of $\tilde{q}q$ are respectively 1 and 0.

The complex conjugate q^* of the complex quaternion is (the star symbol $*$ is used for complex conjugation with respect to the pseudoscalar I , to distinguish it from the asterisk symbol $*$ for complex conjugation with respect to the scalar quantum-mechanical imaginary i)

$$q^* = a^* + \mathbf{b}^* = a^* + \iota_a b_a^* , \quad (14.21)$$

in which the complex coefficients a and b_a are conjugated with respect to the imaginary I , but the quaternionic imaginaries ι_a are *not* conjugated.

A non-zero complex quaternion can have zero modulus (unlike a real quaternion), in which case it is **null**. The null condition

$$\tilde{q}q = a^2 + b_a b_a = 0 \quad (14.22)$$

is a complex condition. The product of two null complex quaternions is a null quaternion. Under multiplication, null quaternions form a 6-dimensional subsemigroup (not a subgroup, because null quaternions do not have inverses) of the 8-dimensional semigroup of complex quaternions.

Exercise 14.1 Null complex quaternions. Show that any non-trivial null complex quaternion q can be written uniquely in the form

$$q = p(1 + I\mathbf{n}) = p(1 + I\iota_a n_a) , \quad (14.23)$$

where p is a real quaternion, and $\mathbf{n} = \iota_a n_a$ is a real unimodular vector quaternion, with real components $\{n_1, n_2, n_3\}$ satisfying $n_a n_a = 1$. Equivalently,

$$q = (1 + I\mathbf{n}')p = (1 + I\iota_a n'_a)p , \quad (14.24)$$

where $\mathbf{n}'$ is the real unimodular vector quaternion

$$\mathbf{n}' = \frac{p\mathbf{n}\tilde{p}}{|p|^2} , \quad (14.25)$$

with real components $\{n'_1, n'_2, n'_3\}$ satisfying $n'_a n'_a = 1$.

Solution. Write the null quaternion q as

$$q = p + Ir \quad (14.26)$$

where p and r are real quaternions, both of which must be non-zero if q is non-trivial. Then equation (14.23) is true with

$$\mathbf{n} = \iota_a n_a = \frac{\tilde{p}r}{|p|^2} . \quad (14.27)$$

The null condition is $\tilde{q}q = 0$. The vanishing of the real part, $\text{Re}(\tilde{q}q) = \tilde{p}p - \tilde{r}r = 0$, shows that $|p|^2 = |r|^2$. The vanishing of the imaginary (I) part, $\text{Im}(\tilde{q}q) = \tilde{p}r + \tilde{r}p = \tilde{p}r + \tilde{\tilde{p}r} = 0$ shows that the $\tilde{p}r$ must be a pure quaternionic imaginary, since the quaternionic conjugate of $\tilde{p}r$ is minus itself, so $\tilde{p}r/|p|^2$ must be of the form $\mathbf{n} = \iota_a n_a$. Its squared modulus $\mathbf{n}\tilde{\mathbf{n}} = n_a n_a = \tilde{p}r\tilde{r}p/|p|^4 = 1$ is unity, so $\mathbf{n}$ is a unimodular 3-vector quaternion. It follows immediately from the manner of construction that the expression (14.23) is unique, as long as q is non-trivial.

Exercise 14.2 Nilpotent complex quaternions. An object whose square is zero is said to be **nilpotent**. Show that a complex quaternion of the form

$$\mathbf{q} = \iota_a q_a \quad \text{with} \quad \mathbf{q} \cdot \mathbf{q} = q_a q_a = 0 \quad (14.28)$$

is nilpotent,

$$\mathbf{q}^2 = 0 . \quad (14.29)$$

Prove that a nilpotent complex quaternion must take the form (14.28). The set of nilpotent complex quaternions forms a 4-dimensional subspace of complex quaternions, since the complex condition $q_a q_a = 0$ eliminates 2 of the 6 degrees of freedom in the quaternionic components q_a . The product of two nilpotent complex quaternions is not necessarily nilpotent, so the nilpotent set does not form a semigroup. The set of nilpotent complex quaternions consists of the subset of null complex quaternions that are purely quaternionic.

14.3 Lorentz transformations and complex quaternions

Lorentz transformations are rotations of spacetime. The rotor group of spacetime rotations in 3+1 dimensions is, as usual, the Lie group generated by the Lie algebra of bivectors. The rotor group in 3+1 dimensions is called $\text{Spin}(3,1)$.

The basis elements of the even spacetime algebra are

$$\begin{array}{ccc} 1, & \boldsymbol{\sigma}_a, I\boldsymbol{\sigma}_a, & I, \\ 1 \text{ scalar} & 6 \text{ bivectors} & 1 \text{ pseudoscalar} \end{array} \quad (14.30)$$

forming a linear space of dimension $1 + 6 + 1 = 8$ over the real numbers. However, it is more elegant to treat the even spacetime algebra as a linear space of dimension $8/2 = 4$ over complex numbers of the form $\lambda = \lambda_R + I\lambda_I$. The pseudoscalar I qualifies as an imaginary because $I^2 = -1$, and because it commutes with all elements of the even spacetime algebra. It is convenient to take the basis elements of the even spacetime algebra over the complex numbers to be

$$\begin{array}{ccc} 1, & I\boldsymbol{\sigma}_a, & \\ 1 \text{ scalar} & 3 \text{ bivectors} & \end{array} \quad (14.31)$$

forming a linear space of dimension $1 + 3 = 4$. The reason for choosing $I\boldsymbol{\sigma}_a$ rather than $\boldsymbol{\sigma}_a$ as the elements of the basis (14.31) is that the basis $\{1, I\boldsymbol{\sigma}_a\}$ is equivalent to the basis (13.99) of the even algebra of 3-dimensional Euclidean space through the isomorphism (14.14) and (14.15). This basis in turn is equivalent to the quaternionic basis $\{1, \iota_a\}$ through the isomorphism (13.105):

$$I\boldsymbol{\sigma}_a \leftrightarrow I_3 \boldsymbol{\gamma}_a^{(3)} \leftrightarrow \iota_a \quad (a = 1, 2, 3) . \quad (14.32)$$

In other words, the even spacetime algebra is isomorphic to the algebra of quaternions with complex coefficients:

$$a + I\boldsymbol{\sigma}_a b_a \leftrightarrow a + \iota_a b_a \quad (14.33)$$

where $a = a_R + Ia_I$ is a complex number, and $b_a \equiv b_{a,R} + Ib_{a,I}$ is a triple of complex numbers.

The isomorphism (14.33) between even elements of the spacetime algebra and complex quaternions implies

that the group $\text{Spin}(3, 1)$ of Lorentz rotors, which are unimodular elements of the even spacetime algebra, is isomorphic to the group of unimodular complex quaternions

$$\text{spacetime rotors} \cong \text{unimodular complex quaternions} . \quad (14.34)$$

In §13.13 it was found that the group of 3D spatial rotors is isomorphic to the group of unimodular real quaternions. Thus Lorentz transformations are mathematically equivalent to complexified spatial rotations.

The Lorentz rotor that produces a rotation by complex angle θ about the unimodular complex direction n_a is, according to equation (13.49),

$$R = e^{-\theta/2} = e^{-\mathbf{n}\theta/2} = \cos \frac{\theta}{2} - \mathbf{n} \sin \frac{\theta}{2} , \quad (14.35)$$

generalizing the 3D rotor (13.102). Here θ is a bivector

$$\theta = \mathbf{n}\theta = I\sigma_a n_a \theta , \quad (14.36)$$

whose modulus is the complex angle $(\tilde{\theta}\theta)^{1/2} = \theta \equiv \theta_R + I\theta_I$, and whose direction is the unimodular complex bivector $\mathbf{n} = \mathbf{n}_R + I\mathbf{n}_I$. The unimodular condition $\tilde{\mathbf{n}}\mathbf{n} = 1$ on $\mathbf{n}$ is equivalent to the complex condition $n_a n_a = 1$ on the complex components $n_a \equiv \{n_1, n_2, n_3\}$. The real and imaginary parts of the unimodular condition imply the two conditions

$$n_{a,R} n_{a,R} - n_{a,I} n_{a,I} = 1 , \quad 2 n_{R,a} n_{I,a} = 0 . \quad (14.37)$$

The complex angle θ has 2 degrees of freedom, while the complex unimodular bivector $\mathbf{n}$ has 4 degrees of freedom, so the Lorentz rotor R has 6 degrees of freedom, which is the correct number of degrees of freedom of the group of Lorentz transformations.

With the equivalence (14.32), the Lorentz rotor R given by equation (14.35) can be reinterpreted as a complex quaternion, with θ the complex quaternion

$$\theta = \mathbf{n}\theta = \imath_a n_a \theta , \quad (14.38)$$

whose complex modulus is $\theta = |\theta| \equiv (\tilde{\theta}\theta)^{1/2}$ and whose complex unimodular direction is $\mathbf{n} \equiv \imath_a n_a$. The associated reverse rotor $\tilde{R}$ is

$$\tilde{R} = e^{\theta/2} = e^{\mathbf{n}\theta/2} = \cos \frac{\theta}{2} + \mathbf{n} \sin \frac{\theta}{2} \quad (14.39)$$

the quaternionic conjugate of R . Note that θ and $\mathbf{n}$ in equation (14.39) are *not* conjugated with respect to the imaginary I . The sine and cosine of the complex angle θ appearing in equations (14.35) and (14.39) are related to its real and imaginary parts in the usual way,

$$\cos \frac{\theta}{2} = \cos \frac{\theta_R}{2} \cosh \frac{\theta_I}{2} - I \sin \frac{\theta_R}{2} \sinh \frac{\theta_I}{2} , \quad \sin \frac{\theta}{2} = \sin \frac{\theta_R}{2} \cosh \frac{\theta_I}{2} + I \cos \frac{\theta_R}{2} \sinh \frac{\theta_I}{2} . \quad (14.40)$$

For the case of a pure spatial rotation, the angle $\theta = \theta_R$ and axis $\mathbf{n} = \mathbf{n}_R$ in the rotor (14.35) are both

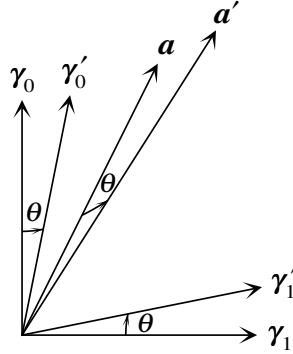


Figure 14.1 Lorentz boost of a vector $\mathbf{a}$ by rapidity θ in the $\gamma_0\gamma_1$ -plane. See Exercise 14.3.

real. The rotor corresponding to a pure spatial rotation by angle θ_R right-handedly about unit real axis $\mathbf{n}_R \equiv I\sigma_a n_{a,R} = \iota_a n_{a,R}$ is the real quaternion

$$R = e^{-\mathbf{n}_R \theta_R/2} = \cos \frac{\theta_R}{2} - \mathbf{n}_R \sin \frac{\theta_R}{2} . \quad (14.41)$$

A Lorentz boost is a change of velocity in some direction, without any spatial rotation, and represents a rotation of spacetime about some time-space plane. For example, a Lorentz boost along the γ_1 -axis (the x -axis) is a rotation of spacetime in the $\gamma_0\gamma_1$ -plane (the tx -plane). In the case of a pure Lorentz boost, the angle $\theta = I\theta_I$ is pure imaginary, but the axis $\mathbf{n} = \mathbf{n}_R$ remains pure real (alternatively, the angle is pure real and the axis is pure imaginary). The rotor corresponding to a boost by rapidity θ_I , or equivalently by velocity $v = \tanh \theta_I$, in unit real direction $\mathbf{n}_R \equiv I\sigma_a n_{a,R} = \iota_a n_{a,R}$ is the complex quaternion

$$R = e^{-I\mathbf{n}_R \theta_I/2} = \cosh \frac{\theta_I}{2} - I\mathbf{n}_R \sinh \frac{\theta_I}{2} . \quad (14.42)$$

Exercise 14.3 Lorentz boost. A Lorentz boost by rapidity $\theta = \operatorname{atanh} v$ along the γ_1 -axis (x -axis) (that is, a rotation in the $\gamma_0\gamma_1$ -plane) is given by the Lorentz rotor

$$R = e^{-I\mathbf{n}_1 \theta/2} = \cosh \frac{\theta}{2} + \gamma_0 \wedge \gamma_1 \sinh \frac{\theta}{2} . \quad (14.43)$$

Confirm that the Lorentz boost transforms the axes γ_m as

$$R : \gamma_0 \rightarrow R\gamma_0\tilde{R} = \gamma_0 \cosh \theta + \gamma_1 \sinh \theta , \quad (14.44a)$$

$$R : \gamma_1 \rightarrow R\gamma_1\tilde{R} = \gamma_1 \cosh \theta + \gamma_0 \sinh \theta , \quad (14.44b)$$

$$R : \gamma_a \rightarrow R\gamma_a\tilde{R} = \gamma_a \quad (a \neq 0, 1) . \quad (14.44c)$$

The boost is illustrated in Figure 14.1.

Exercise 14.4 Factor a Lorentz rotor into a boost and a rotation. Factor a general Lorentz rotor $R = e^{-\boldsymbol{\gamma}_a n_a \theta/2}$ into the product LU of a pure spatial rotation U followed by a pure Lorentz boost L . Do the two factors commute?

Solution. Expand the rotor R as

$$R = p + Iq \quad (14.45)$$

where p and q are real quaternions. Then R can be expressed as the composition of a pure spatial rotation U followed by a pure Lorentz boost L

$$R = LU \quad (14.46)$$

in which

$$U = \frac{p}{|p|}, \quad L = |p| + I \frac{q\tilde{p}}{|p|} \quad (14.47)$$

where $|p| = (\tilde{p}p)^{1/2}$ is the (real) absolute value of the real quaternion p . It is straightforward to check that U and L satisfy the requirements to be pure spatial and boost rotors. The spatial rotor U is by construction unimodular, $U\tilde{U} = 1$, and it follows that the boost rotor $L = R\tilde{U}$ is also unimodular, since R is unimodular. The spatial rotor U is a real quaternion, and therefore satisfies the form (14.41) of a pure spatial rotation. The real part $|p|$ of the boost rotor L is pure real, while the imaginary part $q\tilde{p}/|p|$ is a pure quaternionic imaginary, since unimodularity $R\tilde{R} = 1$ implies that $\text{Im}(R\tilde{R}) = q\tilde{p} + p\tilde{q} = q\tilde{p} + \tilde{q}p = 0$, i.e. the quaternionic conjugate of $q\tilde{p}$ is minus itself. Thus L satisfies the form (14.42) of a pure Lorentz boost.

The factors U and L commute if the boost and rotation axes are in the same direction, but not in general. The expression for the rotor R as the composition of a Lorentz boost followed by a spatial rotation, the opposite order to (14.46), is

$$R = UL' \quad (14.48)$$

where U is the same spatial rotor as before, but the boost rotor L' is

$$L' = |p| + I \frac{\tilde{p}q}{|p|} = \tilde{U}LU \quad (14.49)$$

whose real part $|p|$ is the same as for L , but whose imaginary part $\tilde{p}q/|p|$ differs in direction, though not magnitude, from that of L .

Exercise 14.5 Topology of the group of Lorentz rotors. Show that the geometry of the group of Lorentz rotors is the product of the geometries of the spatial rotation group and the boost group, which is a 3-sphere times Euclidean 3-space, $S^3 \times \mathbb{R}^3$.

14.4 Spatial inversion (P) and Time reversal (T)

Spatial inversion, or P for parity, is the operation of reflecting a (single) spatial direction, $\boldsymbol{\gamma}_a \rightarrow -\boldsymbol{\gamma}_a$. Spatial inversion leaves the scalar product of orthonormal vectors unchanged. A rotation in N spatial dimensions

can be represented by a matrix in the orthogonal group $O(N)$ of matrices satisfying the condition that their inverses are their transposes, $O^{-1} = O^T$. Since transposing a matrix leaves its determinant unchanged, orthogonal matrices have squared determinant equal to 1. The orthogonal group $O(N)$ thus splits into two disconnected pieces, **proper** and **improper** rotations represented by orthogonal matrices of determinant respectively $+1$ and -1 . The subgroup group of proper rotations is designated $SO(N)$, the S signifying Special, meaning matrices of determinant 1.

Inversion of one spatial direction can be represented by a diagonal orthogonal matrix with one of its diagonal elements equal to -1 and the remainder all 1. Thus spatial inversion is a discrete transformation of the geometric algebra, which splits the geometric algebra into two disconnected parts that cannot be transformed into each other by any continuous rotation.

Inversion may be accomplished by reflecting through any odd number of spatial axes. In spacetimes with an odd number of spatial dimensions, as here (where there are 3 spatial dimensions), spatial inversion may be accomplished by reflecting all spatial vector basis elements $\gamma_a \rightarrow -\gamma_a$, while keeping the time vector basis element γ_0 unchanged. This results in $\sigma_a \rightarrow -\sigma_a$ and $I \rightarrow -I$. The equivalence $I\sigma_a \leftrightarrow \iota_a$ means that the quaternionic imaginaries ι_a are unchanged. Thus, if multivectors in the spacetime algebra are written as linear combinations of products of γ_0 , ι_a , and I , then spatial inversion P corresponds to the transformation

$$P: \gamma_0 \rightarrow \gamma_0, \quad \iota_a \rightarrow \iota_a, \quad I \rightarrow -I. \quad (14.50)$$

In other words spatial inversion may be accomplished by the rule, take the complex conjugate (with respect to I) of a multivector.

Time reversal, or T , is the operation of reversing the time direction while keeping all spatial directions unchanged. Time reversal, like spatial inversion, leaves the scalar product of orthonormal vectors unchanged. Time reversal cannot be accomplished by any continuous Lorentz transformation starting from the unit element, nor can it be accomplished by spatial inversion accompanied by any continuous Lorentz transformation starting from the unit element. Thus the Lorentz group contains 4 disconnected components that cannot be transformed into each other by any continuous Lorentz transformation starting from the unit element. The normal and reversed time components of the Lorentz group are sometimes called respectively **orthochronous** and **antichronous**.

Time reversal may be accomplished by reflecting the time vector basis element $\gamma_0 \rightarrow -\gamma_0$, while keeping the spatial vector basis elements γ_a unchanged. As with spatial inversion, this results in $\sigma_a \rightarrow -\sigma_a$ and $I \rightarrow -I$, which keeps $I\sigma_a$ hence ι_a unchanged. If multivectors in the spacetime algebra are written as linear combinations of products of γ_0 , ι_a , and I , then time inversion T corresponds to the transformation

$$T: \gamma_0 \rightarrow -\gamma_0, \quad \iota_a \rightarrow \iota_a, \quad I \rightarrow -I. \quad (14.51)$$

For any multivector, time inversion corresponds to the instruction to flip γ_0 and take the complex conjugate (with respect to I).

The combined operation PT of inverting both space and time directions corresponds to

$$PT: \gamma_0 \rightarrow -\gamma_0, \quad \iota \rightarrow \iota, \quad I \rightarrow I. \quad (14.52)$$

For any multivector, spacetime inversion corresponds to the instruction to flip γ_0 , while keeping ι_a and I unchanged.

14.5 How to implement Lorentz transformations on a computer

The advantages of quaternions for implementing spatial rotations are well-known to 3D game programmers. Compared to standard rotation matrices, quaternions offer increased speed and require less storage, and their algebraic properties simplify interpolation and splining. Complex quaternions retain similar advantages for implementing Lorentz transformations. They are fast, compact, and straightforward to interpolate or spline (Exercises 14.6 and 14.8). Moreover, since complex quaternions contain real quaternions, Lorentz transformations can be implemented simply as an extension of spatial rotations in 3D programs that use quaternions to implement spatial rotations.

Lorentz rotors, 4-vectors, spacetime bivectors, and spinors (spin- $\frac{1}{2}$ objects) can all be implemented as complex quaternions. A complex quaternion

$$q = w + \iota_1 x + \iota_2 y + \iota_3 z \quad (14.53)$$

with complex coefficients w, x, y, z (so $w = w_R + I w_I$, etc.) can be stored as the 8-component object

$$q = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{Bmatrix} . \quad (14.54)$$

Actually, OpenGL and other computer software store the scalar (w) component of a quaternion in the last (fourth) place, but here the scalar components are put in the zeroth position to conform to standard physics convention. The quaternion conjugate $\tilde{q}$ of the quaternion (14.54) is

$$\tilde{q} = \begin{Bmatrix} w_R & -x_R & -y_R & -z_R \\ w_I & -x_I & -y_I & -z_I \end{Bmatrix} , \quad (14.55)$$

while its complex conjugate q^* (with respect to I) is

$$q^* = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ -w_I & -x_I & -y_I & -z_I \end{Bmatrix} . \quad (14.56)$$

A Lorentz rotor R corresponds to a complex quaternion of unit modulus. The unimodular condition $R\tilde{R} = 1$, a complex condition, removes 2 degrees of freedom from the 8 degrees of freedom of complex quaternions, leaving the Lorentz group with 6 degrees of freedom, which is as it should be. Spatial rotations correspond to real unimodular quaternions, and account for 3 of the 6 degrees of freedom of Lorentz transformations. A spatial rotation by angle θ right-handedly about the 1-axis (the x -axis) is the real Lorentz rotor

$$R = e^{-\iota_1 \theta/2} = \cos(\theta/2) - \iota_1 \sin(\theta/2) , \quad (14.57)$$

or, stored as a complex quaternion,

$$R = \begin{Bmatrix} \cos(\theta/2) & -\sin(\theta/2) & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Bmatrix} . \quad (14.58)$$

Note that this is the physics convention, where a right-handed rotation corresponds to $R = e^{-\iota_a n_a \theta/2}$ and rotations accumulate to the left. The convention in OpenGL and other graphics software is that $R = e^{\iota_a n_a \theta/2}$ and rotations accumulate to the right. To change to OpenGL convention, omit the minus sign in equation (14.58). Lorentz boosts account for the remaining 3 of the 6 degrees of freedom of Lorentz transformations. A Lorentz boost by velocity v , or equivalently by rapidity $\theta = \text{atanh}(v)$, along the 1-axis (the x -axis) is the complex Lorentz rotor

$$R = e^{-I\iota_1 \theta/2} = \cosh(\theta/2) - I\iota_1 \sinh(\theta/2) , \quad (14.59)$$

or, stored as a complex quaternion,

$$R = \begin{Bmatrix} \cosh(\theta/2) & 0 & 0 & 0 \\ 0 & -\sinh(\theta/2) & 0 & 0 \end{Bmatrix} . \quad (14.60)$$

Again, this is the physics convention. To change to OpenGL convention, omit the minus sign in equation (14.60). The rule for composing Lorentz transformations is simple: a Lorentz transformation R followed by a Lorentz transformation S is just the product SR of the corresponding complex quaternions. This is the physics convention, where rotations accumulate to the left. In the OpenGL convention, where rotations accumulate to the right, R followed by S is RS .

The inverse of a Lorentz rotor R is its quaternionic conjugate $\tilde{R}$.

Any even multivector $\mathbf{q}$ is equivalent to a complex quaternion by the isomorphism (14.33). According to the usual transformation law (13.56) for multivectors, the rule for Lorentz transforming an even multivector $\mathbf{q}$ is

$$\boxed{R : \mathbf{q} \rightarrow R\mathbf{q}\tilde{R}} \quad (\text{even multivector}) . \quad (14.61)$$

The transformation (14.61) instructs to multiply three complex quaternions R , $\mathbf{q}$, and $\tilde{R}$, a one-line expression in a c++ program. In OpenGL convention, the transformation rule is $\mathbf{q} \rightarrow \tilde{R}\mathbf{q}R$.

As an example of an even multivector, the electromagnetic field $\mathbf{F}$ is a bivector in the spacetime algebra,

$$\mathbf{F} = \frac{1}{2} F^{mn} \gamma_m \wedge \gamma_n , \quad (14.62)$$

the factor of $\frac{1}{2}$ compensating for the double-counting over indices m and n (the $\frac{1}{2}$ could be omitted if the counting were over distinct bivector indices only). The imaginary and real parts of $\mathbf{F}$ constitute the electric and magnetic bivectors $\mathbf{E} = E_a \iota_a$ and $\mathbf{B} = B_a \iota_a$

$$\mathbf{F} = -I(\mathbf{E} + I\mathbf{B}) . \quad (14.63)$$

Under the parity transformation P (14.50), the electric field $\mathbf{E}$ changes sign, whereas the magnetic field $\mathbf{B}$ does not, which is as it should be:

$$P : \mathbf{E} \rightarrow -\mathbf{E} , \quad \mathbf{B} \rightarrow \mathbf{B} . \quad (14.64)$$

In view of the isomorphism (14.33), the electromagnetic field bivector $\mathbf{F}$ can be written as the complex quaternion

$$\mathbf{F} = \left\{ \begin{array}{cccc} 0 & B_1 & B_2 & B_3 \\ 0 & -E_1 & -E_2 & -E_3 \end{array} \right\} . \quad (14.65)$$

According to the rule (14.61), the electromagnetic field bivector $\mathbf{F}$ Lorentz transforms as $\mathbf{F} \rightarrow R\mathbf{F}\tilde{R}$, which is a powerful and elegant way to Lorentz transform the electromagnetic field.

A 4-vector $\mathbf{a} \equiv \gamma_m a^m$ is a multivector of grade 1 in the spacetime algebra. A general odd multivector in the spacetime algebra is the sum of a vector (grade 1) part $\mathbf{a}$ and a pseudovector (grade 3) part $I\mathbf{b} = I\gamma_m b^m$. The odd multivector can be written as the product of the time basis vector γ_0 and an even multivector $\mathbf{q}$

$$\mathbf{a} + I\mathbf{b} = \gamma_0 \mathbf{q} = \gamma_0 (a^0 + I\iota_a a^a - Ib^0 + \iota_a b^a) . \quad (14.66)$$

By the isomorphism (14.33), the even multivector $\mathbf{q}$ is equivalent to the complex quaternion

$$\mathbf{q} = \left\{ \begin{array}{cccc} a^0 & b^1 & b^2 & b^3 \\ -b^0 & a^1 & a^2 & a^3 \end{array} \right\} . \quad (14.67)$$

According to the usual transformation law (13.56) for multivectors, the rule for Lorentz transforming the odd multivector $\gamma_0 \mathbf{q}$ is

$$R : \gamma_0 \mathbf{q} \rightarrow R\gamma_0 \mathbf{q}\tilde{R} = \gamma_0 R^* \mathbf{q}\tilde{R} . \quad (14.68)$$

In the last expression of (14.68), the factor γ_0 has been brought to the left, to be consistent with the convention (14.66) that an odd multivector is γ_0 on the left times an even multivector on the right. Notice that commuting γ_0 through R converts the latter to its complex conjugate R^* (with respect to I), which is true because γ_0 commutes with the quaternionic imaginaries ι_a , but anticommutes with the pseudoscalar I . Thus if the components of an odd multivector are stored as a complex quaternion (14.67), then that complex quaternion $\mathbf{q}$ Lorentz transforms as

$$\boxed{R : \mathbf{q} \rightarrow R^* \mathbf{q}\tilde{R}} \quad (\text{odd multivector}) . \quad (14.69)$$

In OpenGL convention, $\mathbf{q} \rightarrow \tilde{R}^* \mathbf{q} R$. The rule (14.69) again instructs to multiply three complex quaternions R^* , $\mathbf{q}$, and $\tilde{R}$, a one-line expression in a c++ program. The transformation rule (14.69) for an odd multivector encoded as a complex quaternion differs from that (14.61) for an even multivector in that the first factor R is complex conjugated (with respect to I).

A vector $\mathbf{a}$ differs from a pseudovector $I\mathbf{b}$ in that the vector $\mathbf{a}$ changes sign under a parity transformation P whereas the pseudovector $I\mathbf{b}$ does not. However, the behaviour of a pseudovector under a normal Lorentz transformation (which preserves parity) is identical to that of a vector. Thus in practical situations two 4-vectors $\mathbf{a}$ and $\mathbf{b}$ can be encoded into a single complex quaternion (14.67), and Lorentz transformed simultaneously, enabling two transformations to be done for the price of one.

Finally, a Dirac spinor ψ is equivalent to a complex quaternion q (§14.9). It Lorentz transforms as

$$\boxed{R : q \rightarrow Rq} \quad (\text{spinor}) . \quad (14.70)$$

In OpenGL convention, where rotations accumulate to the right instead of left, $q \rightarrow qR$.

Exercise 14.6 Interpolate a Lorentz transformation. Argue that the interpolating Lorentz rotor $R(t)$ that corresponds to uniform rotation and acceleration between initial and final Lorentz rotors R_0 and R_1 as the parameter t varies uniformly from 0 to 1 is

$$R(t) = R_0 \exp [t \ln(R_1/R_0)] . \quad (14.71)$$

Exercise 14.7 Exponential and logarithm of a complex quaternion. What are the (1) exponential and (2) logarithm of a complex quaternion in terms of its components? Address the issue of the multi-valued character of the logarithm.

Solution.

1. **Exponential of a complex quaternion.** Decompose the complex quaternion p into the sum of a complex number w and a complex bivector $\mathbf{n}\theta$ of complex modulus θ and unimodular complex direction $\mathbf{n}$ (satisfying $\tilde{\mathbf{n}}\mathbf{n} = 1$). Then

$$e^p = e^{w+\mathbf{n}\theta} = e^w (\cos \theta + \mathbf{n} \sin \theta) . \quad (14.72)$$

2. **Logarithm of a complex quaternion.** Essentially, reverse the procedure for exponentiation. Denote the logarithm of the complex quaternion q by $\ln q \equiv p \equiv w + \mathbf{n}\theta$. The non-quaternionic part of the logarithm is the complex number w given by the (complex) logarithm of the (complex) modulus of q ,

$$w = \frac{1}{2} \ln(\tilde{q}q) . \quad (14.73)$$

The complex quaternion q scaled to unit modulus is then

$$\frac{q}{\sqrt{\tilde{q}q}} = \cos \theta + \mathbf{n} \sin \theta , \quad (14.74)$$

whose non-quaternionic part $\cos \theta$ defines the (complex) angle θ , and whose quaternionic part $\mathbf{n} \sin \theta$, when divided by $\sin \theta$, yields the unimodular complex quaternion $\mathbf{n}$. The complex logarithm w is as usual ambiguous by additive multiples of $2\pi I$, while the complex argument θ of the $\cos$ and $\sin$ is ambiguous by additive multiples of 2π . But in addition there is (a) an ambiguity of a choice of sign between $\mathbf{n}$ and $\sin \theta$, and (b) an ambiguity of a choice of sign between e^w and the sign of $\cos \theta + \mathbf{n} \sin \theta$. The first ambiguity may be resolved by fixing the real part of θ to lie in the interval $[0, \pi)$. The second ambiguity may be resolved by fixing the real part of e^w to be positive, achieved by setting the imaginary part of w to lie in the interval $(-\pi/2, \pi/2]$. For rotors, which are unimodular by definition, $e^w = 1$ and $w = 0$.

Exercise 14.8 Spline a Lorentz transformation. A spline is a polynomial that interpolates between two points with given values and derivatives at the two points. Confirm that the cubic spline of a real function $f(x)$ with given initial and final values f_0 and f_1 and given initial and final derivatives f'_0 and f'_1 at $x = 0$ and $x = 1$ is

$$f(x) = f_0 + f'_0 x + [3(f_1 - f_0) - 2f'_0 - f'_1] x^2 + [2(f_0 - f_1) + f'_0 + f'_1] x^3 . \quad (14.75)$$

The case in which the derivatives at the endpoints are set to zero, $f'_0 = f'_1 = 0$, is called the “natural” spline.

Argue that a Lorentz rotor can be splined by splining the quaternionic components of the logarithm of the Lorentz rotor.

Exercise 14.9 The wrong way to implement a Lorentz transformation. The principal purpose of this exercise is to persuade you that Lorentz transforming a 4-vector by the rule (14.69) is a much better idea than Lorentz transforming by multiplying by an explicit 4×4 matrix. Suppose that the Lorentz rotor R is the complex quaternion

$$R = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{Bmatrix}. \quad (14.76)$$

Show that the Lorentz transformation (14.69) transforms the 4-vector $a^m \gamma_m = \{a^0 \gamma_0, a^1 \gamma_1, a^2 \gamma_2, a^3 \gamma_3\}$ as (note that the 4×4 rotation matrix is written to the left of the 4-vector in accordance with the physics convention that rotations accumulate to the left):

$$R : \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix} \rightarrow \begin{pmatrix} |w|^2 + |x|^2 + |y|^2 + |z|^2 & 2(-w_R x_I + w_I x_R + y_R z_I - y_I z_R) \\ 2(-w_R x_I + w_I x_R - y_R z_I + y_I z_R) & |w|^2 + |x|^2 - |y|^2 - |z|^2 \\ 2(-w_R y_I + w_I y_R - z_R x_I + z_I x_R) & 2(x_R y_R + x_I y_I + w_R z_R + w_I z_I) \\ 2(-w_R z_I + w_I z_R - x_R y_I + x_I y_R) & 2(z_R x_R + z_I x_I - w_R y_R - w_I y_I) \\ 2(-w_R y_I + w_I y_R + z_R x_I - z_I x_R) & 2(-w_R z_I + w_I z_R + x_R y_I - x_I y_R) \\ 2(x_R y_R + x_I y_I - w_R z_R - w_I z_I) & 2(z_R x_R + z_I x_I + w_R y_R + w_I y_I) \\ |w|^2 - |x|^2 + |y|^2 - |z|^2 & 2(y_R z_R + y_I z_I - w_R x_R - w_I x_I) \\ 2(y_R z_R + y_I z_I + w_R x_R + w_I x_I) & |w|^2 - |x|^2 - |y|^2 + |z|^2 \end{pmatrix} \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix}, \quad (14.77)$$

where $||$ signifies the absolute value of a complex number, as in $|w|^2 = w_R^2 + w_I^2$. As a simple example, show that the transformation (14.77) in the case of a Lorentz boost by velocity v along the 1-axis, where the rotor R takes the form (14.43), is

$$R : \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix} \rightarrow \begin{pmatrix} \gamma & \gamma v & 0 & 0 \\ \gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix}, \quad (14.78)$$

with γ the familiar Lorentz gamma factor

$$\gamma = \cosh \theta = \frac{1}{\sqrt{1-v^2}}, \quad \gamma v = \sinh \theta = \frac{v}{\sqrt{1-v^2}}. \quad (14.79)$$

Exercise 14.10 Transform a 4-vector into a desired frame. Find Lorentz boosts that transform respectively (1) a timelike 4-vector a^k to point along the 0-axis, and (2) a null 4-vector a^k to point along the 0-1 null axis. Find a spatial rotation that transforms (3) a 4-vector $\{a^0, a^a\}$ so that its spatial part points along the 1-axis, leaving the time component a^0 unchanged.

Solution.

1. **Lorentz boost of a timelike 4-vector.** Let $a \equiv \pm\sqrt{-a_k a^k}$ be the magnitude of the timelike 4-vector a^k , with sign chosen to be that of a^0 . The Lorentz boost

$$\{w_R, x_I, y_I, z_I\} = \frac{1}{\sqrt{2a(a^0 + a)}} \{a^0 + a, a^1, a^2, a^3\} \quad (14.80)$$

transforms a^k to $\{a, 0, 0, 0\}$.

2. **Lorentz boost of a null 4-vector.** Choose a to be a non-zero real number with sign equal to that of a^0 . The Lorentz boost

$$\{w_R, x_I, y_I, z_I\} = \frac{1}{\sqrt{2a(a^0 \pm a^1)}} \{a^0 + a, a^1 \mp a, a^2, a^3\} \quad (14.81)$$

transforms a^k to $\{a, \pm a, 0, 0\}$.

3. **Spatial rotation of a 4-vector.** Let $a \equiv \sqrt{a_a a^a}$ be the spatial magnitude of the spatial 4-vector $a^k = \{a^0, a^a\}$. The spatial rotation

$$\{w_R, x_R, y_R, z_R\} = \frac{1}{\sqrt{2a(a^1 + a)}} \{a^1 + a, 0, a^3, -a^2\} \quad (14.82)$$

transforms a^k to $\{a^0, a, 0, 0\}$, leaving the time component a^0 unchanged.

14.6 Killing vector fields of Minkowski space

The geometry of Minkowski space is unchanged under two continuous groups of symmetries, the 4-dimensional group of translations, and the 6-dimensional group of Lorentz transformations. A symmetry transformation is a transformation of the coordinates that, with a suitable choice of coordinates, leaves the metric unchanged. Independent of the choice of coordinates, a symmetry transformation is a transformation that leaves the proper spacetime distance between any two points unchanged.

Any infinitesimal symmetry transformation defines a Killing vector ξ^μ , §7.32, which shifts the coordinates by an infinitesimal amount,

$$x^\mu \rightarrow x^\mu + \epsilon \xi^\mu, \quad (14.83)$$

with ϵ an infinitesimal real number. The infinitesimal transformation defines a flow field, called a Killing vector field, in the spacetime. The basic Killing vector fields of Minkowski space have been met earlier in this book. The Killing field associated with a translation is a set of parallel straight lines (timelike, null, or spacelike) in Minkowski space. The Killing field associated with a spatial rotation is a set of nested spacelike circles about a spatial axis, Figure 1.13. The Killing field associated with a pure Lorentz boost is a set of nested timelike, null, and spacelike hyperbolae, Figure 1.14.

The most general Killing vector field of Minkowski space is a linear combination of translational and Lorentz Killing vectors with constant coefficients. The Killing field associated with a pure Lorentz transformation (no translational component) always has at least one fixed point, the origin, which is unchanged by the

Lorentz transformation. The addition of a translational component corresponds to uniform translational motion (possibly superluminal) of the origin of the Lorentz transformation. In some cases the composition of a translation and a Lorentz transformation simplifies to a Lorentz transformation. For example, a Lorentz transformation (either a spatial rotation or a Lorentz boost) in a given 2-dimensional plane, coupled with a translation in the same plane, always has a fixed point, and is equivalent to another Lorentz transformation in the same plane with origin at the fixed point.

The remainder of this section considers the Killing field of a pure Lorentz transformation (no translational component). The Killing vector associated with a Lorentz transformation is its generator, which is a bivector, or equivalently complex quaternion, $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_R + I\boldsymbol{\theta}_I$. The real part $\boldsymbol{\theta}_R$ of the bivector is the generator of a spatial rotation, while the imaginary part $\boldsymbol{\theta}_I$ is the generator of a Lorentz boost. The decomposition of the bivector into real and imaginary parts is analogous to the decomposition of the electromagnetic field into magnetic and electric parts, equation (14.63). The complex modulus squared $|\boldsymbol{\theta}|^2$ of the bivector,

$$|\boldsymbol{\theta}|^2 \equiv \tilde{\boldsymbol{\theta}}\boldsymbol{\theta} = -\boldsymbol{\theta}^2 = \theta_I^2 - \theta_R^2 - 2I\boldsymbol{\theta}_R \cdot \boldsymbol{\theta}_I , \quad (14.84)$$

is invariant under Lorentz transformations. By a suitable Lorentz transformation, the bivector $\boldsymbol{\theta}$ may be adjusted arbitrarily, subject only to the condition that its complex modulus is fixed, that is, $\theta_I^2 - \theta_R^2$ and $\boldsymbol{\theta}_R \cdot \boldsymbol{\theta}_I$ are constant.

If the bivector is non-null, $|\boldsymbol{\theta}| \neq 0$, then by a suitable Lorentz transformation the real (magnetic) $\boldsymbol{\theta}_R$ and imaginary (electric) $\boldsymbol{\theta}_I$ parts can be taken to be parallel, directed along a common unimodular spatial direction, $\boldsymbol{n}$, say. So transformed, the bivector $\boldsymbol{\theta}$ is the complex quaternion

$$\boldsymbol{\theta} = (\theta_R + I\theta_I)\boldsymbol{\imath} \cdot \boldsymbol{n} . \quad (14.85)$$

The bivector (14.85) generates a uniform proper spatial rotation about the $\boldsymbol{n}$ axis, coupled with a uniform proper acceleration along the $\boldsymbol{n}$ axis. A Killing trajectory $\boldsymbol{x}(\lambda) \equiv x^m(\lambda)\boldsymbol{\gamma}_m$, parametrized by affine parameter λ along the trajectory, starting at $\lambda = 0$, is obtained by Lorentz transforming an initial 4-vector $\boldsymbol{x}_0 \equiv \boldsymbol{x}(0)$ by a rotor $R \equiv e^{-\lambda\boldsymbol{\theta}/2}$, equation (14.35),

$$\boldsymbol{x} = R\boldsymbol{x}_0\tilde{R} . \quad (14.86)$$

Define Killing coordinates α and ϕ by

$$\alpha \equiv \lambda\theta_I , \quad \phi \equiv \lambda\theta_R , \quad (14.87)$$

which like the affine parameter λ start at $\alpha = \phi = 0$. If the unimodular direction $\boldsymbol{n}$ is taken to be the x -direction, then Minkowski coordinates $x^m \equiv \{t, x, y, z\}$ along a Killing trajectory (14.86) starting at $x_0^m = \{0, l, r, 0\}$ are

$$\{t, x, y, z\} = \{l \sinh \alpha, l \cosh \alpha, r \cos \phi, r \sin \phi\} . \quad (14.88)$$

The Killing trajectory (14.88) is arranged, without loss of generality, such that it is initially at rest in the parallel x -direction, and moving with some initial velocity $v_\perp$ in the perpendicular z -direction,

$$v_\perp = \left. \frac{dz}{dt} \right|_{\text{init}} = \frac{r}{l} \frac{d\phi}{d\alpha} = \frac{r\theta_R}{l\theta_I} . \quad (14.89)$$

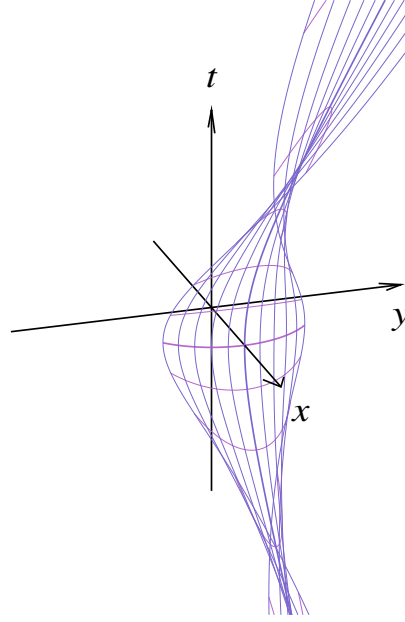


Figure 14.2 3D spacetime diagram of a sample of (blue) timelike Killing trajectories in Minkowski space. The two outermost of the trajectories shown lie on the light cylinder, and are lightlike. Motion in the z spatial direction is suppressed. The trajectories accelerate with uniform proper linear acceleration in the x -direction, and with uniform rotation in the yz -plane. The trajectories shown are for the case of a Killing vector with equal acceleration and rotational components, $|\theta_I| = |\theta_R|$ (corresponding to the motion of charges in equal electric and magnetic fields, $|E| = |B|$, Exercise 14.11). The crossing (purple) lines are spacelike lines of constant affine parameter λ .

The velocity $v_\perp$ is constant, and is the circular velocity from the perspective of an observer who is at the centre $r = 0$ of gyration, moving with the same instantaneous rapidity α in the x -direction as the Killing trajectory. A trajectory is timelike provided that

$$|v_\perp| < 1 . \quad (14.90)$$

Null trajectories, with $|v_\perp| = 1$, define the **light cylinder**. Killing trajectories outside the light cylinder are spacelike. The line-element with respect to Killing coordinates $\{\alpha, \phi\}$ and comoving coordinates $\{l, r\}$ is

$$ds^2 = -l^2 d\alpha^2 + dl^2 + dr^2 + r^2 d\phi^2 . \quad (14.91)$$

The proper time along a Killing trajectory $dl = dr = 0$ is

$$d\tau = \sqrt{l^2 d\alpha^2 - r^2 d\phi^2} = l \sqrt{1 - v_\perp^2} d\lambda . \quad (14.92)$$

The condition that λ be an affine parameter, $d\lambda = d\tau/m$, implies that the lengths l and r are related to

$$l = m\gamma_\perp , \quad r = m\gamma_\perp |v_\perp| , \quad (14.93)$$

where $\gamma_{\perp} \equiv 1/\sqrt{1-v_{\perp}^2}$ is the Lorentz gamma-factor corresponding to the velocity $v_{\perp}$. The 4-momentum $\mathbf{p} \equiv d\mathbf{x}/d\lambda$ and 4-acceleration $\boldsymbol{\kappa} \equiv d\mathbf{p}/d\lambda$ along the Killing trajectory are

$$\mathbf{p} = R\mathbf{p}_0\tilde{R}, \quad \boldsymbol{\kappa} = R\boldsymbol{\kappa}_0\tilde{R}, \quad (14.94)$$

with initial values

$$\mathbf{p}_0 = \left. \frac{d\mathbf{x}}{d\lambda} \right|_0 = -\frac{1}{2}[\boldsymbol{\theta}, \mathbf{x}_0], \quad \boldsymbol{\kappa}_0 = \left. \frac{d\mathbf{p}}{d\lambda} \right|_0 = -\frac{1}{2}[\boldsymbol{\theta}, \mathbf{p}_0]. \quad (14.95)$$

Figure 14.2 illustrates a sample of Killing trajectories for the case of equal boost and rotational components, $|\theta_I| = |\theta_R|$.

The above was for the case where the generating bivector $\boldsymbol{\theta}$ of the symmetry transformation was non-null. Alternatively, the generating bivector may be null, $\tilde{\boldsymbol{\theta}}\boldsymbol{\theta} = 0$. In this case the real and imaginary parts of the bivector are orthogonal, $\boldsymbol{\theta}_R \cdot \boldsymbol{\theta}_I = 0$, and their magnitudes are equal, $|\theta_R| = |\theta_I|$, equation (14.84). A null bivector is also nilpotent, $\boldsymbol{\theta}^2 = 0$, so the rotor R obtained by exponentiating $\boldsymbol{\theta}$ is linear in $\boldsymbol{\theta}$,

$$R \equiv e^{-\lambda\boldsymbol{\theta}/2} = 1 - \lambda\boldsymbol{\theta}/2. \quad (14.96)$$

A Killing trajectory $\mathbf{x}$ starting from an initial 4-vector $\mathbf{x}_0$ is

$$\mathbf{x} \equiv R\mathbf{x}_0\tilde{R} = \mathbf{x}_0 - \frac{\lambda}{2}[\boldsymbol{\theta}, \mathbf{x}_0], \quad (14.97)$$

which is a straight line passing through $\mathbf{x}_0$. It can be checked that the line may be spacelike or null, but never timelike. It is not clear whether this is a useful result.

Exercise 14.11 Motion of a charged particle in uniform parallel electric and magnetic fields.

Calculate the trajectory in Minkowski space of a particle of mass m and charge q in an electromagnetic field where the electric and magnetic fields are uniform and parallel, $\mathbf{E} = E\mathbf{n}$ and $\mathbf{B} = B\mathbf{n}$ (Landau and Lifshitz, 1975, §22, Problem 1).

Solution. See Figure 14.2. As long as the electromagnetic field $\mathbf{F} = \mathbf{B} - I\mathbf{E}$, equation (14.63), is non-null, $|\mathbf{F}| \neq 0$, the electric and magnetic fields can be made parallel by a suitable Lorentz transformation. The electric and magnetic fields are unchanged by a complex (with respect to I) Lorentz transformation along the common direction $\mathbf{n}$, that is, by a combination of a spatial rotation about $\mathbf{n}$ and a Lorentz boost along $\mathbf{n}$. Thus the symmetry of Minkowski space under Lorentz transformations along $\mathbf{n}$ is preserved by the introduction of uniform electric and magnetic fields along $\mathbf{n}$. The trajectories of charged particles are Killing trajectories of Lorentz transformations along the direction $\mathbf{n}$. The equation of motion (4.44),

$$\frac{d\mathbf{p}}{d\lambda} = \frac{1}{2}q[\mathbf{F}, \mathbf{p}], \quad (14.98)$$

implies that the Killing bivector is $\boldsymbol{\theta} = -q\mathbf{F}$, or equivalently

$$\theta_I = qE, \quad \theta_R = -qB. \quad (14.99)$$

14.7 Dirac matrices

The multiplication rules (14.1) for the basis vectors γ_m of the spacetime algebra are identical to the rules (14.5) governing the Clifford algebra of the **Dirac γ -matrices** used in the Dirac theory of relativistic spin- $\frac{1}{2}$ particles.

The Dirac γ -matrices are conventionally represented by 4×4 complex (with respect to the quantum-mechanical imaginary i) unitary matrices. The matrices act on 4-component complex (with respect to i) Dirac spinors, which are spin- $\frac{1}{2}$ particles, §14.8. Four complex components are precisely what is needed to represent a complex quaternion, or Dirac spinor, §14.9.

An essential feature of a successful theory of relativistic spinors is the existence of an inner product of spinors, necessary to allow a scalar probability to be defined. The systematic construction of a scalar product of spinors is deferred to Chapter 41, §41.6. Meanwhile, in the traditional Dirac approach, a spinor ψ is represented as a column vector with 4 complex (with respect to i) components, while its Hermitian conjugate $\psi^\dagger$ is a row vector with 4 complex components that are the complex conjugates (with respect to i) of the components of ψ . The product $\psi^\dagger \psi$ is a real number, but is not satisfactory as a scalar product since it is not Lorentz invariant. Rather, $\psi^\dagger \psi$ proves to be the time component n^0 of a 4-vector number current n^k , which the Dirac equation then shows to be covariantly conserved, $\dot{D}_k n^k = 0$, equation (44.21). The number current n^k is interpreted as a conserved probability current. The requirement that the Dirac number current density n^0 be positive imposes the condition, equation (41.104), that taking the Hermitian conjugate of any of the basis vectors γ_m be equivalent to raising its index,

$$\gamma_m^\dagger = \gamma^m. \quad (14.100)$$

Condition (14.100) is the same as requiring that the basis vectors be unitary matrices, $\gamma_m^{-1} = \gamma_m^\dagger$.

The high-energy physics community commonly adopts the $+---$ metric signature, which is opposite to the convention adopted here. With the high-energy $+---$ signature, the traditional Dirac representation of unitary γ -matrices satisfying the scalar product condition (14.5) is

$$\gamma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma_a = \begin{pmatrix} 0 & -\sigma_a \\ \sigma_a & 0 \end{pmatrix}, \quad (14.101)$$

where 1 denotes the unit 2×2 matrix, and σ_a denote the three 2×2 Pauli matrices (13.112). The choice of γ_0 as a diagonal matrix is motivated by Dirac's discovery that eigenvectors of the time basis vector γ_0 with eigenvalues of opposite sign define particles and antiparticles in their rest frames (see §14.8).

With the $-+++$ metric signature adopted here, the **Dirac representation** of the γ -matrices can be taken to be

$$\gamma_0 = i \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma_a = \begin{pmatrix} 0 & \sigma_a \\ \sigma_a & 0 \end{pmatrix}. \quad (14.102)$$

The representation (14.102) has the advantage that the resulting chiral basis vectors are all real, equations (41.15). In the representation (14.102), the bivectors σ_a and $I\sigma_a$ and the pseudoscalar I of the space-

time algebra are

$$\gamma_0 \gamma_a = \sigma_a = i \begin{pmatrix} 0 & \sigma_a \\ -\sigma_a & 0 \end{pmatrix}, \quad \frac{1}{2} \varepsilon_{abc} \gamma_b \gamma_c = I \sigma_a = i \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_a \end{pmatrix}, \quad I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \quad (14.103)$$

The Hermitian conjugates of the bivector and pseudoscalar basis elements are

$$\sigma_a^\dagger = \sigma_a, \quad (I \sigma_a)^\dagger = -I \sigma_a, \quad I^\dagger = -I. \quad (14.104)$$

The conventional chiral matrix γ_5 of Dirac theory is defined to be $-i$ times the pseudoscalar,

$$\gamma_5 \equiv -i \gamma_0 \gamma_1 \gamma_2 \gamma_3 = -i I = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}. \quad (14.105)$$

The chiral matrix γ_5 is Hermitian ($\gamma_5^\dagger = \gamma_5$) and unitary ($\gamma_5^{-1} = \gamma_5^\dagger$), so its square is the unit matrix,

$$\gamma_5^\dagger = \gamma_5, \quad \gamma_5^2 = 1. \quad (14.106)$$

14.8 Dirac spinors

A **Dirac spinor** is a spin- $\frac{1}{2}$ particle in Dirac's theory of relativistic spin- $\frac{1}{2}$ particles. A Dirac spinor ψ is a complex (with respect to the quantum-mechanical imaginary i) linear combination of 4 basis spinors ϵ_a with indices a running over $\{\uparrow, \downarrow, \uparrow, \downarrow\}$, a total of 8 degrees of freedom in all,

$$\psi = \psi^a \epsilon_a. \quad (14.107)$$

The basis spinors ϵ_a are basis elements of a super spacetime algebra, Chapter 41. In the Dirac representation (14.102), the four basis spinors are the column spinors

$$\epsilon_{\uparrow\uparrow} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon_{\uparrow\downarrow} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon_{\downarrow\uparrow} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \epsilon_{\downarrow\downarrow} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \quad (14.108)$$

The Dirac γ -matrices operate by pre-multiplication on Dirac spinors ψ , yielding other Dirac spinors. The basis spinors are eigenvectors of the time basis vector γ_0 and of the bivector $I\sigma_3$, with $\epsilon_\uparrow$ and $\epsilon_\downarrow$ denoting eigenvectors of γ_0 , and $\epsilon_\uparrow$ and $\epsilon_\downarrow$ eigenvectors of $I\sigma_3$,

$$\gamma_0 \epsilon_\uparrow = i \epsilon_\uparrow, \quad \gamma_0 \epsilon_\downarrow = -i \epsilon_\downarrow, \quad I\sigma_3 \epsilon_\uparrow = i \epsilon_\uparrow, \quad I\sigma_3 \epsilon_\downarrow = -i \epsilon_\downarrow. \quad (14.109)$$

The bivector $I\sigma_3$ is the generator of a spatial rotation about the 3-axis (z -axis), equation (14.32). Simultaneous eigenvectors of γ_0 and $I\sigma_3$ exist because γ_0 and $I\sigma_3$ commute.

A pure spin-up Dirac spinor $\epsilon_\uparrow$ can be rotated into a pure spin-down spinor $\epsilon_\downarrow$, or vice versa, by a spatial rotation about the 1-axis or 2-axis. By contrast, a pure time-up spinor $\epsilon_\uparrow$ cannot be rotated into a pure time-down spinor $\epsilon_\downarrow$, or vice versa, by any Lorentz transformation. Consider for example trying to rotate the pure time-up spin-up $\epsilon_{\uparrow\uparrow}$ spinor into any combination of pure time-down $\epsilon_\downarrow$ spinors. According to the

expression (14.122), the Dirac spinor ψ obtained by Lorentz transforming the $\epsilon_{\uparrow}$ spinor is pure $\epsilon_{\perp}$ only if the corresponding complex quaternion q is pure imaginary (with respect to I). But a pure imaginary quaternion has negative squared modulus $\tilde{q}q$, so cannot be equivalent to any unimodular rotor.

Thus the pure time-up and pure time-down spinors $\epsilon_{\uparrow}$ and $\epsilon_{\downarrow}$ are distinct spinors that cannot be transformed into each other by any Lorentz transformation. The two spinors represent distinct species, particles and antiparticles (see §14.10).

Although a pure time-up spinor cannot be transformed into a pure time-down spinor or vice versa by any Lorentz transformation, the time-up and time-down spinors $\epsilon_{\uparrow}$ and $\epsilon_{\downarrow}$ do mix under Lorentz transformations. The manner in which Dirac spinors transform is described in §14.9.

The choice of time-axis γ_0 and spin-axis γ_3 with respect to which the eigenvectors are defined can of course be adjusted arbitrarily by a Lorentz boost and a spatial rotation. The eigenvectors of a particular time-axis γ_0 correspond to either particles or antiparticles that are at rest in that frame. The eigenvectors associated with a particular spin-axis γ_3 correspond to particles or antiparticles that are either pure spin-up or pure spin-down in that frame.

14.9 Dirac spinors as complex quaternions

In §13.15 it was found that a spin- $\frac{1}{2}$ object in 3D space, a Pauli spinor, is isomorphic to a real quaternion, or equivalently scaled 3D rotor, equation (13.122). In the relativistic theory, the corresponding spin- $\frac{1}{2}$ object, a Dirac spinor ψ , is isomorphic (14.113) to a complex quaternion. The 4 complex degrees of freedom of the Dirac spinor ψ are equivalent to the 8 degrees of freedom of a complex quaternion. A physically interesting complication arises in the relativistic case because a non-trivial Dirac spinor can be null, isomorphic to a null complex quaternion, whereas any non-trivial Pauli spinor is necessarily non-null. The cases of non-null (massive) and null (massless) Dirac spinors are considered respectively in §14.10 and §14.11. If the Dirac spinor is non-null, then it is equivalent to a scaled rotor, equation (14.140), but if the Dirac spinor is null, then it is not simply a scaled rotor. The present section establishes an isomorphism (14.113) between Dirac spinors and complex quaternions that is valid in general, regardless of whether the Dirac spinor is null or not.

If $\mathbf{a}$ is a spacetime multivector, equivalent to an element of the Clifford algebra of Dirac γ -matrices, then under rotation by Lorentz rotor R , the multivector $\mathbf{a}$ operating on the Dirac spinor ψ transforms as

$$R : \mathbf{a}\psi \rightarrow (R\mathbf{a}\tilde{R})(R\psi) = R\mathbf{a}\psi . \quad (14.110)$$

This shows that a Dirac spinor ψ Lorentz transforms, by construction, as

$$R : \psi \rightarrow R\psi . \quad (14.111)$$

The rule (14.111) is precisely the transformation rule (13.75) for spacetime rotors under Lorentz transformations: under a rotation by rotor R , a rotor S transforms as $S \rightarrow RS$. More generally, the transformation law (14.111) holds for any linear combination of Dirac spinors ψ . The isomorphism (14.34) between spacetime rotors and unimodular quaternions, coupled with linearity, shows that the vector space of Dirac spinors is

isomorphic to the vector space of complex quaternions. Specifically, any Dirac spinor ψ can be expressed uniquely in the form of a 4×4 matrix $\mathbf{q}$, the Dirac representation of a complex quaternion q , acting on the time-up spin-up column vector $\epsilon_{\uparrow\uparrow}$ (the precise translation between Dirac spinors and complex quaternions is left as Exercises 14.12 and 14.13):

$$\psi = \mathbf{q} \epsilon_{\uparrow\uparrow} . \quad (14.112)$$

In this section (including the Exercises) the 4×4 matrix $\mathbf{q}$ is written in boldface to distinguish it from the quaternion q that it represents; but the distinction is not fundamental. The mapping (14.112) establishes an isomorphism between the vector spaces of Dirac spinors and quaternions

$$\psi \leftrightarrow q . \quad (14.113)$$

The isomorphism means that there is a one-to-one correspondence between Dirac spinors ψ and complex quaternions q , and that they transform in the same way under Lorentz transformations.

The isomorphism between the vector spaces of Dirac spinors and complex quaternions does not extend to multiplication; that is, the product of two Dirac spinors ψ_1 and ψ_2 equivalent to the complex 4×4 matrices $\mathbf{q}_1$ and $\mathbf{q}_2$ does not equal the Dirac spinor equivalent to the product $\mathbf{q}_1 \mathbf{q}_2$. The problem is that the Dirac representation of a Dirac spinor ψ is a column vector, and two column vectors cannot be multiplied. The question of how to multiply Dirac spinors is resumed in Chapter 41 on the super spacetime algebra.

14.9.1 Reverse Dirac spinor

An essential feature of any viable theory of spinors is the existence of a scalar product of spinors. The scalar product must be a complex (with respect to the quantum mechanical imaginary i) number that is invariant under Lorentz transformations. Now the product $\tilde{q}q$ of the conjugate of a quaternion with itself is a Lorentz-invariant complex (with respect to I) number. This suggests defining a row Dirac spinor $\tilde{\psi}$ reverse to the column Dirac spinor ψ defined by equation (14.112) by

$$\tilde{\psi} \equiv \epsilon_{\uparrow\uparrow}^\top \tilde{\mathbf{q}} , \quad (14.114)$$

where $\tilde{\mathbf{q}}$ is the matrix representation of the quaternionic conjugate $\tilde{q}$ of the complex quaternion q , and $\epsilon_a^\top$ denotes the basis of row Dirac spinors obtained by transposing the basis of column Dirac spinors defined by equations (14.108),

$$\epsilon_{\uparrow\uparrow}^\top = (1 \ 0 \ 0 \ 0) , \quad \epsilon_{\uparrow\downarrow}^\top = (0 \ 1 \ 0 \ 0) , \quad \epsilon_{\downarrow\uparrow}^\top = (0 \ 0 \ 1 \ 0) , \quad \epsilon_{\downarrow\downarrow}^\top = (0 \ 0 \ 0 \ 1) . \quad (14.115)$$

The reverse Dirac spinor $\tilde{\psi}$ is also called the Dirac adjoint spinor. It is related to the Hermitian conjugate Dirac spinor $\psi^\dagger$ by equation (14.130), and is the same as the Dirac row conjugate spinor $\bar{\psi}$ discussed in Chapter 41, equation (41.102).

As found in equation (14.125a), $\tilde{\psi}\psi$ is a Lorentz-invariant real number. More generally, the product $\tilde{\chi}\psi$ of a row spinor $\tilde{\chi}$ with a column spinor ψ is a Lorentz-invariant complex (with respect to i) number, and therefore provides a viable definition of a scalar product of Dirac spinors. The problem of defining a scalar product of Dirac spinors is resumed in §41.6.1.

Exercise 14.12 Translate a Dirac spinor into a complex quaternion. Given any Dirac spinor in the Dirac representation (14.102),

$$\psi = \psi^a \epsilon_a = \begin{pmatrix} \psi^{\top\uparrow} \\ \psi^{\top\downarrow} \\ \psi^{\perp\uparrow} \\ \psi^{\perp\downarrow} \end{pmatrix}, \quad (14.116)$$

show that the corresponding complex quaternion q , and the equivalent 4×4 matrix $\mathbf{q}$ such that $\psi = \mathbf{q} \epsilon_{\top\uparrow}$, are (the complex conjugates ψ^{a*} of the components ψ^a of the spinor are with respect to the quantum-mechanical imaginary i)

$$q = \left\{ \begin{array}{cccc} \text{Re } \psi^{\top\uparrow} & \text{Im } \psi^{\top\downarrow} & -\text{Re } \psi^{\top\downarrow} & \text{Im } \psi^{\top\uparrow} \\ \text{Re } \psi^{\perp\uparrow} & \text{Im } \psi^{\perp\downarrow} & -\text{Re } \psi^{\perp\downarrow} & \text{Im } \psi^{\perp\uparrow} \end{array} \right\} \leftrightarrow \mathbf{q} = \begin{pmatrix} \psi^{\top\uparrow} & -\psi^{\top\downarrow*} & -\psi^{\perp\uparrow} & \psi^{\perp\downarrow*} \\ \psi^{\top\downarrow} & \psi^{\top\uparrow*} & -\psi^{\perp\downarrow} & -\psi^{\perp\uparrow*} \\ \psi^{\perp\uparrow} & -\psi^{\perp\downarrow*} & \psi^{\top\uparrow} & -\psi^{\top\downarrow*} \\ \psi^{\perp\downarrow} & \psi^{\perp\uparrow*} & \psi^{\top\downarrow} & \psi^{\top\uparrow*} \end{pmatrix}. \quad (14.117)$$

Show that the quaternionic conjugate $\tilde{q}$ and the equivalent 4×4 matrix $\tilde{\mathbf{q}}$ in the Dirac representation (14.102), are

$$\tilde{q} = \left\{ \begin{array}{cccc} \text{Re } \psi^{\top\uparrow} & -\text{Im } \psi^{\top\downarrow} & \text{Re } \psi^{\top\downarrow} & -\text{Im } \psi^{\top\uparrow} \\ \text{Re } \psi^{\perp\uparrow} & -\text{Im } \psi^{\perp\downarrow} & \text{Re } \psi^{\perp\downarrow} & -\text{Im } \psi^{\perp\uparrow} \end{array} \right\} \leftrightarrow \tilde{\mathbf{q}} = \begin{pmatrix} \psi^{\top\uparrow*} & \psi^{\top\downarrow*} & -\psi^{\perp\uparrow*} & -\psi^{\perp\downarrow*} \\ -\psi^{\top\downarrow} & \psi^{\top\uparrow} & \psi^{\perp\downarrow} & -\psi^{\perp\uparrow} \\ \psi^{\perp\uparrow*} & \psi^{\perp\downarrow*} & \psi^{\top\uparrow*} & \psi^{\top\downarrow*} \\ -\psi^{\perp\downarrow} & \psi^{\perp\uparrow} & -\psi^{\top\downarrow} & \psi^{\top\uparrow} \end{pmatrix}. \quad (14.118)$$

Conclude that the reverse spinor $\tilde{\psi}$ defined by equation (14.114) is

$$\tilde{\psi} \equiv \epsilon_{\top\uparrow}^\top \tilde{\mathbf{q}} = \begin{pmatrix} \psi^{\top\uparrow*} & \psi^{\top\downarrow*} & -\psi^{\perp\uparrow*} & -\psi^{\perp\downarrow*} \end{pmatrix}. \quad (14.119)$$

Exercise 14.13 Translate a complex quaternion into a Dirac spinor. Show that the complex quaternion $q \equiv w + ix + jy + kz$ is equivalent in the Dirac representation (14.102) to the 4×4 matrix $\mathbf{q}$

$$q = \left\{ \begin{array}{cccc} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{array} \right\} \leftrightarrow \mathbf{q} = \begin{pmatrix} w_R + iz_R & ix_R + y_R & -w_I - iz_I & -ix_I - y_I \\ ix_R - y_R & w_R - iz_R & -ix_I + y_I & -w_I + iz_I \\ w_I + iz_I & ix_I + y_I & w_R + iz_R & ix_R + y_R \\ ix_I - y_I & w_I - iz_I & ix_R - y_R & w_R - iz_R \end{pmatrix}. \quad (14.120)$$

Show that the quaternionic conjugate $\tilde{q}$, the complex conjugate (with respect to I) quaternion q^* , and the quaternionic and complex conjugate quaternion $\tilde{q}^*$ are respectively equivalent to the 4×4 matrices

$$\tilde{q} \leftrightarrow \tilde{\mathbf{q}} = -\gamma_0 \mathbf{q}^\dagger \gamma_0, \quad (14.121a)$$

$$q^* \leftrightarrow \tilde{\mathbf{q}}^\dagger = -\gamma_0 \mathbf{q} \gamma_0, \quad (14.121b)$$

$$\tilde{q}^* \leftrightarrow \mathbf{q}^\dagger = -\gamma_0 \tilde{\mathbf{q}} \gamma_0, \quad (14.121c)$$

where γ_0 is the Dirac γ -matrix given by equation (14.102). Conclude that the Dirac spinor $\psi \equiv \mathbf{q} \epsilon_{\uparrow\uparrow}$ corresponding to the complex quaternion q is

$$\psi \equiv \mathbf{q} \epsilon_{\uparrow\uparrow} = \begin{pmatrix} w_R + iz_R \\ ix_R - y_R \\ w_I + iz_I \\ ix_I - y_I \end{pmatrix}, \quad (14.122)$$

that the reverse spinor $\tilde{\psi}$, equation (14.114), is

$$\tilde{\psi} \equiv \epsilon_{\uparrow\uparrow}^\top \tilde{\mathbf{q}} = \begin{pmatrix} w_R - iz_R & -ix_R - y_R & -w_I + iz_I & ix_I + y_I \end{pmatrix}, \quad (14.123)$$

and that the Hermitian conjugate spinor $\psi^\dagger$ is

$$\psi^\dagger \equiv \epsilon_{\uparrow\uparrow}^\top \mathbf{q}^\dagger = \begin{pmatrix} w_R - iz_R & -ix_R - y_R & w_I - iz_I & -ix_I - y_I \end{pmatrix}. \quad (14.124)$$

Hence conclude that $\tilde{\psi}\psi$ and $\psi^\dagger\psi$ are

$$\tilde{\psi}\psi = \text{Re}(\tilde{q}q) = \tilde{q}_R q_R - \tilde{q}_I q_I, \quad (14.125a)$$

$$\psi^\dagger\psi = \tilde{q}_R q_R + \tilde{q}_I q_I, \quad (14.125b)$$

with

$$\tilde{q}_R q_R = w_R^2 + x_R^2 + y_R^2 + z_R^2, \quad (14.126a)$$

$$\tilde{q}_I q_I = w_I^2 + x_I^2 + y_I^2 + z_I^2. \quad (14.126b)$$

Exercise 14.14 Pseudoscalar times a Dirac spinor. In §14.10 it will be found that multiplying a Dirac spinor ψ by the pseudoscalar I converts it to an antispinor. In Chapter 41, equation (41.136), it will be seen that $I\psi$ is the PT conjugate of I , the spinor obtained by reversing all 4 axes of space and time. Show that the product Iq of the pseudoscalar I with the complex quaternion $q \equiv w + ix + jy + kz$ is equivalent in the Dirac representation (14.102) to the 4×4 matrix $I\mathbf{q}$

$$Iq = \left\{ \begin{array}{cccc} -w_I & -x_I & -y_I & -z_I \\ w_R & x_R & y_R & z_R \end{array} \right\} \leftrightarrow I\mathbf{q} = \begin{pmatrix} -w_I - iz_I & -ix_I - y_I & -w_R - iz_R & -ix_R - y_R \\ -ix_I + y_I & -w_I + iz_I & -ix_R + y_R & -w_R + iz_R \\ w_R + iz_R & ix_R + y_R & -w_I - iz_I & -ix_I - y_I \\ ix_R - y_R & w_R - iz_R & -ix_I + y_I & -w_I + iz_I \end{pmatrix}. \quad (14.127)$$

Conclude that the Dirac antispinor $I\psi \equiv I\mathbf{q} \epsilon_{\uparrow\uparrow}$ corresponding to the complex quaternion Iq is

$$I\psi = I\mathbf{q} \epsilon_{\uparrow\uparrow} = \begin{pmatrix} -w_I - iz_I \\ -ix_I + y_I \\ w_R + iz_R \\ ix_R - y_R \end{pmatrix}, \quad (14.128)$$

Conclude that the pseudomagnitude $\tilde{\psi}I\psi$ is

$$\tilde{\psi}I\psi = -\text{Im}(\tilde{q}q) = -2(w_R w_I + x_R x_I + y_R y_I + z_R z_I). \quad (14.129)$$

Exercise 14.15 Relation between $\tilde{\psi}$ and $\psi^\dagger$. Show that $\tilde{\psi}$ and $\psi^\dagger$ are related by

$$\tilde{\psi} = -i\psi^\dagger\gamma_0, \quad \psi^\dagger = -i\tilde{\psi}\gamma_0, \quad (14.130)$$

by showing from equations (14.123) and (14.124) that

$$\psi^\dagger = -i\epsilon_{\uparrow\uparrow}^\top \tilde{q}\gamma_0. \quad (14.131)$$

The same result follows from equation (14.121b). The Hermitian conjugate matrix is $q^\dagger = -\gamma_0 \tilde{q} \gamma_0$, and $\epsilon_{\uparrow\uparrow}^\top \gamma_0 = i\epsilon_{\uparrow\uparrow}^\top$, so $\psi^\dagger \equiv \epsilon_{\uparrow\uparrow}^\top q^\dagger = -i\epsilon_{\uparrow\uparrow}^\top \tilde{q} \gamma_0$.

Exercise 14.16 Translate a Dirac spinor into a pair of Pauli spinors. Show that in terms of the real and imaginary (with respect to I) parts of the complex quaternion q , the equivalent 4×4 matrix $\mathbf{q}$ is

$$q = q_R + Iq_I \leftrightarrow \mathbf{q} = \begin{pmatrix} q_R & -q_I \\ q_I & q_R \end{pmatrix}, \quad (14.132)$$

where q_R and q_I are the complex 2×2 special unitary matrices equivalent to the real quaternions q_R and q_I , equation (13.130). Show that the reverse quaternion $\tilde{q}$, the complex conjugate (with respect to I) quaternion q^* , and the reverse complex conjugate (with respect to I) quaternion $\tilde{q}^*$ are respectively

$$\tilde{q} \leftrightarrow \tilde{\mathbf{q}} = \begin{pmatrix} q_R^\dagger & -q_I^\dagger \\ q_I^\dagger & q_R^\dagger \end{pmatrix}, \quad (14.133a)$$

$$q^* \leftrightarrow \tilde{\mathbf{q}}^\dagger = \begin{pmatrix} q_R & q_I \\ -q_I & q_R \end{pmatrix}, \quad (14.133b)$$

$$\tilde{q}^* \leftrightarrow \mathbf{q}^\dagger = \begin{pmatrix} q_R^\dagger & q_I^\dagger \\ -q_I^\dagger & q_R^\dagger \end{pmatrix}. \quad (14.133c)$$

Conclude that the Dirac spinor $\psi \equiv \mathbf{q} \epsilon_{\uparrow\uparrow}$ corresponding to the complex quaternion q is

$$\psi \equiv \mathbf{q} \epsilon_{\uparrow\uparrow} = \begin{pmatrix} \psi_R \\ \psi_I \end{pmatrix}, \quad (14.134)$$

where ψ_R and ψ_I are the Pauli spinors corresponding to the real quaternions q_R and q_I , equation (13.131). Conclude further that the antispinor $I\psi$ is

$$I\psi \equiv I\mathbf{q} \epsilon_{\uparrow\uparrow} = \begin{pmatrix} -\psi_I \\ \psi_R \end{pmatrix}, \quad (14.135)$$

that the reverse spinor $\tilde{\psi}$, equation (14.114), is

$$\tilde{\psi} \equiv \epsilon_{\uparrow\uparrow}^\top \tilde{\mathbf{q}} = \begin{pmatrix} \psi_R^\dagger & -\psi_I^\dagger \end{pmatrix}, \quad (14.136)$$

and that the Hermitian conjugate spinor $\psi^\dagger$ is

$$\psi^\dagger \equiv \epsilon_{\uparrow\uparrow}^\top \mathbf{q}^\dagger = \begin{pmatrix} \psi_R^\dagger & \psi_I^\dagger \end{pmatrix}. \quad (14.137)$$

Hence conclude that $\tilde{\psi}\psi$, $\tilde{\psi}I\psi$, and $\psi^\dagger\psi$ are given by

$$\tilde{\psi}\psi = \psi_R^\dagger\psi_R - \psi_I^\dagger\psi_I , \quad (14.138a)$$

$$\tilde{\psi}I\psi = -(\psi_R^\dagger\psi_I + \psi_I^\dagger\psi_R) , \quad (14.138b)$$

$$\psi^\dagger\psi = \psi_R^\dagger\psi_R + \psi_I^\dagger\psi_I . \quad (14.138c)$$

Exercise 14.17 Is the group of Lorentz rotors isomorphic to $\text{SU}(4)$? Previously, Exercise 13.10, it was found that the group $\text{Spin}(3)$ of spatial rotors in 3 dimensions is isomorphic to $\text{SU}(2)$. Is the group $\text{Spin}(3,1)$ of Lorentz rotors isomorphic to the group $\text{SU}(4)$ of complex 4×4 unitary matrices with unit determinant?

Solution. No. The Dirac representation of the group $\text{Spin}(3,1)$ of Lorentz rotors shares with $\text{SU}(4)$ the property that its matrices are complex 4×4 matrices with unit determinant. From the equivalence (14.120), the determinant of the 4×4 complex matrix $\mathbf{q}$ equivalent to a complex quaternion q is

$$\det \mathbf{q} = (\tilde{q}q)^*(\tilde{q}q) . \quad (14.139)$$

Since a Lorentz rotor is unimodular, with $\tilde{q}q = 1$, its Dirac representation has unit determinant. However, the Dirac representation of a Lorentz rotor is not unitary (its inverse is not its Hermitian conjugate), despite the fact that all the generators of the group, namely the 6 bivectors σ_a and $I\sigma_a$, are unitary. Rather, the inverse of a rotor R is its reverse $\tilde{R}$, related to its Hermitian conjugate by equation (14.121a). The condition for the matrices of a group to be unitary is that the generators be skew-Hermitian (they equal minus their Hermitian conjugates). The 3 spatial generators $I\sigma_a$ are indeed skew-Hermitian, but the 3 boost generators σ_a are Hermitian.

14.10 Non-null Dirac spinor

A non-null, or massive, Dirac spinor ψ is one that is isomorphic (14.113) to a non-null complex quaternion q . A non-null complex quaternion can be factored as a non-zero complex (with respect to I) scalar $\lambda = \lambda_R + I\lambda_I$ times a unimodular complex (with respect to I) quaternion R , a Lorentz rotor. Thus a non-null Dirac spinor can be expressed as, equation (14.112) (the boldface for q , adopted in §14.9 to distinguish a quaternion q from its matrix representation $\mathbf{q}$, is dropped henceforth, since the distinction is not fundamental),

$$\psi = q\epsilon_{\uparrow\uparrow} , \quad q = \lambda R . \quad (14.140)$$

The complex scalar λ can be taken without loss of generality to lie in the right hemisphere of the complex plane (positive real part), since a minus sign can be absorbed into a spatial rotation by 2π of the rotor R . There is no further ambiguity in the decomposition (14.140) into scalar and rotor, because the squared modulus $\tilde{\lambda}R\lambda R = \lambda^2$ of the scaled rotor λR is the same for any decomposition (do not confuse reversion with complex conjugation; the reverse of a scalar is itself, $\tilde{\lambda} = \lambda$; the product λ^2 is a complex (with respect to I) number).

The fact that a non-null Dirac spinor ψ encodes a Lorentz rotor shows that a non-null Dirac spinor in

some sense “knows” about the Lorentz structure of spacetime. It is profound that the Lorentz structure of spacetime is built in to a non-null Dirac particle.

As discussed in §14.8, a pure time-up eigenvector $\epsilon_{\top}$ represents a particle in its own rest frame, while a pure time-down eigenvector $\epsilon_{\perp}$ represents an antiparticle in its own rest frame. The time-up spin-up eigenvector $\epsilon_{\top\uparrow}$ is by definition (14.140) equivalent to the unit scaled rotor, $\lambda R = 1$, so in this case the scalar λ is pure real. Lorentz transforming the eigenvector multiplies it by a rotor, but leaves the scalar λ unchanged, therefore pure real. Conversely, if the time-up spin-up eigenvector $\epsilon_{\top\uparrow}$ is multiplied by the imaginary I , then according to the expression (14.122) the resulting spinor can be Lorentz transformed into a pure $\epsilon_{\perp}$ spinor, corresponding to a pure antiparticle. Thus one may conclude that the real and imaginary parts (with respect to I) of the complex scalar $\lambda = \lambda_R + I\lambda_I$ correspond respectively to particles and antiparticles.

The Lorentz-invariant decomposition of a non-null Dirac spinor ψ into its particle $\psi_{\top}$ and antiparticle $\psi_{\perp}$ parts is accomplished by

$$\psi = \psi_{\top} + I\psi_{\perp}, \quad \psi_{\top} \equiv \frac{\text{Re } \lambda}{\lambda} \psi, \quad \psi_{\perp} \equiv \frac{\text{Im } \lambda}{\lambda} \psi, \quad \lambda = \sqrt{\tilde{\psi}\psi - I(\tilde{\psi}I\psi)}. \quad (14.141)$$

The decomposition (14.141) is not the same as the decomposition (14.134) of the Dirac spinor into a pair of Pauli spinors. The decomposition (14.141) into particle and antiparticle parts is Lorentz-invariant, whereas the Pauli spinors of the decomposition (14.134) mix under Lorentz boosts. The Lorentz-invariant magnitude $\tilde{\psi}\psi$ of the Dirac spinor, equation (14.125a), is the difference between the probabilities λ_R^2 of particles and λ_I^2 of antiparticles,

$$\tilde{\psi}\psi = \tilde{\psi}_{\top}\psi_{\top} - \tilde{\psi}_{\perp}\psi_{\perp}, \quad \tilde{\psi}_{\top}\psi_{\top} = \lambda_R^2, \quad \tilde{\psi}_{\perp}\psi_{\perp} = \lambda_I^2. \quad (14.142)$$

Thus $\tilde{\psi}\psi$ is positive for particles, negative for antiparticles. The Lorentz-invariant pseudomagnitude $\tilde{\psi}I\psi$, equation (14.129), is minus twice the product $\lambda_R\lambda_I$ of the amplitudes of particles and antiparticles,

$$\tilde{\psi}I\psi = -\tilde{\psi}_{\top}\psi_{\perp} - \tilde{\psi}_{\perp}\psi_{\top}, \quad \tilde{\psi}_{\top}\psi_{\perp} = \tilde{\psi}_{\perp}\psi_{\top} = \lambda_R\lambda_I. \quad (14.143)$$

The sum of the probabilities λ_R^2 of particles and λ_I^2 of antiparticles equals the number density in the rest frame, which can be written in the manifestly Lorentz-invariant form

$$\tilde{\psi}_{\top}\psi_{\top} + \tilde{\psi}_{\perp}\psi_{\perp} = \lambda_R^2 + \lambda_I^2 = \sqrt{(\tilde{\psi}\boldsymbol{\gamma}^m\psi)(\tilde{\psi}\boldsymbol{\gamma}_m\psi)}. \quad (14.144)$$

Since λ_R and λ_I are invariant under Lorentz transformations, all three terms $\tilde{\psi}_{\top}\psi_{\top}$, $\tilde{\psi}_{\perp}\psi_{\perp}$, and $\tilde{\psi}_{\top}\psi_{\perp} = \tilde{\psi}_{\perp}\psi_{\top}$ are Lorentz-invariant scalars.

Concept question 14.18 Is $\tilde{\psi}\psi$ real or complex? If $\psi \equiv \lambda\epsilon_{\top\uparrow}$ is a Dirac spinor corresponding to a complex quaternion $\lambda = \lambda_R + I\lambda_I$ with no quaternionic part (so $\tilde{\lambda} = \lambda$), should it not be that

$$\tilde{\psi}\psi = \epsilon_{\top\uparrow}^{\top} \tilde{\lambda} \lambda \epsilon_{\top\uparrow} = \epsilon_{\top\uparrow}^{\top} \lambda^2 \epsilon_{\top\uparrow} = \lambda^2, \quad (14.145)$$

which is a complex number, not a real number? **Answer.** No. Do not confuse the quantum-mechanical

imaginary i with the pseudoscalar I . The combination $\epsilon_{\uparrow\uparrow}^\top I \epsilon_{\uparrow\uparrow}$ does *not* equal $I \epsilon_{\uparrow\uparrow}^\top \epsilon_{\uparrow\uparrow}$. The complex (with respect to I) number λ , and its matrix representation $\boldsymbol{\lambda}$ are, equation (14.120),

$$\lambda = \lambda_R + I\lambda_I \leftrightarrow \boldsymbol{\lambda} = \lambda_R \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + i\lambda_I \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad (14.146)$$

where the 1's in the matrices on the right hand side denote the 2×2 unit matrix. The product $\tilde{\psi}\psi$ is

$$\tilde{\psi}\psi = \begin{pmatrix} \lambda_R & 0 & i\lambda_I & 0 \end{pmatrix} \begin{pmatrix} \lambda_R \\ 0 \\ i\lambda_I \\ 0 \end{pmatrix} = \lambda_R^2 - \lambda_I^2, \quad (14.147)$$

in agreement with equation (14.142), not equation (14.145).

Concept question 14.19 Is $\tilde{\psi}\boldsymbol{\gamma}_m\psi$ a scalar or a 4-vector? Under a Lorentz transformation $\tilde{\psi}\boldsymbol{\gamma}_m\psi$ transforms as

$$R: \tilde{\psi}\boldsymbol{\gamma}_m\psi \rightarrow \tilde{\psi}\tilde{R}R\boldsymbol{\gamma}_m\tilde{R}R\psi = \tilde{\psi}\boldsymbol{\gamma}_m\psi, \quad (14.148)$$

which appears to be a scalar. Yet $\tilde{\psi}\boldsymbol{\gamma}_m\psi$ also looks like it transforms as a 4-vector. Which is it? **Answer.** The transformations of spinors $\psi = \psi^a \epsilon_a$ and vectors $\boldsymbol{a} = a^m \boldsymbol{\gamma}_m$ considered in this Chapter are active transformations, §13.9, which rotate the basis spinors ϵ_a and vectors $\boldsymbol{\gamma}_m$ while keeping coefficients ψ^a and a^m fixed. Under active transformations the combination $\tilde{\psi}\boldsymbol{a}\psi$ is indeed a scalar, transforming as

$$R: \tilde{\psi}\boldsymbol{a}\psi \rightarrow \tilde{\psi}\tilde{R}R\boldsymbol{a}\tilde{R}R\psi = \tilde{\psi}\boldsymbol{a}\psi. \quad (14.149)$$

In fact $\tilde{\psi}\boldsymbol{a}\psi$ is a scalar product by construction, as will be explored in greater depth in a later Chapter, §41.6, so the fact that it transforms like a scalar should not be a surprise. However, as usual, one is free to make choices as to whether a transformation is active (bodily rotates an object) or passive (rotates the frame while leaving the object itself unchanged), §13.9. In most of this book, the convention is that transformations are passive, meaning that a transformation rotates both the coefficients and basis elements of a spinor $\psi = \psi^a \epsilon_a$ or vector $\boldsymbol{a} = a^m \boldsymbol{\gamma}_m$, while leaving the spinor or vector itself unchanged. With the passive convention, $\tilde{\psi}\boldsymbol{\gamma}_m\psi$ indeed transforms as a covariant vector (while $\tilde{\psi}\boldsymbol{a}\psi = \tilde{\psi}a^m \boldsymbol{\gamma}_m\psi$ transforms as a scalar, the transformation of the covariant vector $\boldsymbol{\gamma}_m$ cancelling against the transformation of the contravariant vector a^m). The advantage of the passive convention is that the transformation properties of an object are evident from the indices attached to it. However, the active convention of the present Chapter is needed in order to establish the fundamentals of how spinors transform.

14.11 Null Dirac Spinor

A **null Dirac spinor** is a Dirac spinor ψ constructed from a null complex quaternion q acting on the rest-frame eigenvector $\epsilon_{\uparrow\uparrow}$,

$$\psi = q \epsilon_{\uparrow\uparrow}, \quad \tilde{q}q = 0. \quad (14.150)$$

Physically, a null Dirac spinor represents a spin- $\frac{1}{2}$ particle moving at the speed of light. A non-trivial null spinor must be moving at the speed of light because if it were not, then there would be a rest frame where the rotor part of the spinor $\psi = \lambda R \epsilon_{\uparrow}$ would be unity, $R = 1$, and the spinor, being non-trivial, $\lambda \neq 0$, would not be null. The null condition (14.150) is a complex constraint, which eliminates 2 of the 8 degrees of freedom of a complex quaternion, so that a null spinor has 6 degrees of freedom. The null condition $\tilde{q}q = 0$ is equivalent to the two conditions

$$\tilde{\psi}\psi = \tilde{\psi}I\psi = 0 . \quad (14.151)$$

Any non-trivial null complex quaternion q can be written uniquely as the product of a real quaternion λU and a null factor $(1 - I\mathbf{n})$ (Exercise 14.1):

$$q = \lambda U(1 - I\mathbf{n}) . \quad (14.152)$$

Here λ is a positive real scalar, U is a purely spatial (i.e. real, with no I part) rotor, and $\mathbf{n} = \iota_a n_a$, $a = 1, 2, 3$, is a real unimodular vector quaternion, satisfying $n_a n_a = 1$ with real n_a . Physically, equation (14.152) contains the instruction to boost to light speed in the direction $\mathbf{n}$, then scale by the real scalar λ and rotate spatially by U . The minus sign in front of $I\mathbf{n}$ comes from the fact that a boost in direction $\mathbf{n}$ is described by a rotor $R = \cosh(\theta/2) - I\mathbf{n} \sinh(\theta/2)$, equation (14.42), which becomes proportional to $1 - I\mathbf{n}$ as the boost tends to infinity, $\theta \rightarrow \infty$. The $1 + 3 + 2 = 6$ degrees of freedom from the real scalar λ , the spatial rotor U , and the real unimodular vector $\mathbf{n}$ in the expression (14.152) are precisely the number needed to specify a null quaternion. The boost axis $\mathbf{n}$ is Lorentz-invariant. For if the boost factor $1 - I\mathbf{n}$ is Lorentz transformed by pre-multiplying by any complex quaternion $p + Ir$, then the result

$$(p + Ir)(1 - I\mathbf{n}) = (p + r\mathbf{n})(1 - I\mathbf{n}) \quad (14.153)$$

is the same unchanged boost factor $1 - I\mathbf{n}$ pre-multiplied by the real quaternion $p + r\mathbf{n}$, the latter being a product of a real scalar and a pure spatial rotation. Equation (14.153) is true because $\mathbf{n}^2 = -1$. The null Dirac spinor ψ corresponding to the null complex quaternion q , equation (14.152), is

$$\psi \equiv q \epsilon_{\uparrow} = \lambda U(1 - I\mathbf{n}) \epsilon_{\uparrow} . \quad (14.154)$$

The boost axis $\mathbf{n}$ specifies the direction of the boost relative to the spin rest frame, where the spin is pure up $\uparrow$. Because the boost axis $\mathbf{n}$ is Lorentz-invariant, Lorentz transforming a given null Dirac spinor fills out only 4 of the 6 degrees of freedom of null spinors.

Concept question 14.20 The boost axis of a null spinor is Lorentz-invariant. It may seem counter-intuitive that the boost axis $\mathbf{n}$ of a null spinor is Lorentz-invariant. Should not a spatial rotation rotate the boost direction, the direction in which the null spinor moves? **Answer.** The direction $\mathbf{n}$ specifies the direction of the boost axis relative to the spin axis. A Lorentz transformation of a null spinor effectively rotates both boost and spin directions simultaneously. For example, if the boost and spin axes are parallel in one frame, then they are parallel in any frame.

Equation (14.153) shows that a Lorentz transformation of the null Dirac spinor ψ (14.154) is equivalent to a scaling and a spatial rotation of that spinor,

$$(p + Ir)\psi = (p + Ir)\lambda U(1 - I\mathbf{n})\epsilon_{\uparrow} = (p + r\mathbf{n}')\psi, \quad (14.155)$$

where

$$\mathbf{n}' = U\mathbf{n}\tilde{U}. \quad (14.156)$$

The real quaternion $p + r\mathbf{n}'$ on the right hand side of equation (14.155) is not necessarily unimodular (a spatial rotor) even if the complex quaternion $p + Ir$ on the left hand side is unimodular (a Lorentz rotor). As a simple example, a Lorentz boost $e^{-I\mathbf{n}\theta/2}$ by rapidity θ along the boost axis $\mathbf{n}$, equation (14.42), multiplies the null spinor $(1 - I\mathbf{n})\epsilon_{\uparrow}$ by the real scalar $e^{\theta/2}$. Physically, when a null spinor is Lorentz transformed, it gets blueshifted (multiplied by a real scalar).

The spinor reverse to the spinor (14.154) is

$$\tilde{\psi} \equiv \epsilon_{\uparrow}^{\top} \tilde{q} = \epsilon_{\uparrow}^{\top} (1 + I\mathbf{n})\lambda \tilde{U}. \quad (14.157)$$

The spinor is null, $\tilde{q}q = 0$, because the boost factor is null, $(1 + I\mathbf{n})(1 - I\mathbf{n}) = 0$.

14.11.1 Weyl spinor

A Weyl spinor is a null Dirac spinor in the special case where the boost axis $\mathbf{n}$ in equation (14.154) aligns with the spin axis. For a right-handed spinor, the boost and spin axes point in the same direction. For a left-handed spinor, the boost and spin axes point in opposite directions. If the spin axis is taken along the positive 3-direction (z -axis), as in the Dirac representation (14.102), then for a right-handed spinor, the boost direction is $\mathbf{n} = \mathbf{e}_3$, while for a left-handed spinor, the boost direction is $\mathbf{n} = -\mathbf{e}_3$.

The bivector $\mathbf{e}_3$ generates a spatial rotation about the 3-axis, yielding, in the Dirac representation, i when acting on the spin-up eigenvector, $\mathbf{e}_3 \epsilon_{\uparrow} = i \epsilon_{\uparrow}$, equation (14.109). For right- and left-handed Weyl spinors, the null boost factor $1 - I\mathbf{e} \cdot \mathbf{n}$ acting on the rest-frame spinor $\epsilon_{\uparrow}$ becomes

$$(1 - I\mathbf{e} \cdot \mathbf{n})\epsilon_{\uparrow} = (1 \mp I\mathbf{e}_3)\epsilon_{\uparrow} = (1 \mp Ii)\epsilon_{\uparrow} = (1 \pm \gamma_5)\epsilon_{\uparrow}, \quad (14.158)$$

where $\gamma_5 \equiv -iI$ is the chiral operator. A general right- or left-handed Weyl spinor may be written uniquely as the right- or left-handed basis spinor defined by equation (14.158) pre-multiplied by a positive real scalar λ and a purely spatial rotor U ,

$$\psi_R = \lambda U(1 \pm \gamma_5)\epsilon_{\uparrow}. \quad (14.159)$$

A Weyl spinor has definite chirality, positive for a right-handed spinor ψ_R , negative for a left-handed spinor ψ_L ,

$$\gamma_5 \psi_R = \pm \psi_R. \quad (14.160)$$

The complex quaternionic components of the right- or left-handed basis Weyl spinors (14.158) are

$$(1 \mp I_3) \epsilon_{\uparrow} = (1 \pm \sigma_3) \epsilon_{\uparrow} = \begin{Bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pm 1 \end{Bmatrix}, \quad (14.161)$$

the Dirac representation of the bivector σ_3 being given by equations (14.103), which translates to a complex quaternion in accordance with the equivalence (14.117). If the components of the real quaternion in equation (14.159) are $\lambda U = \{w, x, y, z\}$, then the complex quaternionic components of the right- or left-handed Weyl spinor are

$$\psi_{\begin{smallmatrix} R \\ L \end{smallmatrix}} = \begin{Bmatrix} w & x & y & z \\ \mp z & \mp y & \pm x & \pm w \end{Bmatrix}. \quad (14.162)$$

Concept question 14.21 What makes Weyl spinors special? What is special about choosing the boost axis $\mathbf{n}$ of a null spinor to align with the spin axis? Why not consider null spinors with arbitrary boost axis $\mathbf{n}$? **Answer.** The property that the boost axis aligns with the spin axis is Lorentz invariant. If the boost aligns with the spin in one frame, then it does so in any Lorentz-transformed frame. This is the same thing as saying that chirality is a Lorentz invariant. In the Standard Model of Physics, Chapter 49, the fundamental fermions are natively massless right- or left-handed Weyl spinors. The fermions acquire their masses through interaction with a scalar Higgs field. Right- and left-handed fermions are distinctly different because only left-handed fermions (and right-handed antifermions) feel weak interactions.

The extension of the spacetime algebra to a super spacetime algebra, wherein the spacetime algebra of multivectors is shown to be isomorphic to the algebra of outer products of spinors, is resumed in Chapter 41.

Geometric Differentiation and Integration

The problem of integrating over a curved hypersurface crops up routinely in general relativity, for example in developing the Lagrangian or Hamiltonian mechanics of a field, Chapter 16. The apparatus developed by mathematicians to allow integration over curved hypersurfaces is called differential forms, §15.6. The geometric algebra provides an elegant way to understand differential forms.

In standard calculus, integration is inverse to differentiation. In the theory of differential forms, integration is inverse to something called exterior differentiation, §15.9. The exterior derivative, conventionally written d (distinguished here by latin font), is the (coordinate and tetrad) scalar derivative operator

$$d \equiv dx^\nu \frac{\partial}{\partial x^\nu} \wedge , \quad (15.1)$$

the wedge $\wedge$ signifying that the derivative is a curl. A more explicit definition of the exterior derivative is given by equation (15.63). A closely related derivative is the covariant spacetime derivative $\mathbf{D}$ defined by

$$\mathbf{D} \equiv e^\nu D_\nu = \gamma^n D_n , \quad (15.2)$$

where D_ν and D_n are respectively the coordinate- and tetrad-frame covariant derivatives. The exterior derivative d is isomorphic to the torsion-free covariant spacetime curl $\mathring{\mathbf{D}} \wedge$ (see equation (15.67) for a more precise statement of the isomorphism),

$$d \leftrightarrow \mathring{\mathbf{D}} \wedge . \quad (15.3)$$

The first part of this Chapter shows how to take the covariant derivative of a multivector, and defines the covariant spacetime derivative $\mathbf{D}$. The second part, starting from §15.6, shows how these ideas relate to differential forms and the exterior derivative, and derives the main result of the theory, the generalized Stokes' theorem.

If torsion is present, then the torsion-full covariant derivative differs from the torsion-free covariant derivative, equation (2.68). In sections 15.1–15.4, the covariant derivative D_n and the covariant spacetime derivative $\mathbf{D}$ signify either the torsion-full or the torsion-free derivative; all the results hold either way. In the theory of differential forms, however, starting at §15.6, the covariant spacetime derivative is the torsion-free derivative $\mathring{\mathbf{D}}$ even when torsion is present.

In this Chapter N denotes the dimension of the parent manifold in which the hypersurface of integration is embedded. In the standard spacetime of general relativity, N equals 4 and the signature is $-+++$, but all the results extend to manifolds of arbitrary dimension and arbitrary signature.

15.1 Covariant derivative of a multivector

The geometric algebra suggests an alternative approach to covariant differentiation in general relativity, in which the connection is treated as a vector of operators $\hat{\Gamma}_n$, the covariant derivative D_n being written

$$D_n = \partial_n + \hat{\Gamma}_n . \quad (15.4)$$

Acting on any object, the connection operator $\hat{\Gamma}_n$ generates a Lorentz transformation.

In the spacetime algebra, a Lorentz transformation (13.48) by rotor R transforms a multivector $\mathbf{a}$ by $\mathbf{a} \rightarrow R\mathbf{a}\tilde{R}$. The generator of a Lorentz transformation is a bivector. The rotor corresponding to an infinitesimal Lorentz transformation generated by a bivector $\mathbf{\Gamma}$ is $R = e^{\epsilon\mathbf{\Gamma}/2} = 1 + \frac{1}{2}\epsilon\mathbf{\Gamma}$. The resulting infinitesimal Lorentz transformation transforms the multivector $\mathbf{a}$ by $\mathbf{a} \rightarrow \mathbf{a} + \frac{1}{2}\epsilon[\mathbf{\Gamma}, \mathbf{a}]$, where $[\mathbf{\Gamma}, \mathbf{a}] \equiv \mathbf{\Gamma}\mathbf{a} - \mathbf{a}\mathbf{\Gamma}$ is the commutator. It follows that the action of the connection operator $\hat{\Gamma}_n$ on a multivector $\mathbf{a}$ must take the form

$$\hat{\Gamma}_n \mathbf{a} = \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a}] \quad (15.5)$$

for some set of bivectors $\mathbf{\Gamma}_n$. Since rotation does not change the grade of a multivector, $[\mathbf{\Gamma}_n, \mathbf{a}]$ for each n is a multivector with the same grade as $\mathbf{a}$.

Concept question 15.1 Commutator versus wedge product of multivectors. Is the commutator $\frac{1}{2}[\mathbf{a}, \mathbf{b}]$ of two multivectors the same as their wedge product $\mathbf{a} \wedge \mathbf{b}$? **Answer.** No. In the first place, the wedge product anticommutes only if both $\mathbf{a}$ and $\mathbf{b}$ have odd grade, equation (13.32). In the second place, the anti-commutator selects all grade components of the geometric product that anticommute, per equation (13.28). The only case where $\mathbf{a} \wedge \mathbf{b} = \frac{1}{2}[\mathbf{a}, \mathbf{b}]$ is true is where either $\mathbf{a}$ or $\mathbf{b}$ is a vector (a multivector of grade 1), and both $\mathbf{a}$ and $\mathbf{b}$ are odd.

To establish the relation between the bivectors $\mathbf{\Gamma}_n$ and the usual tetrad connections Γ_{kmn} , consider the covariant derivative of the vector $\mathbf{a} = a^m \boldsymbol{\gamma}_m$:

$$D_n \mathbf{a} = \partial_n \mathbf{a} + \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a}] = \boldsymbol{\gamma}_m \partial_n a^m + \frac{1}{2}[\mathbf{\Gamma}_n, \boldsymbol{\gamma}_m] a^m . \quad (15.6)$$

Notice that the directed derivative ∂_n in equation (15.6) is to be interpreted as acting only on the components a^m of the vector, not on the tetrad $\boldsymbol{\gamma}_m$; rather, the variation of the tetrad under parallel transport is embodied in the $\frac{1}{2}[\mathbf{\Gamma}_n, \boldsymbol{\gamma}_m]$ term. The expression (15.6) must agree with the expression (11.35) obtained in the earlier treatment, namely

$$D_n \mathbf{a} = \boldsymbol{\gamma}_m \partial_n a^m + \Gamma_{mn}^k \boldsymbol{\gamma}_k a^m . \quad (15.7)$$

Comparison of equations (15.6) and (15.7) shows that

$$\frac{1}{2}[\mathbf{\Gamma}_n, \boldsymbol{\gamma}_m] = \Gamma_{mn}^k \boldsymbol{\gamma}_k . \quad (15.8)$$

The N -tuple (not vector) of bivectors $\mathbf{\Gamma}_n$ satisfying equation (15.8) is

$$\boxed{\mathbf{\Gamma}_n \equiv \frac{1}{2} \Gamma_{kl n} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l} \quad (15.9)$$

(the factor of $\frac{1}{2}$ would disappear if the implicit summation were over distinct antisymmetric pairs kl of indices). Equation (15.9) can be proved with the help of the identity

$$\frac{1}{2}[\boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l, \boldsymbol{\gamma}_m] = \delta_m^l \boldsymbol{\gamma}^k - \delta_m^k \boldsymbol{\gamma}^l . \quad (15.10)$$

The same formula (15.5) applies, with the bivector $\mathbf{\Gamma}_n$ given by the same equation (15.9), if the vector $\mathbf{a}$ is expressed as a sum $\mathbf{a} = a_m \boldsymbol{\gamma}^m$ over its covariant a_m rather than contravariant a^m components. In this case

$$D_n \mathbf{a} = \boldsymbol{\gamma}^m \partial_n a_m + \frac{1}{2}[\mathbf{\Gamma}_n, \boldsymbol{\gamma}^m] a_m , \quad (15.11)$$

which reproduces the earlier equations (11.40) and (11.41),

$$D_n \mathbf{a} = \boldsymbol{\gamma}^m \partial_n a_m - \Gamma_{kn}^m \boldsymbol{\gamma}^k a_m , \quad (15.12)$$

since

$$\frac{1}{2}[\mathbf{\Gamma}_n, \boldsymbol{\gamma}^m] = -\Gamma_{kn}^m \boldsymbol{\gamma}^k . \quad (15.13)$$

The same formula (15.5) with the same bivector (15.9) applies to any multivector, which follows because the connection operator $\hat{\Gamma}_n$ is additive over any product of vectors or multivectors:

$$\hat{\Gamma}_n \mathbf{a} \mathbf{b} = \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a} \mathbf{b}] = \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a}] \mathbf{b} + \frac{1}{2} \mathbf{a} [\mathbf{\Gamma}_n, \mathbf{b}] = (\hat{\Gamma}_n \mathbf{a}) \mathbf{b} + \mathbf{a} (\hat{\Gamma}_n \mathbf{b}) . \quad (15.14)$$

To summarize, the covariant derivative of a multivector $\mathbf{a}$ can be written

$$\boxed{D_n \mathbf{a} = \partial_n \mathbf{a} + \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a}]} , \quad (15.15)$$

with the N -tuple of bivectors $\mathbf{\Gamma}_n$ given by equation (15.9). In equation (15.15), as previously in equation (15.6), for a multivector $\mathbf{a} = \boldsymbol{\gamma}_A a^A$, the directed derivative ∂_n is to be interpreted as acting only on the components a^A of the multivector, $\partial_n \mathbf{a} = \boldsymbol{\gamma}_A \partial_n a^A$. Equation (15.15) is just another way to write the covariant derivative of a multivector, yielding exactly the same result as the earlier method from §11.9.

The earlier (§11.9) and multivector approaches to covariant differentiation can be combined as needed. For example, the covariant derivative of a covariant vector $\mathbf{a}_m$ of multivectors is

$$D_n \mathbf{a}_m = \partial_n \mathbf{a}_m - \Gamma_{mn}^k \mathbf{a}_k + \frac{1}{2}[\mathbf{\Gamma}_n, \mathbf{a}_m] . \quad (15.16)$$

As always, covariant differentiation is defined so that it commutes with the tetrad basis elements; that is, covariant derivatives of the tetrad basis elements vanish by construction,

$$D_n \boldsymbol{\gamma}_m = 0 . \quad (15.17)$$

For example, equation (15.17) is implied by the equality

$$D_n(\gamma_m a^m) = \gamma_m D_n a^m , \quad (15.18)$$

which is true by construction.

The covariant derivative of a multivector $\mathbf{a}$ can also be expressed as a coordinate derivative

$$D_\nu \mathbf{a} = \frac{\partial \mathbf{a}}{\partial x^\nu} + \frac{1}{2} [\Gamma_\nu, \mathbf{a}] , \quad (15.19)$$

where the coordinate and directed derivatives are related as usual by $\partial/\partial x^\nu = e^\nu_\nu \partial_n$, and where the connection vector Γ_ν is related to the tetrad connection Γ_n defined by equation (15.9) by

$$\Gamma_\nu \equiv e^\nu_\nu \Gamma_n . \quad (15.20)$$

The components $\Gamma_{kl\nu} \equiv e^\nu_\nu \Gamma_{kln}$ of $\Gamma_\nu \equiv \frac{1}{2} \Gamma_{kl\nu} \gamma^k \wedge \gamma^l$ constitute a coordinate-frame vector, but not a tetrad-frame tensor. The connection Γ_ν is given by equation (15.20), *not* by a direct relation to the coordinate-frame connections $\Gamma_{\mu\nu\kappa}$, that is, $\Gamma_\nu \neq \frac{1}{2} \Gamma_{\kappa\lambda\nu} e^\kappa \wedge e^\lambda$.

15.2 Riemann tensor of bivectors

As discussed in §2.19.2, the commutator of the covariant derivative defines two fundamental geometric objects, the torsion tensor S_{kl}^n and the Riemann curvature tensor R_{klmn} . The commutator can be written

$$[D_k, D_l] = S_{kl}^n D_n + \hat{R}_{kl} , \quad (15.21)$$

where S_{kl}^n is the torsion tensor, and the Riemann curvature operator $\hat{R}_{kl}$ is an operator whose action on any tensor was given previously by equation (2.114). Define the Riemann antisymmetric tensor of bivectors $\mathbf{R}_{kl}$ by

$$\boxed{\mathbf{R}_{kl} \equiv \frac{1}{2} R_{klmn} \gamma^m \wedge \gamma^n} \quad (15.22)$$

(again, the factor of $\frac{1}{2}$ would disappear if the implicit summation were over distinct antisymmetric pairs mn of indices). Acting on any multivector $\mathbf{a}$, the Riemann curvature operator yields

$$\hat{R}_{kl} \mathbf{a} = \frac{1}{2} [\mathbf{R}_{kl}, \mathbf{a}] , \quad (15.23)$$

which is an antisymmetric tensor of multivectors of the same grade as $\mathbf{a}$. The Riemann tensor of bivectors $\mathbf{R}_{kl}$, equation (15.22), is related to the connection N -tuple of bivectors Γ_k , equation (15.9), by

$$\mathbf{R}_{kl} = \partial_k \Gamma_l - \partial_l \Gamma_k + \frac{1}{2} [\Gamma_k, \Gamma_l] + (\Gamma_{kl}^m - \Gamma_{lk}^m - S_{kl}^m) \Gamma_m , \quad (15.24)$$

where, in conformity with the convention of equation (15.15), directed derivatives $\partial_k \Gamma_l$ are to be interpreted as acting only on the components Γ_{mnl} of $\Gamma_l \equiv \frac{1}{2} \Gamma_{mnl} \gamma^m \wedge \gamma^n$, not on the tetrad axes γ_m . Equation (15.24) can be derived either from the tetrad-frame expression (11.60) for the Riemann tensor, or from the expression (15.15) for the covariant derivative of a multivector.

Transforming equation (15.24) into a coordinate frame, $\mathbf{R}_{\kappa\lambda} = e^k_{\kappa} e^l_{\lambda} \mathbf{R}_{kl}$, and substituting equation (11.76) gives, with or without torsion, the elegant expression

$$\boxed{\mathbf{R}_{\kappa\lambda} = \frac{\partial \Gamma_{\lambda}}{\partial x^{\kappa}} - \frac{\partial \Gamma_{\kappa}}{\partial x^{\lambda}} + \frac{1}{2}[\Gamma_{\kappa}, \Gamma_{\lambda}]}, \quad (15.25)$$

which can also be written as the commutator

$$\frac{1}{2}\mathbf{R}_{\kappa\lambda} = \left[\frac{\partial}{\partial x^{\kappa}} + \frac{1}{2}\Gamma_{\kappa}, \frac{\partial}{\partial x^{\lambda}} + \frac{1}{2}\Gamma_{\lambda} \right]. \quad (15.26)$$

Equation (15.25) is Cartan's second equation of structure, explored in depth in §16.14.2. The components of $\mathbf{R}_{\kappa\lambda} = \frac{1}{2}R_{\kappa\lambda mn}\boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n$ constitute the Riemann tensor $R_{\kappa\lambda mn}$ in the mixed coordinate-tetrad basis, equation (11.77).

15.3 Torsion tensor of vectors

Define the torsion antisymmetric tensor of vectors $\mathbf{S}_{\kappa\lambda}$ by (the minus sign is chosen so that equation (15.29) resembles equation (15.25))

$$\mathbf{S}_{\kappa\lambda} \equiv -S_{m\kappa\lambda} \boldsymbol{\gamma}^m. \quad (15.27)$$

In components, the torsion tensor of vectors $\mathbf{S}_{\kappa\lambda}$ is, from equation (11.49),

$$\mathbf{S}_{\kappa\lambda} = \left(\frac{\partial e^m_{\lambda}}{\partial x^{\kappa}} - \frac{\partial e^m_{\kappa}}{\partial x^{\lambda}} + \Gamma_{\kappa\lambda}^m - \Gamma_{\lambda\kappa}^m \right) \boldsymbol{\gamma}_m, \quad (15.28)$$

which can be written elegantly

$$\boxed{\mathbf{S}_{\kappa\lambda} = \frac{\partial \mathbf{e}_{\lambda}}{\partial x^{\kappa}} - \frac{\partial \mathbf{e}_{\kappa}}{\partial x^{\lambda}} + \frac{1}{2}[\Gamma_{\kappa}, \mathbf{e}_{\lambda}] - \frac{1}{2}[\Gamma_{\lambda}, \mathbf{e}_{\kappa}]}, \quad (15.29)$$

where $\mathbf{e}_{\kappa} \equiv e^k_{\kappa} \boldsymbol{\gamma}_k$ are the usual tangent basis vectors, and again the coordinate derivative $\partial/\partial x^{\kappa}$ is to be interpreted as acting only on the components e^k_{κ} of $\mathbf{e}_{\kappa}$, not on the tetrad axes $\boldsymbol{\gamma}_k$. Equation (15.29) is Cartan's first equation of structure, §16.14.2. Equation (15.29) can also be written in terms of covariant derivatives

$$\mathbf{S}_{\kappa\lambda} = D_{\kappa} \mathbf{e}_{\lambda} - D_{\lambda} \mathbf{e}_{\kappa}. \quad (15.30)$$

15.4 Covariant spacetime derivative

The covariant derivative D_n , equation (15.4), acts on multivectors, but it does not yield a multivector (it yields a vector of multivectors). A covariant derivative that does map multivectors to multivectors is the **covariant spacetime derivative** $\mathbf{D}$ defined by

$$\boxed{\mathbf{D} \equiv \boldsymbol{\gamma}^n D_n}. \quad (15.31)$$

The covariant spacetime derivative $\mathbf{D}$ is a sum of a directed derivative ∂ and a connection operator $\hat{\Gamma}$,

$$\mathbf{D} = \partial + \hat{\Gamma} , \quad \partial \equiv \gamma^n \partial_n , \quad \hat{\Gamma} \equiv \gamma^n \hat{\Gamma}_n . \quad (15.32)$$

The action of the connection operator $\hat{\Gamma}$ on a multivector $\mathbf{a}$ is

$$\hat{\Gamma} \mathbf{a} = \frac{1}{2} \gamma^n [\Gamma_n, \mathbf{a}] \quad (15.33)$$

(not $\hat{\Gamma} \mathbf{a} = \frac{1}{2} [\Gamma, \mathbf{a}]$). The covariant spacetime derivative of a multivector $\mathbf{a}$ is

$$\mathbf{D} \mathbf{a} = \gamma^n D_n \mathbf{a} = \gamma^n \left(\partial_n \mathbf{a} + \frac{1}{2} [\Gamma_n, \mathbf{a}] \right) . \quad (15.34)$$

The covariant spacetime derivative (15.31) can equally well be written in terms of the coordinate derivatives,

$$\mathbf{D} \equiv e^\nu D_\nu . \quad (15.35)$$

The covariant spacetime derivative (15.34) of a multivector can then also be written

$$\mathbf{D} \mathbf{a} = e^\nu D_\nu \mathbf{a} = e^\nu \left(\frac{\partial \mathbf{a}}{\partial x^\nu} + \frac{1}{2} [\Gamma_\nu, \mathbf{a}] \right) . \quad (15.36)$$

Acting on a multivector $\mathbf{a}$, the covariant spacetime derivative $\mathbf{D}$ yields a sum of two multivectors, a covariant divergence $\mathbf{D} \cdot \mathbf{a}$ with one grade lower than $\mathbf{a}$, and a covariant curl $\mathbf{D} \wedge \mathbf{a}$ with one grade higher than $\mathbf{a}$,

$$\mathbf{D} \mathbf{a} = \mathbf{D} \cdot \mathbf{a} + \mathbf{D} \wedge \mathbf{a} \quad \text{multivector} . \quad (15.37)$$

In the particular case that $\mathbf{a}$ is a scalar a (a multivector of grade 0), the covariant divergence (defined to be one grade lower than a) is zero, $\mathbf{D} \cdot a = 0$. If torsion vanishes, the curl $\mathbf{D} \wedge \mathbf{a}$ is essentially the same as the exterior derivative in the mathematics of differential forms, §15.9.

The covariant spacetime divergence and curl of a grade- p multivector $\mathbf{a} = (1/p!) \gamma^{lm\dots n} a_{lm\dots n}$ are

$$\mathbf{D} \cdot \mathbf{a} = \frac{1}{(p-1)!} \gamma^{m\dots n} (\mathbf{D} \cdot \mathbf{a})_{m\dots n} , \quad (\mathbf{D} \cdot \mathbf{a})_{m\dots n} = D^l a_{lm\dots n} , \quad (15.38a)$$

$$\mathbf{D} \wedge \mathbf{a} = \frac{1}{(p+1)!} \gamma^{klm\dots n} (\mathbf{D} \wedge \mathbf{a})_{klm\dots n} , \quad (\mathbf{D} \wedge \mathbf{a})_{klm\dots n} = (p+1) D_{[k} a_{lm\dots n]} . \quad (15.38b)$$

The factorial factors could be dropped if the implicit summations were taken over distinct antisymmetric sequences of indices, but are retained here for explicitness. For example, the components of the covariant divergences and curls of a scalar φ , a vector $\mathbf{A} = \gamma^n A_n$, and a bivector $\mathbf{F} = \frac{1}{2} \gamma^m \wedge \gamma^n F_{mn}$, are respectively

$$\mathbf{D} \cdot \varphi = 0 , \quad (\mathbf{D} \wedge \varphi)_n = D_n \varphi = \partial_n \varphi , \quad (15.39a)$$

$$\mathbf{D} \cdot \mathbf{A} = D^n A_n , \quad (\mathbf{D} \wedge \mathbf{A})_{mn} = D_m A_n - D_n A_m , \quad (15.39b)$$

$$(\mathbf{D} \cdot \mathbf{F})_n = D^m F_{mn} , \quad (\mathbf{D} \wedge \mathbf{F})_{lmn} = D_l F_{mn} + D_m F_{nl} + D_n F_{lm} . \quad (15.39c)$$

A divergence can be converted to a curl, and vice versa, by post-multiplying by the pseudoscalar I_N ,

$$\mathbf{D} \wedge (\mathbf{a} I_N) = (\mathbf{D} \cdot \mathbf{a}) I_N , \quad \mathbf{D} \cdot (\mathbf{a} I_N) = (\mathbf{D} \wedge \mathbf{a}) I_N , \quad (15.40)$$

which works because the pseudoscalar I_N is covariantly constant, and multiplying by it flips the grade of a multivector from p to $N-p$.

The curl of the wedge product of a grade- p multivector $\mathbf{a}$ with a multivector $\mathbf{b}$ satisfies the Leibniz-like rule

$$\mathbf{D} \wedge (\mathbf{a} \wedge \mathbf{b}) = (\mathbf{D} \wedge \mathbf{a}) \wedge \mathbf{b} + (-)^p \mathbf{a} \wedge (\mathbf{D} \wedge \mathbf{b}) . \quad (15.41)$$

The square of the covariant spacetime derivative is

$$\mathbf{D}\mathbf{D} = \mathbf{D} \cdot \mathbf{D} + \mathbf{D} \wedge \mathbf{D} = D_k D^k + \frac{1}{2} \gamma^k \wedge \gamma^l [D_k, D_l] , \quad (15.42)$$

which is a sum of the scalar d'Alembertian wave operator $\square \equiv D_k D^k$, and a bivector operator whose components constitute the commutator of the covariant derivative, equation (15.21).

For vanishing torsion, the squared spacetime curl of a multivector $\mathbf{a}$ vanishes. For example, for a grade 1 multivector $\mathbf{a} = a^n \gamma_n$,

$$\mathbf{D} \wedge \mathbf{D} \wedge \mathbf{a} = \frac{1}{2} \gamma^k \wedge \gamma^l \wedge \gamma^m R_{klmn} a^n = 0 , \quad (15.43)$$

which vanishes thanks to the cyclic symmetry of the Riemann tensor, $R_{[klm]n} = 0$, valid for vanishing torsion.

Exercise 15.2 Leibniz rule for the covariant spacetime derivative.

1. What is the covariant derivative $D_m(\mathbf{a}\mathbf{b})$ of a geometric product of multivectors $\mathbf{a}$ and $\mathbf{b}$ in terms of covariant derivatives of each of $\mathbf{a}$ and $\mathbf{b}$?
2. What is the covariant spacetime derivative $\mathbf{D}(\mathbf{a}\mathbf{b})$ of a geometric product of multivectors $\mathbf{a}$ and $\mathbf{b}$ in terms of covariant spacetime derivatives of each of $\mathbf{a}$ and $\mathbf{b}$?

Solution.

1. The covariant derivative $D_m(\mathbf{a}\mathbf{b})$ satisfies the Leibniz rule

$$D_m(\mathbf{a}\mathbf{b}) = (D_m \mathbf{a})\mathbf{b} + \mathbf{a}D_m \mathbf{b} . \quad (15.44)$$

2. If $\mathbf{a}$ is a multivector of grade p , then the covariant spacetime derivative $\mathbf{D}(\mathbf{a}\mathbf{b})$ satisfies the Leibniz-like rule

$$\mathbf{D}(\mathbf{a}\mathbf{b}) \equiv \gamma^m D_m(\mathbf{a}\mathbf{b}) = \gamma^m ((D_m \mathbf{a})\mathbf{b} + \mathbf{a}D_m \mathbf{b}) = (\mathbf{D}\mathbf{a})\mathbf{b} + (-)^p (-(\mathbf{a} \cdot \mathbf{D})\mathbf{b} + (\mathbf{a} \wedge \mathbf{D})\mathbf{b}) . \quad (15.45)$$

15.5 Torsion-full and torsion-free covariant spacetime derivative

The results of §15.1–§15.4 hold with or without torsion.

As in §2.12 and §11.15, when torsion is present and one wishes to make the torsion part explicit, it is convenient to distinguish torsion-free quantities with a $\circ$ overscript. The torsion-full and torsion-free connection N -tuples $\mathbf{\Gamma}_n$ and $\mathbf{\overset{\circ}{\Gamma}}_n$ are related by

$$\mathbf{\Gamma}_n = \mathbf{\overset{\circ}{\Gamma}}_n + \mathbf{K}_n , \quad (15.46)$$

where the contortion vector of bivectors $\mathbf{K}_n$ is defined, analogously to equation (15.9), in terms of the contortion tensor K_{kln} equation (11.56), by

$$\mathbf{K}_n = \frac{1}{2} K_{kln} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l, \quad (15.47)$$

implicitly summed over distinct indices k and l . Acting on a multivector $\mathbf{a}$, the torsion-full and torsion-free covariant spacetime derivatives $\mathbf{D}$ and $\mathring{\mathbf{D}}$ are related by

$$\mathbf{D}\mathbf{a} = \mathring{\mathbf{D}}\mathbf{a} + \frac{1}{2} \boldsymbol{\gamma}^n [\mathbf{K}_n, \mathbf{a}]. \quad (15.48)$$

From equation (15.25), the Riemann tensor of bivectors $\mathbf{R}_{\kappa\lambda}$ splits into torsion-free and contortion parts,

$$\begin{aligned} \mathbf{R}_{\kappa\lambda} &= \frac{\partial(\mathring{\Gamma}_\lambda + \mathbf{K}_\lambda)}{\partial x^\kappa} - \frac{\partial(\mathring{\Gamma}_\kappa + \mathbf{K}_\kappa)}{\partial x^\lambda} + \frac{1}{2} [\mathring{\Gamma}_\kappa + \mathbf{K}_\kappa, \mathring{\Gamma}_\lambda + \mathbf{K}_\lambda] \\ &= \mathring{\mathbf{R}}_{\kappa\lambda} + \mathring{\mathbf{D}}_\kappa \mathbf{K}_\lambda - \mathring{\mathbf{D}}_\lambda \mathbf{K}_\kappa + \frac{1}{2} [\mathbf{K}_\kappa, \mathbf{K}_\lambda]. \end{aligned} \quad (15.49)$$

15.6 Differential forms

Differential forms, or p -forms, pioneered by Cartan (1945), are invariant measures of integration over a p -dimensional hypersurface in an N -dimensional manifold. In §13.1 it was seen that the wedge product of p vectors defines a directed p -dimensional volume, illustrated in Figure 13.1. A p -form is essentially the same thing, but with the vectors taken to be infinitesimals. The purpose of p -forms is to allow integration over p -dimensional hypersurfaces in a coordinate-independent fashion. By construction, a differential form is a coordinate (and tetrad) scalar, as is essential for integration to be coordinate-independent.

In an N -dimensional manifold with coordinates x^μ , a 1-form expressed in the coordinate frame is

$$\mathbf{a} = a_\mu dx^\mu. \quad (15.50)$$

By definition, the differential dx^μ transforms under coordinate transformations like a contravariant coordinate vector. Requiring that the 1-form $\mathbf{a}$ defined by equation (15.50) be a coordinate scalar imposes that a_μ must be a covariant coordinate vector. When the 1-form $\mathbf{a}$ is integrated over any line (= 1-dimensional hypersurface) in the manifold, the result is independent of the choice of coordinates, as desired.

A 2-form expressed in a coordinate frame is

$$\mathbf{a} = \frac{1}{2} a_{\mu\nu} dx^\mu \wedge dx^\nu, \quad (15.51)$$

implicitly summed over all antisymmetric pairs $\mu\nu$. The factor of $\frac{1}{2}$ cancels the double-counting of pairs, ensuring that each distinct antisymmetric pair $\mu\nu$ counts once. The factor of $\frac{1}{2}$ could be omitted if the implicit sum were taken over only distinct antisymmetric pairs $\mu\nu$. The wedge product $dx^\mu \wedge dx^\nu$ of differentials defines a parallelogram, a directed infinitesimal element of area, whose 2-dimensional direction is the $(dx^\mu - dx^\nu)$ -plane, and whose magnitude is the area of the parallelogram. The wedge product is antisymmetric,

$$dx^\mu \wedge dx^\nu = -dx^\nu \wedge dx^\mu. \quad (15.52)$$

The wedge product $dx^\mu \wedge dx^\nu$ transforms as an antisymmetric contravariant rank-2 coordinate tensor. Requiring that the 2-form $\mathbf{a}$ defined by equation (15.51) be a coordinate scalar imposes that $a_{\mu\nu}$ must be a covariant rank-2 coordinate tensor, which can be taken to be antisymmetric without loss of generality. To see that the antisymmetric prescription recovers correctly the usual behaviour of areal elements of integration, consider the particular case where the 2-dimensional surface of integration is spanned by just two coordinates, x and y , all other coordinates being constant on the surface. Under a coordinate transformation $\{x, y\} \rightarrow \{x', y'\}$, the wedge product of differentials transforms as

$$dx' \wedge dy' = \left(\frac{\partial x'}{\partial x} dx + \frac{\partial x'}{\partial y} dy \right) \wedge \left(\frac{\partial y'}{\partial x} dx + \frac{\partial y'}{\partial y} dy \right) = \left(\frac{\partial x'}{\partial x} \frac{\partial y'}{\partial y} - \frac{\partial x'}{\partial y} \frac{\partial y'}{\partial x} \right) dx \wedge dy . \quad (15.53)$$

The factor relating the two areal elements is the familiar Jacobian determinant $|\partial\{x', y'\}/\partial\{x, y\}|$. The definition (15.51) of the 2-form $\mathbf{a}$ is by construction coordinate-invariant, and is therefore valid when more than 2 coordinates vary over the surface of integration. However, it is always possible to erect a local coordinate system in which only 2 of the coordinates vary over the 2-dimensional surface of integration.

In general, a p -form expressed in a coordinate frame is

$$\mathbf{a} = \frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} . \quad (15.54)$$

The factor of $1/p!$ ensures that each distinct index sequence $\mu_1 \dots \mu_p$ is counted only once. The $1/p!$ factor could be dropped if the implicit sum were taken over distinct antisymmetric sequences of indices. Thus equation (15.54) could also be written

$$\mathbf{a} = a_\Lambda dx^\Lambda , \quad (15.55)$$

where the sum is only over distinct antisymmetric sequences Λ of p indices. The wedge product $dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}$ of differentials is totally antisymmetric. It transforms like an antisymmetric contravariant rank- p tensor. Requiring that the p -form $\mathbf{a}$ defined by equation (15.54) be coordinate-invariant imposes that $a_{\mu_1 \dots \mu_p}$ be a (without loss of generality antisymmetric) covariant rank- p coordinate tensor.

The definition (15.54) of a p -form extends to the case $p = 0$. A 0-form is simply a scalar a .

15.7 Differential forms in an arbitrary frame

Differential forms are not restricted to coordinate frames. In any arbitrary tetrad frame, which may or may not be a coordinate frame, and which may or may not be orthonormal, the invariant expression (15.55) for a p -form may be written

$$\mathbf{a} = a_K dx^K , \quad (15.56)$$

implicitly summed over distinct antisymmetric sequences K of p tetrad indices. Coordinate indices $\Lambda = \kappa \dots \lambda$ are converted in the usual way to tetrad indices $K = k \dots l$ using the vielbein $e^k{}_\kappa$ and its inverse $e_\kappa{}^k$,

$$a_{k \dots l} = e_k{}^\kappa \dots e_l{}^\lambda a_{\kappa \dots \lambda} , \quad dx^{k \dots l} = e^k{}_\kappa \dots e^l{}_\lambda dx^{\kappa \dots \lambda} . \quad (15.57)$$

The entire apparatus of differential forms translates into any arbitrary frame.

15.8 Wedge product of differential forms

The wedge product of differential forms is defined consistent with the wedge product of multivectors, equation (13.31). The wedge product of a 1-form $\mathbf{a}$ with a 1-form $\mathbf{b}$ defines a 2-form

$$\mathbf{a} \wedge \mathbf{b} = a_{\mu} dx^{\mu} \wedge b_{\nu} dx^{\nu} = a_{[\mu} b_{\nu]} dx^{\mu} \wedge dx^{\nu} = \frac{1}{2} (a_{\mu} b_{\nu} - a_{\nu} b_{\mu}) dx^{\mu} \wedge dx^{\nu} , \quad (15.58)$$

implicitly summed over both indices μ and ν . If instead the implicit sum were taken over distinct antisymmetric pairs $\mu\nu$ of indices, then there would be an extra factor of 2 in the third expression, and the $\frac{1}{2}$ in the last expression would disappear. In general, the wedge product of a p -form $\mathbf{a}$ with a q -form $\mathbf{b}$ defines a $(p+q)$ -form $\mathbf{a} \wedge \mathbf{b}$,

$$\begin{aligned} \mathbf{a} \wedge \mathbf{b} &= \left(\frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \right) \wedge \left(\frac{1}{q!} b_{\nu_1 \dots \nu_q} dx^{\nu_1} \wedge \dots \wedge dx^{\nu_q} \right) \\ &= \frac{1}{p!q!} a_{[\mu_1 \dots \mu_p} b_{\nu_1 \dots \nu_q]} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \wedge dx^{\nu_1} \wedge \dots \wedge dx^{\nu_q} . \end{aligned} \quad (15.59)$$

If the forms are expressed as sums $\mathbf{a} \equiv a_{\Lambda} d^p x^{\Lambda}$ and $\mathbf{b} \equiv b_{\Pi} d^q x^{\Pi}$ over distinct antisymmetric sequences Λ and Π of respectively p and q indices, then their wedge product is

$$\mathbf{a} \wedge \mathbf{b} = a_{\Lambda} b_{\Pi} d^{p+q} x^{\Lambda\Pi} = \frac{(p+q)!}{p!q!} a_{[\Lambda} b_{\Pi]} d^{p+q} x^{\Lambda\Pi} , \quad (15.60)$$

where the second expression is implicitly summed over distinct antisymmetric sequences Λ and Π of p and q indices, while the last expression is implicitly summed over distinct antisymmetric sequences $\Lambda\Pi$ of $p+q$ indices.

The wedge product is symmetric or antisymmetric as pq is even or odd,

$$\mathbf{a} \wedge \mathbf{b} = (-1)^{pq} \mathbf{b} \wedge \mathbf{a} , \quad (15.61)$$

consistent with the wedge product (13.31) of two multivectors.

The wedge product of a 0-form (scalar) a with a differential form $\mathbf{b}$ is just their ordinary product,

$$a \wedge \mathbf{b} = a\mathbf{b} \quad \text{if } a \text{ is a scalar} , \quad (15.62)$$

consistent with the result (13.34) for multivectors.

15.9 Exterior derivative

The exterior derivative of a differential form is constructed so that integration and exterior differentiation are inverse to each other, §15.12. In the abstract language of differential forms, the exterior derivative is denoted d , and the exterior derivative of a p -form $\mathbf{a}$ is the $(p+1)$ -form $d\mathbf{a}$ defined by

$$d\mathbf{a} \equiv d \left(\frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \right) \equiv \frac{1}{p!} \frac{\partial a_{\mu_1 \dots \mu_p}}{\partial x^{\nu}} dx^{\nu} \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} , \quad (15.63)$$

in which the left and right hand sides are implicitly summed over all antisymmetric sets of indices $\mu_1 \dots \mu_p$ and $\nu \mu_1 \dots \mu_p$ respectively. Equation (15.63) makes explicit the meaning of the definition (15.1) of the exterior derivative d . Thanks to the antisymmetry of the wedge product of differentials, the exterior derivative (15.63) may be rewritten

$$d\mathbf{a} = \frac{1}{p!} \partial_{[\nu} a_{\mu_1 \dots \mu_p]} dx^\nu \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} , \quad (15.64)$$

where $\partial_\nu \equiv \partial/\partial x^\nu$. If the p -form is expressed as a sum $\mathbf{a} \equiv a_\Lambda dx^\Lambda$ over distinct antisymmetric sequences Λ of p indices, then its exterior derivative is the $(p+1)$ -form

$$d\mathbf{a} = \partial_\nu a_\Lambda dx^{\nu\Lambda} = (p+1) \partial_{[\nu} a_{\Lambda]} dx^{\nu\Lambda} , \quad (15.65)$$

where the second expression is implicitly summed over indices ν and over distinct antisymmetric sequences Λ of p indices, while the last expression is implicitly summed over distinct antisymmetric sequences $\nu\Lambda$ of $p+1$ indices.

The antisymmetrized coordinate derivative is just equal to the antisymmetrized torsion-free covariant derivative (Exercise 2.7),

$$\partial_{[\nu} a_{\mu_1 \dots \mu_p]} = \mathring{D}_{[\nu} a_{\mu_1 \dots \mu_p]} , \quad (15.66)$$

which is true even when torsion is present (that is, the antisymmetrized coordinate derivative equals the antisymmetrized torsion-free covariant derivative, not the antisymmetrized torsion-full covariant derivative). The antisymmetrized coordinate derivative is a covariant coordinate tensor despite the fact that the derivatives are coordinate not covariant derivatives, and this is true whether or not torsion is present. Thus the exterior derivative $d\mathbf{a}$ is coordinate-invariant, with or without torsion. In an arbitrary frame, not necessarily a coordinate frame or an orthonormal frame, the exterior derivative of a p -form $\mathbf{a}$ is its torsion-free covariant curl,

$$d\mathbf{a} = (p+1) \mathring{D}_{[n} a_{K]} dx^{nK} , \quad (15.67)$$

implicitly summed over distinct antisymmetric sequences nK of $p+1$ tetrad indices.

The simplest case is the exterior derivative of a 0-form (scalar) φ , which according to the definition (15.63) is the one-form

$$d\varphi \equiv \frac{\partial \varphi}{\partial x^\nu} dx^\nu . \quad (15.68)$$

The next simplest case is the exterior derivative of a one-form $\mathbf{a}$, which according to the definition (15.63) is the 2-form

$$\begin{aligned} d\mathbf{a} &= d(a_\nu dx^\nu) = \frac{\partial a_\nu}{\partial x^\mu} dx^\mu \wedge dx^\nu \\ &= \frac{1}{2} \left(\frac{\partial a_\nu}{\partial x^\mu} - \frac{\partial a_\mu}{\partial x^\nu} \right) dx^\mu \wedge dx^\nu \end{aligned} \quad (15.69)$$

$$= \frac{1}{2} (\mathring{D}_\mu a_\nu - \mathring{D}_\nu a_\mu) dx^\mu \wedge dx^\nu , \quad (15.70)$$

implicitly summed over both indices μ and ν . The factors of $\frac{1}{2}$ would disappear if the sum were only over distinct antisymmetric pairs $\mu\nu$.

The exterior derivative of the wedge product of a p -form $\mathbf{a}$ with a q -form $\mathbf{b}$ satisfies the same Leibniz-like rule (15.41) as the spacetime curl,

$$d(\mathbf{a} \wedge \mathbf{b}) \equiv (d\mathbf{a}) \wedge \mathbf{b} + (-)^p \mathbf{a} \wedge (d\mathbf{b}) . \quad (15.71)$$

15.9.1 The square of the exterior derivative vanishes

The exterior derivative has the notable property that its square vanishes,

$$\boxed{dd\mathbf{a} = \frac{1}{p!} \partial_{[\nu_1 \nu_2} a_{\mu_1 \dots \mu_p]} dx^{\nu_1} \wedge dx^{\nu_2} \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} = 0} , \quad (15.72)$$

because coordinate derivatives commute. The analogous statement in the geometric algebra is that the torsion-free covariant spacetime curl squared of a multivector $\mathbf{a}$ vanishes, equation (15.43),

$$\overset{\circ}{D} \wedge \overset{\circ}{D} \mathbf{a} = 0 . \quad (15.73)$$

15.10 Hodge dual form

The Hodge dual $^*\mathbf{a}$ of a differential form $\mathbf{a}$ is most easily defined by taking advantage of the isomorphism between the geometric algebra and differential forms. The Hodge dual of a multivector $\mathbf{a}$ is defined to be the multivector $I_N \mathbf{a}$ obtained by premultiplying by the pseudoscalar I_N , equation (13.24).

The pseudoscalar I_N can be expressed as

$$I_N = \varepsilon^M \boldsymbol{\gamma}_M , \quad (15.74)$$

where M runs over the one distinct antisymmetric sequence $1\dots N$ of N indices, and ε^M is the total antisymmetric tensor normalized to $\varepsilon^{1\dots N} = 1$ in an orthonormal frame, as is the convention of this book. Thus the dual $I_N \mathbf{a}$ of a grade- p multivector $\mathbf{a} = a^K \boldsymbol{\gamma}_K$ may be written

$$I_N \mathbf{a} = I_N a^K \boldsymbol{\gamma}_K = \varepsilon^{LK} \boldsymbol{\gamma}_{LK} a^K \boldsymbol{\gamma}_K = (-)^{[p/2]} \varepsilon^{LK} a_K \boldsymbol{\gamma}_L = (-)^{[p/2]} \varepsilon_{LK} a^K \boldsymbol{\gamma}^L , \quad (15.75)$$

implicitly summed over distinct antisymmetric sequences K of p indices, and the one distinct sequence L of $q = N-p$ indices complementary to K . In the third expression of equations (15.75), the indices LK of the pseudoscalar $\boldsymbol{\gamma}_{LK}$ have been ordered without loss of generality to end with the sequence K . The associativity of the multivector product means that $\boldsymbol{\gamma}_{LK} \boldsymbol{\gamma}_K = \boldsymbol{\gamma}_L (\boldsymbol{\gamma}_K \boldsymbol{\gamma}_K)$; the $(-)^{[p/2]}$ factor comes from the square $\boldsymbol{\gamma}_K \boldsymbol{\gamma}_K$ of a grade- p multivector, which in an orthonormal frame is

$$\boldsymbol{\gamma}_K \boldsymbol{\gamma}_K = \boldsymbol{\gamma}_{k_1 \dots k_p} \boldsymbol{\gamma}_{k_1 \dots k_p} = (-)^{[p/2]} \gamma_{k_1 k_1} \dots \gamma_{k_p k_p} , \quad (15.76)$$

with γ_{kl} the orthonormal tetrad metric (the Euclidean metric if all dimensions are spatial, or the Minkowski

metric if one dimension is a time dimension). Equation (15.75) can be cast as a sum over dual multivectors $I_N \gamma_K$,

$$I_N \mathbf{a} = a^K (I_N \gamma_K) , \quad I_N \gamma_K = (-)^{[p/2]} \varepsilon_{LK} \gamma^L = (-)^{[p/2]+pq} \varepsilon_{KL} \gamma^L , \quad (15.77)$$

where L runs over the one distinct antisymmetric sequence complementary to K . The $(-)^{pq}$ factor in the rightmost expression of equations (15.77) comes from permuting the p indices K and q indices L through each other, $\varepsilon_{LK} = (-)^{pq} \varepsilon_{KL}$. Alternatively, equation (15.75) can be cast as a sum over dual coefficients $*a^L$,

$$I_N \mathbf{a} = *a^L \gamma_L , \quad *a^L = (-)^{[p/2]} \varepsilon^{LK} a_K , \quad (15.78)$$

where K runs over the one distinct antisymmetric sequence complementary to L .

The dual $*\mathbf{a}$ of a p -form $\mathbf{a} = a_\Lambda d^p x^\Lambda$ is defined to be the q -form, analogously to the multivector dual (15.75),

$$\boxed{* \mathbf{a} \equiv (-)^{[p/2]} \varepsilon_{\Pi \Lambda} a^\Lambda d^q x^\Pi} , \quad (15.79)$$

implicitly summed over distinct antisymmetric sequences Λ of p indices, and the one distinct sequence Π of $q = N-p$ indices complementary to Λ . As with the multivector expression (15.77), the form dual (15.79) can be cast as a sum over dual volume elements $*d^q x^\Lambda$,

$$*\mathbf{a} = a_\Lambda *d^q x^\Lambda , \quad *d^q x^\Lambda = (-)^{[p/2]} \varepsilon_{\Pi \Lambda} d^q x^\Pi = (-)^{[p/2]+pq} \varepsilon_{\Lambda \Pi} d^q x^\Pi . \quad (15.80)$$

The dual volume element $*d^q x^\Lambda$ is an element of a q -dimensional space, but its index Λ is a totally antisymmetric sequence of $p = N-q$ indices. Alternatively, as with the multivector expression (15.78), the form dual (15.79) can be cast as a sum over dual coefficients, $*a_\Pi$,

$$*\mathbf{a} = *a_\Pi d^q x^\Pi , \quad *a_\Pi = (-)^{[p/2]} \varepsilon_{\Pi \Lambda} a^\Lambda . \quad (15.81)$$

Taking the double dual of a multivector $\mathbf{a}$ multiplies it by the pseudoscalar squared I_N^2 ,

$$**\mathbf{a} = I_N^2 \mathbf{a} = \pm (-)^{[N/2]} \mathbf{a} , \quad (15.82)$$

where the $\pm$ sign is the determinant of the orthonormal tetrad metric (+ for the Euclidean metric, - for the Minkowski metric). The same result (15.82) holds for the double dual $**\mathbf{a}$ of a differential form $\mathbf{a}$. The same factor $\pm (-)^{[N/2]}$ can be deduced in a lengthier fashion by taking the double dual along the lines of equation (15.75). There is a factor of $(-)^{[p/2]}$ from taking the dual of the grade- p vector $\mathbf{a}$, as in equation (15.75); a further factor of $(-)^{[q/2]}$ comes from taking the dual of the grade- q dual vector $*\mathbf{a}$; a factor of $(-)^{pq}$ comes from permuting indices of the pseudoscalar, $\varepsilon^{LK} = (-)^{pq} \varepsilon^{KL}$; and a final $\pm$ sign, the determinant of the tetrad metric, comes from $\varepsilon_M \varepsilon^M = \pm$. The overall sign is, for any p and $q = N-p$,

$$\pm (-)^{[p/2]+[q/2]+pq} = \pm (-)^{[N/2]} . \quad (15.83)$$

The reader may check that equation (15.83) holds for all values of $p+q = N$, with each of p and q either even or odd.

Concept question 15.3 Calculating with the totally antisymmetric tensor. The most difficult thing in all of mathematics is getting the sign right. This is certainly true with the totally antisymmetric tensor. Is there a sure fire way to get the sign right? **Answer.** The key point is that there is an isomorphism between the totally antisymmetric tensor and the geometric algebra. In an orthonormal frame,

$$\varepsilon^{k\dots l} \leftrightarrow \gamma^k \wedge \dots \wedge \gamma^l . \quad (15.84)$$

In a general coordinate frame,

$$\varepsilon^{\kappa\dots\lambda} \leftrightarrow e^\kappa \wedge \dots \wedge e^\lambda . \quad (15.85)$$

Indices are raised and lowered, and transformed between tetrad and coordinate frames, in the usual way, using the tetrad and coordinate metrics, and the vielbein.

15.11 Relation between coordinate- and tetrad-frame volume elements

Consider a p -dimensional hypersurface embedded inside an N -dimensional manifold. Choose an orthonormal tetrad such that the first p basis elements $\gamma_1, \dots, \gamma_p$ of the tetrad are tangent to the p -dimensional hypersurface, while the last $N - p$ basis elements $\gamma_{p+1}, \dots, \gamma_N$ are orthogonal to it. (Such a choice is not always possible. An example is the case of an integral along a null geodesic. But even in that case an integral can be defined — the affine distance — by a suitable limiting procedure. Whatever the case, if an integral can be defined, some version of the equations below applies.) With respect to an orthonormal tetrad frame, the components $d^p x^{1\dots p}$ of the p -volume element transform like the p -dimensional pseudoscalar I_p . Thus the orthonormal tetrad-frame p -volume element is invariant, the proper p -volume element. The coordinate- and tetrad-frame p -volume elements, which are tensors, are related by the vielbein in the usual way, leading to the result that

$$e_{\mu_1\dots\mu_p}^{1\dots p} d^p x^{\mu_1\dots\mu_p} = d^p x^{1\dots p} , \quad (15.86)$$

where $e_{\mu_1\dots\mu_p}^{1\dots p}$ is the determinant of the $p \times p$ vielbein matrix e^m_μ with m running from 1 to p and μ running from μ_1 to μ_p ,

$$e_{\mu_1\dots\mu_p}^{1\dots p} \equiv \begin{vmatrix} e^1_{\mu_1} & \dots & e^1_{\mu_p} \\ \vdots & & \vdots \\ e^p_{\mu_1} & \dots & e^p_{\mu_p} \end{vmatrix} = p! e^1_{[\mu_1} \dots e^p_{\mu_p]} . \quad (15.87)$$

Equation (15.86) is summed over the $\binom{N}{p}$ distinct sets of p coordinate indices $\mu_1\dots\mu_p$ drawn from the N coordinate indices. Equation (15.86) implies that $e_{\mu_1\dots\mu_p}^{1\dots p} d^p x^{\mu_1\dots\mu_p}$ is the proper p -volume element.

Dual proper q -volume elements are related similarly,

$$e_{\mu_1\dots\mu_p}^{1\dots p} {}^*d^q x^{\mu_1\dots\mu_p} = {}^*d^q x^{1\dots p} . \quad (15.88)$$

15.12 Generalized Stokes' theorem

The most important result in the mathematics of differential forms is a generalization of the theorems of Cauchy, Gauss, Green, and Stokes relating the integral of a derivative of a function to a surface integral of the function. Katz (1979) gives a history of the discovery of these theorems. In the mathematicians' compact notation, the generalized **Stokes' theorem** says that the integral of the exterior derivative $d\mathbf{a}$ of a p -form $\mathbf{a}$ over a $(p+1)$ -dimensional hypersurface V equals the integral of the p -form $\mathbf{a}$ over the p -dimensional boundary ∂V of the hypersurface:

$$\int_V d\mathbf{a} = \oint_{\partial V} \mathbf{a} . \quad (15.89)$$

More explicitly, if $\mathbf{a} = a_{\mu_1 \dots \mu_p} dx^{\mu_1} \dots dx^{\mu_p}$ is a p -form, Stokes' theorem states

$$\int_V (p+1) \partial_{[\nu} a_{\mu_1 \dots \mu_p]} dx^{\nu \mu_1 \dots \mu_p} = \oint_{\partial V} a_{\mu_1 \dots \mu_p} dx^{\mu_1 \dots \mu_p} , \quad (15.90)$$

implicitly summed over distinct sequences $\nu \mu_1 \dots \mu_p$ and $\mu_1 \dots \mu_p$ of respectively $p+1$ and p indices. In an arbitrary frame, not necessarily either a coordinate frame or an orthonormal frame, Stoke's theorem (15.90) is

$$\boxed{\int_V (\mathring{D} \wedge \mathbf{a})_L dx^{p+1} x^L = \int_V (p+1) \mathring{D}_{[n} a_{K]} dx^{p+1} x^{nK} = \oint_{\partial V} a_K dx^p x^K} , \quad (15.91)$$

implicitly summed over distinct sequences $L = nK$ and K of respectively $p+1$ and p indices. In a coordinate frame, the torsion-free covariant curl reduces to an ordinary curl, $\mathring{D}_{[\nu} a_{\Lambda]} = \partial_{[\nu} a_{\Lambda]}$, Exercise 2.7.

In the case of a 0-form (scalar) φ , the exterior derivative $d\varphi$, equation (15.68), is the total derivative. The integral of the 1-form $d\varphi$ along any line (1-dimensional hypersurface) $x^\mu(\lambda)$ parametrized by an arbitrary differentiable parameter λ , from initial value λ_0 to final value λ_1 , is

$$\int_{\lambda_0}^{\lambda_1} d\varphi = \int_{\lambda_0}^{\lambda_1} \frac{\partial \varphi}{\partial x^\nu} dx^\nu = \int_{\lambda_0}^{\lambda_1} \frac{\partial \varphi}{\partial x^\nu} \frac{dx^\nu}{d\lambda} d\lambda = \int_{\lambda_0}^{\lambda_1} \frac{d\varphi}{d\lambda} d\lambda = \varphi(\lambda_1) - \varphi(\lambda_0) . \quad (15.92)$$

Equation (15.92) can be recognized as the fundamental theorem of calculus. Equation (15.92) is equation (15.89) or (15.91) for the case where $\mathbf{a}$ is the 0-form (scalar) φ . The hypersurface V is the 1-dimensional path of integration. The boundary ∂V is the two endpoints of the path.

Here is a sketch of a proof of the generalized Stokes' theorem (15.89). The key ingredient is that $d\mathbf{a}$ is coordinate-invariant, so one can use any convenient coordinate system to evaluate the integral, and the result will be independent of the choice of coordinates.

First, partition the hypersurface V into rectangular patches. Rectangular means that a system of coordinates can be chosen such that the patch extends over a fixed finite interval $x_0^\mu \leq x^\mu \leq x_1^\mu$ of each coordinate. Figure 15.1 illustrates a partition of a 2-dimensional disk into five rectangular patches. Thanks to the arbitrariness of the choice of coordinates, although each patch appears to be non-rectangular, coordinates can always be chosen so that the patch is rectangular with respect to those coordinates. Notice that the $(p+1)$ -dimensional hypersurface could be embedded in a higher dimensional manifold, so there could potentially be more coordinates available than the dimension of the hypersurface; but again the arbitrariness of

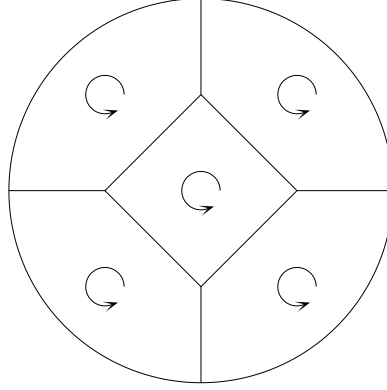


Figure 15.1 Partition of a disk into five rectangular patches. The arrowed circles show the direction of circulation of the integral over the boundary of each patch.

coordinates means that coordinates can always be chosen such that only $p+1$ of them vary over the $(p+1)$ -dimensional hypersurface, the remaining coordinates being constant. With such convenient coordinates, the integral over a patch is a straightforward integration in Euclidean space. For simplicity, consider the integral in 2 dimensions. The integral over a single rectangular patch $x_0 \leq x \leq x_1$ and $y_0 \leq y \leq y_1$ is

$$\begin{aligned}
 \int_{\text{patch}} d\mathbf{a} &= \int_{y_0}^{y_1} \int_{x_0}^{x_1} \left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y} \right) dx \wedge dy \\
 &= \int_{y_0}^{y_1} \left(\int_{x_0}^{x_1} \frac{\partial a_y}{\partial x} dx \right) \wedge dy - \int_{x_0}^{x_1} \left(\int_{y_0}^{y_1} \frac{\partial a_x}{\partial y} dy \right) \wedge dx \\
 &= \int_{y_0}^{y_1} \left(\int_{x_0}^{x_1} \frac{\partial a_y}{\partial x} dx \right) dy - \int_{x_0}^{x_1} \left(\int_{y_0}^{y_1} \frac{\partial a_x}{\partial y} dy \right) dx \\
 &= \int_{y_0}^{y_1} [a_y(x_1) - a_y(x_0)] dy - \int_{x_0}^{x_1} [a_x(y_1) - a_x(y_0)] dx \\
 &= \oint_{\partial \text{patch}} a_\mu dx^\mu = \oint_{\partial \text{patch}} \mathbf{a} . \tag{15.93}
 \end{aligned}$$

The first line of equations (15.93) is the standard expression (15.70) for the exterior derivative of a 1-form $\mathbf{a}$; the double count over pairs of indices eliminates the factor of $\frac{1}{2}$. The second line of equations (15.93) rearranges the first. The third line of equations (15.93) differs from the second by the loss of the $\wedge$ signs; the equality holds because $\int (\partial a_x / \partial y) dy$ is a scalar for any interval of integration, and the wedge product of a scalar with a differential form is just the ordinary product of the scalar with the form, equation (15.62). The fourth line of equations (15.93) follows from the fundamental theorem of calculus, equation (15.92). The integral contains 4 contributions, corresponding to the 4 edges of the rectangular patch. The signs of the 4 contributions are such that they circulate anti-clockwise about the patch, as illustrated in Figure 15.1.

The last line of equations (15.93) expresses the fourth in more compact notation, with ∂patch denoting the boundary, the 4 edges, of the patch. Equations (15.93) prove Stokes' theorem for a patch.

The final step of the proof is to add together the contributions from all the patches of the partition. Where two patches abut, the contributions from the common edge cancel, because consistent circulation about the boundaries causes the integral along the common edge to be in opposite directions, as illustrated in Figure 15.1. Once again, the coordinate-invariant character of the differential form $\mathbf{a}$ ensures that the integral along a prescribed path is independent of the choice of coordinates, so the contributions from abutting edges of patches do indeed cancel.

15.13 Exact and closed forms

Consider the 1-form $d\phi$ defined by the exterior derivative of the azimuthal angle ϕ around a circle. The integral of the angle around the circle is

$$\int_0^{2\pi} d\phi = 2\pi . \quad (15.94)$$

But since the circle has no boundary, should not Stokes' theorem imply that the integral vanishes? The resolution of the problem is that ϕ is not a single-valued scalar. The 1-form $d\phi$ constructed from ϕ is well-defined, being single-valued and continuous everywhere on the circle, but ϕ itself is not. The circle can be cut at any point, and a single-valued scalar ϕ defined on the cut circle. But since the scalar is discontinuous at the cut point, the contributions on the boundary do not cancel, but rather produce a finite contribution, namely 2π .

A differential form $\mathbf{F}$ is said to be **exact** if it can be expressed as the exterior derivative of a differential form $\mathbf{A}$,

$$\mathbf{F} = d\mathbf{A} . \quad (15.95)$$

Stokes' theorem implies that an integral of an exact form over a surface with no boundary must vanish. The condition of exactness is a global condition. The above example 1-form $d\phi$ in equation (15.94) is not exact, because ϕ is not a single-valued 0-form (scalar).

A differential form $\mathbf{F}$ is said to be **closed** if its exterior derivative vanishes,

$$d\mathbf{F} = 0 . \quad (15.96)$$

The rule $dd = 0$, §15.9.1, implies that every exact form is closed. The inverse theorem, that every closed form is exact, is true locally, but not globally. **Poincaré's lemma** states that a form that is closed over a volume V that is continuously contractible to point is exact over that volume. The condition of being closed can be thought of as a local test of exactness. The example form $d\phi$ is closed, but not exact. In the Cartesian xy plane, the 1-form $d\phi$ is

$$d\phi = d\text{atan}(y/x) = \frac{x\,dy - y\,dx}{x^2 + y^2} , \quad (15.97)$$

which is singular at the origin $x = y = 0$. Consistent with Poincaré's lemma, the 1-form $d\phi$ is not continuously contractible to a point.

The above example illustrates that topological properties of differentiable manifolds, such as winding number, can be inferred from the behaviour of integrals.

15.14 Generalized Gauss' theorem

In physics, Stokes' theorem (15.91) is most commonly encountered in the form of Gauss' theorem, which relates the volume integral of the divergence of a vector to the integral of the flux of the vector through the surface of the volume. The relation (15.40) shows that a covariant curl as required by Stokes' theorem (15.91) can be converted to a covariant divergence by post-multiplying by the pseudoscalar I_N ,

$$\mathring{D} \wedge (\mathbf{a} I_N) = (\mathring{D} \cdot \mathbf{a}) I_N . \quad (15.98)$$

If $\mathbf{a} = a_K d^p x^K$ is a p -form, substituting equation (15.98) into Stokes' theorem (15.91) gives the generalized Gauss' theorem in an arbitrary (not necessarily coordinate or orthonormal) frame, with $q \equiv N-p$,

$$\boxed{\int_V (\mathring{D} \cdot \mathbf{a})_L {}^*d^{q+1}x^L = \int_V \mathring{D}^n a_{nL} {}^*d^{q+1}x^L = \oint_{\partial V} a_K {}^*d^q x^K} , \quad (15.99)$$

where $(\mathring{D} \cdot \mathbf{a})_L$ denotes the components of the torsion-free covariant divergence, equation (15.38a), ${}^*d^q x^K$ denotes the dual q -volume element, equation (15.80), and K and L are implicitly summed over distinct antisymmetric sequences of p and $p-1$ indices respectively.

In the mathematicians' notation, Gauss' theorem (15.99) is

$$(-)^{(p+1)(q-1)} \int_V {}^*(d\mathbf{a}) = (-)^{pq} \oint_{\partial V} {}^*\mathbf{a} , \quad (15.100)$$

the signs coming from commuting the pseudoscalar I_N through $d\mathbf{a}$ on the left hand side and through $\mathbf{a}$ on the right hand side. Equivalently,

$$(-)^{N-1} \int_V {}^*(d\mathbf{a}) = \int_V d({}^*\mathbf{a}) = \oint_{\partial V} {}^*\mathbf{a} , \quad (15.101)$$

the $(-)^{N-1}$ sign coming from commuting the pseudoscalar through the 1-form exterior derivative d .

In the remainder of this book, the dual q -volume element ${}^*d^q x^{k\dots l}$ is often abbreviated to $d^q x^{k\dots l}$ without the Hodge star symbol, since the dual nature is usually evident from the number of indices $k\dots l$, which is q for the standard q -volume, or $p \equiv N-q$ for the dual q -volume. The only ambiguity occurs when $q = p = N/2$. For example, the dual N -volume element, which is a scalar, will be abbreviated to $d^N x$, whereas the standard N -volume element, which is a pseudoscalar, is written $d^N x^{1\dots N}$.

Beware that physics texts commonly use $d^N x$ to denote the pseudoscalar N -volume, and $e d^N x$ or equivalently $\sqrt{-g} d^N x$ to denote the dual scalar N -volume. The common physics convention seems designed to

confuse the smart student who expects a notation that manifests, not obscures, the transformational properties of a volume element.

The simplest and most common application of Gauss' theorem is where $\mathbf{a} = a_m dx^m$ is a 1-form, in which case

$$\boxed{\int_V \tilde{D}^n a_n d^N x = \oint_{\partial V} a_n d^{N-1} x^n}, \quad (15.102)$$

where, as just remarked, $d^N x$ and $d^{N-1} x^n$ denote respectively the dual scalar N -volume and the dual vector $(N-1)$ -volume.

15.15 Dirac delta-function

A Dirac delta-function can be thought of as a special function that is infinity at the origin, zero everywhere else, and has unit volume in the sense that it yields one when integrated over any region containing the origin. In curved spacetime, in order that the integral be a scalar, the p -dimensional Dirac delta-function must transform oppositely to the p -dimensional volume element.

The p -dimensional Dirac delta-function $\delta^p(x)$ is defined such that for any scalar function $f(x)$, the integral over any p -volume element containing the origin $x = 0$,

$$\int f(x) \delta_K^p(x) d^p x^K = f(0), \quad (15.103)$$

yields the value $f(0)$ of the function at the origin. The p -dimensional Dirac delta-function is an antisymmetric tensor of rank p , with components $\delta_K^p(x)$, where K runs over distinct antisymmetric sequences of K indices.

The dual q -dimensional Dirac delta-function ${}^*\delta^q(x)$ with $q \equiv N-p$, is defined to behave similarly when integrated over the dual q -volume element ${}^*d^q x^K$ defined by equation (15.80),

$$\int f(x) {}^*\delta_K^q(x) {}^*d^q x^K = f(0). \quad (15.104)$$

The dual q -dimensional Dirac delta-function ${}^*\delta^q(x)$ is an antisymmetric tensor of rank p , with components ${}^*\delta_K^q(x)$ where K runs over distinct antisymmetric sequences of p indices.

As with the dual q -volume, the dual Dirac delta-function ${}^*\delta_{k\dots l}^q(x)$ will often be abbreviated in this book to $\delta_{k\dots l}^q(x)$ without the Hodge star symbol, since the dual nature can usually be inferred from the number p of indices $k\dots l$.

The most common use of the Dirac delta-function is in integration over N -dimensional space,

$$\boxed{\int f(x) \delta^N(x) d^N x = f(0)}, \quad (15.105)$$

where $\delta^N(x)$ and $d^N x$ denote respectively the dual scalar N -dimensional Dirac delta-function, and the dual scalar N -volume. The lack of indices on $\delta^N(x)$ and $d^N x$ signals that they are scalars.

15.16 Integration of multivector-valued forms

In Chapter 16, §16.14, it will be found that the Hilbert action of general relativity takes its most insightful form when expressed in the language of multivector-valued forms. These are forms whose coefficients are themselves multivectors,

$$\mathbf{a} = \mathbf{a}_\Lambda d^p x^\Lambda = a_{K\Lambda} \gamma^K d^p x^\Lambda, \quad (15.106)$$

implicitly summed over distinct sequences K of multivector indices and distinct sequences Λ of p coordinate indices. The advantage of the multivector-valued forms notation is that it makes manifest the two distinct symmetries of general relativity: Lorentz transformations, encoded in the transformation of the multivector, and translations (coordinate transformations), encoded in the transformation of the form.

Stokes' theorem for multivector-valued forms is an immediate generalization of Stokes's theorem (15.89) for forms: the integral of the exterior derivative $d\mathbf{a}$ of a p -form multivector $\mathbf{a}$, equation (15.106), over a $(p+1)$ -dimensional hypersurface V equals the integral of the p -form multivector $\mathbf{a}$ over the p -dimensional boundary ∂V of the hypersurface:

$$\int_V d\mathbf{a} = \oint_{\partial V} \mathbf{a}. \quad (15.107)$$

In other words, the fact that the coordinate components $\mathbf{a}_\Lambda$ of the form $\mathbf{a}$ are themselves multivectors leaves Stokes' theorem intact and unchanged.

Exercise 15.4 Action principle for strings and branes in arbitrary dimensions. The action for a point particle is, up to a factor, the integral of the proper time along the worldline of the particle, equation (4.7). Similarly, a consistent action for a p -dimensional object in N -dimensional spacetime is, up to a factor, the integral of the proper area of the $(p+1)$ -dimensional worldsheet of the object. String theorists call such an object a p -brane, with $p = 0$ for a point particle, $p = 1$ for a string, and $p = -1$ for an instanton, which has no extension in space or time. Let λ^α , $\alpha = 0, \dots, p$, be $p+1$ coordinates on the $(p+1)$ -dimensional worldsheet of the brane. The action of the p -brane is the **Nambu-Goto** action

$$S_p = -\mu \int_i^f d^{p+1} \lambda = -\mu \int_i^f e d^{p+1} \lambda^{0\dots p}, \quad (15.108)$$

where $d^{p+1} \lambda$ is the dual scalar $(p+1)$ -volume element, $d^{p+1} \lambda^{0\dots p}$ is the pseudoscalar $(p+1)$ -volume element, and $e \equiv e_{0\dots p}^{0\dots p}$ is the vielbein determinant, equation (15.87). The action has units of mass $\times$ length (angular momentum), so the constant μ , the tension of the brane, has units of mass/length ^{p} . For example, for a string, $p = 1$, the tension μ has dimensions of mass per unit length. Notice that it is built into the action (15.108) that the tension μ of the brane, its mass per unit proper length ^{p} , is constant. Thus the brane behaves like a thin shell with a vacuum internal equation of state, equation (15.142). The minus sign in the brane action (15.108) arises for the same reason as the minus sign in the action (4.7) for a particle: when one dimension is timelike, the principle of least spatial area is replaced by the principle of most spacetime area. A positive μ implies a positive proper mass/length ^{p} of the brane.

1. Derive the equations of motion that follow from the action (15.108).

2. As previously for a point particle, §4.3, the standard version of the brane Lagrangian, equation (15.113), involves a square root, and is not in (super-)Hamiltonian form. Recast the action (15.108) into (super-)Hamiltonian form.
3. Derive the energy-momentum tensor of the brane.

Solution.

1. **Standard Lagrangian.** The Lagrangian of the p -brane with action (15.108) is $L_p = -\mu e$. The Lagrangian approach requires that the Lagrangian be expressed in terms of coordinates and velocities. Let x^μ be N coordinates of the N -dimensional spacetime in which the brane propagates, with line-element $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$. The coordinates x^μ on the worldsheet are functions $x^\mu(\lambda^\alpha)$ of the worldsheet coordinates λ^α . The induced metric $h_{\alpha\beta}$ on the $(p+1)$ -dimensional worldsheet of the brane, satisfying $ds^2 = h_{\alpha\beta} d\lambda^\alpha d\lambda^\beta$, is related to the metric $g_{\mu\nu}$ of the full spacetime by

$$h_{\alpha\beta} = g_{\mu\nu} \frac{\partial x^\mu}{\partial \lambda^\alpha} \frac{\partial x^\nu}{\partial \lambda^\beta} = g_{\mu\nu} u_\alpha^\mu u_\beta^\nu, \quad (15.109)$$

where the velocities u_α^μ are defined by

$$u_\alpha^\mu \equiv \frac{\partial x^\mu}{\partial \lambda^\alpha}, \quad \mu = 0, \dots, N-1, \quad \alpha = 0, \dots, p. \quad (15.110)$$

In terms of an orthonormal tetrad whose first $(p+1)$ vectors γ_a are tangent to the worldsheet, the metric $h_{\alpha\beta}$ of the worldsheet is

$$h_{\alpha\beta} \equiv \eta_{ab} e^a_\alpha e^b_\beta. \quad (15.111)$$

Coordinate indices $\mu, \nu, \dots$ are raised and lowered with the spacetime metric $g_{\mu\nu}$, while worldsheet indices $\alpha, \beta, \dots$ are raised and lowered with the independent worldsheet metric $h_{\alpha\beta}$. The determinant h of the metric on the worldsheet is

$$h \equiv |h_{\alpha\beta}| = |\eta_{ab}| |e^a_\alpha| |e^b_\beta| = -e^2, \quad (15.112)$$

where e is the vielbein determinant, the same determinant as that in the action (15.108). The minus sign in equation (15.112) assumes that the worldsheet progresses in time, so that one of the dimensions of the worldsheet is timelike, hence $|\eta_{ab}| = -1$. The Lagrangian of the p -brane is then

$$L_p = -\mu e = -\mu \sqrt{-h} = -\mu \sqrt{-|h_{\alpha\beta}|} = -\mu \sqrt{-\left| g_{\mu\nu} \frac{\partial x^\mu}{\partial \lambda^\alpha} \frac{\partial x^\nu}{\partial \lambda^\beta} \right|} = -\mu \sqrt{-|g_{\mu\nu} u_\alpha^\mu u_\beta^\nu|}. \quad (15.113)$$

The variation of the Lagrangian satisfies, from equation (2.77) for the variation of a determinant,

$$e^{-1} \delta L_p = -\mu \delta \ln e = -\frac{1}{2} \mu \delta \ln h = -\frac{1}{2} \mu h^{\alpha\beta} \delta h_{\alpha\beta} = -\frac{1}{2} \mu h^{\alpha\beta} \delta (g_{\mu\nu} u_\alpha^\mu u_\beta^\nu). \quad (15.114)$$

The variations of the Lagrangian with respect to the velocities u_α^κ and coordinates x^κ are therefore

$$p^\alpha_\kappa \equiv \frac{\partial L_p}{\partial u_\alpha^\kappa} = -\mu e h^{\alpha\beta} g_{\kappa\nu} \frac{\partial x^\nu}{\partial \lambda^\beta} = -\mu e u_\kappa^\alpha, \quad (15.115a)$$

$$\frac{\partial L_p}{\partial x^\kappa} = -\frac{1}{2} \mu e h^{\alpha\beta} \frac{\partial g_{\mu\nu}}{\partial x^\kappa} \frac{\partial x^\mu}{\partial \lambda^\alpha} \frac{\partial x^\nu}{\partial \lambda^\beta} = -\mu e \Gamma_{\mu\nu\kappa} u_\alpha^{\alpha\mu} u_\alpha^{\nu}, \quad (15.115b)$$

the variation (15.115a) with respect to velocities u_{α}^{κ} defining the generalized momenta p^{α}_{κ} . In equation (15.115b) the coordinate derivatives of the metric have been replaced by torsion-free coordinate connections (Christoffel symbols) in view of $\partial g_{\mu\nu}/\partial x^{\kappa} = 2\Gamma_{(\mu\nu)\kappa} = 2\overset{\circ}{\Gamma}_{(\mu\nu)\kappa}$, equations (2.59) and (2.64). Linearity of the derivative, equation (4.3), implies that the variation of the velocity equals the velocity of the variation, $\delta(\partial x^{\kappa}/\partial \lambda^{\alpha}) = \partial \delta x^{\kappa}/\partial \lambda^{\alpha}$. The variation of the brane action with respect to coordinates and velocities is then

$$\delta S_p = \int_i^f \left(p^{\alpha}_{\kappa} \frac{\partial \delta x^{\kappa}}{\partial \lambda^{\alpha}} + \frac{\partial L_p}{\partial x^{\kappa}} \delta x^{\kappa} \right) d^p \lambda^{0\dots p} . \quad (15.116)$$

The first term on the right hand side of equation (15.116) can be integrated by parts. To do so, recall that Gauss' theorem (15.99) involves the integral of a torsion-free divergence, which in the present application takes the form

$$\int \overset{\circ}{D}_{\alpha} a^{\alpha} d^{p+1} \lambda = \oint a^{\alpha} d^p \lambda_{\alpha} , \quad (15.117)$$

with $d^{p+1} \lambda = e d^{p+1} \lambda^{0\dots p}$ and $d^p \lambda_{\alpha} = e d^p \lambda^{0\dots \kappa \dots p}$ the dual proper $p+1$ and p volume elements. The integration by parts is accomplished through

$$p^{\alpha}_{\kappa} \frac{\partial \delta x^{\kappa}}{\partial \lambda^{\alpha}} = \frac{\partial}{\partial \lambda^{\alpha}} (p^{\alpha}_{\kappa} \delta x^{\kappa}) - \frac{\partial p^{\alpha}_{\kappa}}{\partial \lambda^{\alpha}} \delta x^{\kappa} = e \overset{\circ}{D}_{\alpha} (e^{-1} p^{\alpha}_{\kappa} \delta x^{\kappa}) - \frac{\partial p^{\alpha}_{\kappa}}{\partial \lambda^{\alpha}} \delta x^{\kappa} . \quad (15.118)$$

The variation of the action (15.116) becomes, after integration by parts,

$$\delta S_p = \oint_i^f p^{\alpha}_{\kappa} \delta x^{\kappa} d^p \lambda^{0\dots \kappa \dots p} - \int_i^f \left(\frac{\partial p^{\alpha}_{\kappa}}{\partial \lambda^{\alpha}} - \frac{\partial L_p}{\partial x^{\kappa}} \right) \delta x^{\kappa} d^{p+1} \lambda^{0\dots p} . \quad (15.119)$$

As usual, application of least action requires the coordinates to be held fixed on the boundary, so δx^{κ} vanishes on the boundary, and the surface term in equation (15.119) vanishes. Requiring the variation of the action to vanish for all possible variations of δx^{κ} on the worldsheet then implies the equation of motion

$$\frac{\partial p^{\alpha}_{\kappa}}{\partial \lambda^{\alpha}} = \frac{\partial L_p}{\partial x^{\kappa}} = -\mu e \overset{\circ}{\Gamma}_{\mu\nu\kappa} u^{\alpha\mu} u_{\alpha}^{\nu} . \quad (15.120)$$

The equation of motion (15.120) may also be written as the vanishing of the torsion-free covariant divergence of the velocity $u^{\alpha}_{\kappa} = -p^{\alpha}_{\kappa}/(\mu e)$,

$$\overset{\circ}{D}_{\alpha} u^{\alpha}_{\kappa} = \partial_{\alpha} u^{\alpha}_{\kappa} + \overset{\circ}{\Gamma}_{\beta\alpha}^{\alpha} u^{\beta}_{\kappa} - \overset{\circ}{\Gamma}_{\mu\nu\kappa} u^{\alpha\mu} u_{\alpha}^{\nu} = 0 , \quad (15.121)$$

in which the connection term $\overset{\circ}{\Gamma}_{\beta\alpha}^{\alpha} = \partial \ln e / \partial \lambda^{\beta}$ (all worldsheet indices), equation (2.79), enforces covariance with respect to the worldsheet coordinate index α of the velocity, while the connection term $\overset{\circ}{\Gamma}_{\mu\nu\kappa}$ (all external indices) enforces covariance with respect to the external coordinate index κ of the velocity.

As a check, for a point particle, $p = 0$, of mass $\mu = m$, the Lagrangian (15.113) is $L_m = -me$ with $e = d\tau/d\lambda$, equation (4.8), the induced metric (15.109) is $h_{00} = -e^2$ with inverse $h^{00} = -e^{-2}$, the

momentum (15.115a) is $p_{\kappa} = me^{-1}g_{\kappa\nu}dx^{\nu}/d\lambda = mg_{\kappa\nu}dx^{\nu}/d\tau$, and the equation of motion (15.120) reduces to

$$\frac{dp_{\kappa}}{d\lambda} = m\mathring{\Gamma}_{\mu\nu\kappa} \frac{dx^{\mu}}{d\lambda} \frac{dx^{\nu}}{d\lambda} \frac{d\lambda}{d\tau}, \quad (15.122)$$

in agreement with equation (4.12).

2. **(Super-)Hamiltonian.** The standard brane Lagrangian (15.113) is not in (super-)Hamiltonian form. To take Hamiltonian form, the brane action must take the form, analogous to the action (4.69) for a point particle,

$$S_p = \int_i^f \left(p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} - H_p \right) d^{p+1}\lambda^{0\dots p}, \quad (15.123)$$

with brane Lagrangian

$$L_p = p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} - H_p. \quad (15.124)$$

The brane Hamiltonian is to be considered as a function $H_p(x^{\mu}, p^{\alpha}_{\mu})$ of independent coordinates x^{μ} and momenta p^{α}_{μ} . However, as in the case of a point particle, §4.6, the equations of motion must be independent of the arbitrary coordinates λ^{α} that label the worldsheet: the Hamiltonian must be reparametrization independent. To achieve independence with respect to the choice of worldsheet coordinates, it is necessary to treat the brane Hamiltonian as a function $H_p(x^{\mu}, p^{\alpha}_{\mu}, h_{\alpha\beta})$ not only of coordinates and momenta, but also of an independent worldsheet metric $h_{\alpha\beta}$. In the Hamiltonian approach one cannot simply assert a priori that the worldsheet metric $h_{\alpha\beta}$ satisfies the condition (15.109) that it be induced from the spacetime metric $g_{\mu\nu}$; rather, that condition (15.109) should emerge as an equation of motion (15.128c).

The variation of the first term in the integrand of the action (15.123) is

$$\begin{aligned} \delta \left(p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} \right) &= \delta p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} + p^{\alpha}_{\mu} \frac{\partial \delta x^{\mu}}{\partial \lambda^{\alpha}} = \delta p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} + \frac{\partial}{\partial \lambda^{\alpha}} (p^{\alpha}_{\mu} \delta x^{\mu}) - \frac{\partial p^{\alpha}_{\mu}}{\partial \lambda^{\alpha}} \delta x^{\mu} \\ &= \delta p^{\alpha}_{\mu} \frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} + e\mathring{D}_{\alpha} (e^{-1} p^{\alpha}_{\mu} \delta x^{\mu}) - \frac{\partial p^{\alpha}_{\mu}}{\partial \lambda^{\alpha}} \delta x^{\mu}. \end{aligned} \quad (15.125)$$

The term involving the torsion-free divergence integrates by parts to a surface term. The variation of the action (15.123) with Hamiltonian $H_p(x^{\mu}, p^{\alpha}_{\mu}, h_{\alpha\beta})$ is

$$\begin{aligned} \delta S_p &= \oint_i^f p^{\alpha}_{\mu} \delta x^{\mu} d^p \lambda^{0\dots\alpha\dots p} \\ &+ \int_i^f \left(\delta p^{\alpha}_{\mu} \left(\frac{\partial x^{\mu}}{\partial \lambda^{\alpha}} - \frac{\partial H_p}{\partial p^{\alpha}_{\mu}} \right) - \left(\frac{\partial p^{\alpha}_{\mu}}{\partial \lambda^{\alpha}} + \frac{\partial H_p}{\partial x^{\mu}} \right) \delta x^{\mu} - \frac{\partial H_p}{\partial h_{\alpha\beta}} \delta h_{\alpha\beta} \right) d^{p+1}\lambda^{0\dots p}. \end{aligned} \quad (15.126)$$

Vanishing of the surface term requires either $p^{\alpha}_{\mu} = 0$ or $\delta x^{\mu} = 0$ on the boundary, which are called Neumann and Dirichlet boundary conditions. Dirichlet boundary conditions $\delta x^{\mu} = 0$ are standard on the spacelike caps at the initial and final times τ_i and τ_f ($\alpha = 0$). For branes with spatial boundaries, such as open strings, boundary conditions must also be chosen on the spatial boundary ($\alpha \neq 0$), and either

Neumann or Dirichlet or a mix of boundary conditions may be appropriate. The (super-)Hamiltonian that correctly recovers the relation (15.115a) between brane velocities and momenta, and the brane equations of motion (15.120), is

$$H_p = -\frac{1}{2\mu e} h_{\alpha\beta} g^{\mu\nu} p^\alpha_\mu p^\beta_\nu - \frac{(p-1)\mu e}{2} , \quad (15.127)$$

in which e is to be considered a function $e = \sqrt{-h} = \sqrt{-|h_{\alpha\beta}|}$ of the worldsheet metric $h_{\alpha\beta}$.

The equations of motion that follow from the vanishing of the variation (15.126) of the brane action with the brane Hamiltonian (15.127) are

$$u_\alpha^\mu \equiv \frac{\partial x^\mu}{\partial \lambda_\alpha} = \frac{\partial H_p}{\partial p^\alpha_\mu} = -\frac{1}{\mu e} h_{\alpha\beta} g^{\mu\nu} p^\beta_\nu , \quad (15.128a)$$

$$\frac{\partial p^\alpha_\kappa}{\partial \lambda^\alpha} = -\frac{\partial H_p}{\partial x^\kappa} = \frac{1}{2\mu e} h_{\alpha\beta} \frac{\partial g^{\mu\nu}}{\partial x^\kappa} p^\alpha_\mu p^\beta_\nu , \quad (15.128b)$$

$$0 = \frac{\partial H_p}{\partial h_{\alpha\beta}} = -\frac{1}{2\mu e} \left(g^{\mu\nu} p^\alpha_\mu p^\beta_\nu - \frac{1}{2} h^{\alpha\beta} (p^\gamma_\mu p_\gamma^\mu - (p-1)(\mu e)^2) \right) . \quad (15.128c)$$

Notice that the equations of motion (15.128a) and (15.128b) for the coordinates x^μ and momenta p^α_μ depend on the worldsheet metric $h_{\alpha\beta}$ only in the combination $h_{\alpha\beta}/e$, which in the special case of strings is independent of the normalization of $h_{\alpha\beta}$. In other words the equations of motion are conformally invariant in the special case of strings. Equation (15.128c) imposes that $h^{\alpha\beta}$ be proportional to $g^{\mu\nu} p^\alpha_\mu p^\beta_\nu$. Taking the trace of equation (15.128c), and bearing in mind that $h_{\alpha\beta} h^{\alpha\beta} = \delta^\alpha_\alpha = p+1$ (the brane dimension), implies the normalization condition

$$0 = \frac{\partial H_p}{\partial \ln |h|} = h_{\alpha\beta} \frac{\partial H_p}{\partial h_{\alpha\beta}} = \frac{p-1}{4\mu e} (p^\gamma_\mu p_\gamma^\mu - (p+1)(\mu e)^2) , \quad (15.129)$$

which, along with the relation (15.128a) between u_α^μ and p^α_μ , implies the desired condition (15.109) that the worldsheet metric $h_{\alpha\beta}$ is indeed induced by the spacetime metric $g_{\mu\nu}$. For all branes *except* strings, equation (15.129) implies the normalization

$$g^{\mu\nu} p^\alpha_\mu p^\beta_\nu = (\mu e)^2 h^{\alpha\beta} \quad \text{for } p \neq 1 . \quad (15.130)$$

The reason there is no normalization condition for strings is that, under a conformal rescaling of the worldsheet metric $h_{\alpha\beta}$ by a scale factor a ,

$$h_{\alpha\beta} \rightarrow a^2 h_{\alpha\beta} , \quad h \equiv |h_{\alpha\beta}| \rightarrow a^{2(p+1)} h , \quad e \equiv \sqrt{-h} \rightarrow a^{p+1} e , \quad (15.131)$$

the brane Hamiltonian (15.127) transforms as

$$H_p \rightarrow -\frac{a^{1-p}}{2\mu e} h_{\alpha\beta} g^{\mu\nu} p^\alpha_\mu p^\beta_\nu - \frac{a^{p+1}(p-1)\mu e}{2} , \quad (15.132)$$

which implies that the Hamiltonian (15.127) is conformally invariant for a string,

$$H_p \rightarrow H_p \quad \text{for } p = 1 . \quad (15.133)$$

The conformal invariance (15.133), commonly called Weyl invariance, of the string Hamiltonian implies that equation (15.129) is satisfied automatically without any normalization condition on $h_{\alpha\beta}$. The conformal invariance of the string Hamiltonian is at the heart of some of the magic of string theory.

For non-strings, $p \neq 1$, after the normalization (15.130) is imposed, the value of the brane Hamiltonian (15.127) is

$$H_p = -\mu e \quad \text{for } p \neq 1 . \quad (15.134)$$

The value of the brane Lagrangian (15.124) is

$$L_p = p^\alpha_\mu u_\alpha^\mu - H_p = -\mu e \quad \text{for } p \neq 1 , \quad (15.135)$$

in agreement with the standard brane Lagrangian (15.113).

For strings, $p = 1$, let the true worldsheet metric be denoted with a happy sign, $\check{h}_{\alpha\beta}$,

$$\check{h}_{\alpha\beta} \equiv g_{\mu\nu} u_\alpha^\mu u_\beta^\nu . \quad (15.136)$$

For strings, equation (15.128c) implies that $h_{\alpha\beta}$ is proportional to the true worldsheet metric $\check{h}_{\alpha\beta}$, but leaves the normalization arbitrary. For strings, the Hamiltonian (15.127), with momenta p^α_μ eliminated in favour of velocities u_α^μ using equation (15.128a), is

$$H_p = -\frac{1}{2}\mu e h^{\alpha\beta} \check{h}_{\alpha\beta} \quad \text{for } p = 1 . \quad (15.137)$$

But for strings, the product $eh^{\alpha\beta}$ with $e = \sqrt{-h}$ is unchanged by the normalization of $h_{\alpha\beta}$, so can be replaced by $\check{e}\check{h}^{\alpha\beta}$ with $\check{e} = \sqrt{-\check{h}}$. Thus the string Hamiltonian (15.137) is, regardless of the normalization of $h_{\alpha\beta}$,

$$H_p = -\mu \check{e} \quad \text{for } p = 1 , \quad (15.138)$$

which agrees with the non-string Hamiltonian (15.134) if e is interpreted as the true vielbein determinant $\check{e}$. The value of the string Lagrangian (15.124) is

$$L_p = -\mu \check{e} \quad \text{for } p = 1 , \quad (15.139)$$

again in agreement with the standard brane Lagrangian (15.113).

Equation (15.128b) with (15.128a) implies

$$\frac{\partial p^\alpha_\kappa}{\partial \lambda^\alpha} = -\mu e h^{\alpha\beta} \overset{\circ}{\Gamma}_{\mu\nu\kappa} u_\alpha^\mu u_\beta^\nu , \quad (15.140)$$

which recovers the brane equation of motion (15.120), for branes of arbitrary dimension p . For strings, $p = 2$, the factor $eh^{\alpha\beta}$ in the equation of motion (15.140) can be replaced by $\check{e}\check{h}^{\alpha\beta}$, affirming that equation (15.140) is correctly normalized also for $p = 1$.

For a point particle, $p = 0$, of mass $\mu = m$, the brane Hamiltonian (15.127) reduces to the nice Hamiltonian (4.96) (absent electromagnetism) if the scale factor a in the latter is identified with

$$a = \frac{e}{m} , \quad (15.141)$$

where $e \equiv \sqrt{-|h_{00}|}$.

For a string, $p = 1$, the brane Lagrangian (15.124) with Hamiltonian (15.127) is essentially the Polyakov (1981) Lagrangian.

3. The energy-momentum tensor $T_{\alpha a}$ of a brane is obtained by varying the brane action S_p with respect to the worldsheet vielbein $e^{a\alpha}$, equation (16.120). The energy-momentum tensor of the brane is proportional to the worldsheet vielbein, so the brane has a vacuum equation of state,

$$T_{\alpha a} = -\mu e_{a\alpha} . \quad (15.142)$$

In the coordinate frame, the brane energy-momentum is

$$T_{\alpha\beta} = -\mu h_{\alpha\beta} . \quad (15.143)$$

The brane Lagrangian is independent of the Lorentz connections $\Gamma_{ab\alpha}$, so the brane carries no torsion, equation (16.121).

15.17 Covariant derivative of a spinor

A Lorentz transformation of a Dirac spinor ψ by rotor R transforms the spinor by $\psi \rightarrow R\psi$. An infinitesimal Lorentz transformation $R = 1 + \epsilon\mathbf{\Gamma}/2$ generated by a bivector $\mathbf{\Gamma}$ transforms $\psi \rightarrow \psi + \frac{1}{2}\epsilon\mathbf{\Gamma}\psi$. Consequently the action of the connection operator $\hat{\Gamma}_n$ on a spinor ψ is

$$\hat{\Gamma}_n\psi = \frac{1}{2}\mathbf{\Gamma}_n\psi , \quad (15.144)$$

where $\mathbf{\Gamma}_n$ is the N -tuple of bivectors (15.9). The covariant derivative of a spinor ψ is thus

$$D_n\psi = \partial_n\psi + \frac{1}{2}\mathbf{\Gamma}_n\psi , \quad (15.145)$$

In equation (15.145), as previously in equations (15.6) and (15.15), for a spinor $\psi = \epsilon_a\psi^a$, the directed derivative ∂_n is to be interpreted as acting only on the components ψ^a of the spinor, $\partial_n\psi = \epsilon_a\partial_n\psi^a$. In the convention (41.78) that multivectors acting to the right of a column spinor yield zero, the connection term in equation (15.145) can be written as a commutator, in the same form as (15.15),

$$\boxed{D_n\psi = \partial_n\psi + \frac{1}{2}[\mathbf{\Gamma}_n, \psi]} . \quad (15.146)$$

Acting on a spinor ψ , the Riemann curvature operator $\hat{R}_{kl}$, equation (15.21), yields another spinor,

$$\hat{R}_{kl}\psi = \frac{1}{2}\mathbf{R}_{kl}\psi . \quad (15.147)$$

In the convention (41.78) that multivectors acting to the right of a column spinor yield zero, equation (15.147) can be written in the same form as equation (15.23),

$$\hat{R}_{kl}\psi = \frac{1}{2}[\mathbf{R}_{kl}, \psi] . \quad (15.148)$$

15.17.1 Covariant derivative of a row spinor

A row Dirac spinor $\psi \cdot$ Lorentz transforms as $\psi \cdot \rightarrow \psi \cdot \bar{R}$, so an infinitesimal Lorentz transformation $\bar{R} = 1 - \epsilon \mathbf{\Gamma}/2$ generated by a bivector $\mathbf{\Gamma}$ transforms $\psi \cdot \rightarrow \psi \cdot - \frac{1}{2} \epsilon \psi \cdot \mathbf{\Gamma}$. Consequently the action of the connection operator $\hat{\Gamma}_n$ on a row spinor $\psi \cdot$ is

$$\hat{\Gamma}_n \psi \cdot = -\frac{1}{2} \psi \cdot \mathbf{\Gamma}_n , \quad (15.149)$$

and the covariant derivative of a row spinor $\psi \cdot$ is then

$$D_n \psi \cdot = \partial_n \psi \cdot - \frac{1}{2} \psi \cdot \mathbf{\Gamma}_n . \quad (15.150)$$

Again in the convention (41.78) that multivectors acting to the left of a row spinor yield zero, the connection term in equation (15.150) can be written as a commutator, in the same form as (15.15),

$$\boxed{D_n \psi \cdot = \partial_n \psi \cdot + \frac{1}{2} [\mathbf{\Gamma}_n, \psi \cdot]} . \quad (15.151)$$

Acting on a row spinor $\psi \cdot$, the Riemann curvature operator $\hat{R}_{kl}$, equation (15.21), yields another row spinor,

$$\hat{R}_{kl} \psi \cdot = -\frac{1}{2} \psi \cdot \mathbf{R}_{kl} . \quad (15.152)$$

Again in the convention (41.78) that multivectors acting to the left of a row spinor yield zero, equation (15.152) can be written in the same form as equation (15.23),

$$\hat{R}_{kl} \psi \cdot = \frac{1}{2} [\mathbf{R}_{kl}, \psi \cdot] . \quad (15.153)$$

Equations (15.15), (15.146), and (15.151) show that if a is any element of the supergeometric algebra, either a multivector or a column or row spinor, or a true scalar, its covariant derivative $D_n a$ can be written in the same form

$$D_n a = \partial_n a + \frac{1}{2} [\mathbf{\Gamma}_n, a] . \quad (15.154)$$

Likewise the action of the Riemann curvature operator $\hat{R}_{kl}$, equation (15.21), on any element a of the supergeometric algebra takes the same form

$$\hat{R}_{kl} a = \frac{1}{2} [\mathbf{R}_{kl}, a] . \quad (15.155)$$

15.18 Covariant derivative in a spinor basis

The covariant derivative D_n can also be expressed in a spinor basis.

The spinor tetrad connections Γ_{an}^b are defined, analogously to the definition (11.37) of the tetrad connections Γ_{mn}^k , to be the coefficients of the change of the spinor axes ϵ_a parallel-transported along the direction γ_n ,

$$\Gamma_{an}^b \epsilon_b \equiv \partial_n \epsilon_a . \quad (15.156)$$

The same equation (15.156) with a trailing dot appended on both sides holds for row spinors. The constancy of the spinor metric,

$$0 = \partial_n \varepsilon_{ab} = \partial_n (\epsilon_a \cdot \epsilon_b) = \Gamma_{bn}^c \epsilon_a \cdot \epsilon_c + \Gamma_{an}^c \epsilon_c \cdot \epsilon_b = \Gamma_{bn}^c \varepsilon_{ac} + \Gamma_{an}^c \varepsilon_{cb} = \Gamma_{abn} - \Gamma_{ban} , \quad (15.157)$$

along with the antisymmetry of the spinor metric, implies that the spinor tetrad connection Γ_{abn} is symmetric in its first two indices,

$$\Gamma_{abn} = \Gamma_{ban} . \quad (15.158)$$

The symmetry of the spinor tetrad connection is analogous to the antisymmetry of the tetrad connection, equation (11.47). The preservation of chirality under parallel transport implies that the spinor connections Γ_{abn} are non-vanishing only when a and b have the same chirality,

$$\Gamma_{abn} = 0 \quad \text{for } a, b \text{ of opposite chirality} . \quad (15.159)$$

In the 4D super spacetime algebra, the non-vanishing spinor connection coefficients comprise 12 right-handed spinor connections, and 12 left-handed spinor connections. The 24 spinor connection coefficients Γ_{abn} are related to the 24 tetrad connection coefficients Γ_{kmn} by

$$\Gamma_{abn} = \gamma_{ab}^{km} \Gamma_{kmn} , \quad (15.160)$$

where γ_{ab}^{km} is the matrix defined by equation (41.70) with km running over the 6 bivector indices, and km are implicitly summed over distinct bivector indices. In the chiral representation, the matrix coefficients are given by equations (41.64) and (41.65).

The connection $\mathbf{\Gamma}_n$ defined by equation (15.9) is in terms of the spinor basis

$$\mathbf{\Gamma}_n = \Gamma_{abn} \epsilon^a \epsilon^b , \quad (15.161)$$

implicitly summed over distinct symmetric self-chiral pairs ab of spinor indices. Expressions (15.15), (15.146), and (15.151) for the covariant derivatives of multivectors and spinors remain valid with the connection $\mathbf{\Gamma}_n$ given by equation (15.161).

15.19 Covariant spacetime derivative of a spinor

Acting on a Dirac column spinor ψ , the covariant spacetime derivative $\mathbf{D} \equiv \boldsymbol{\gamma}^n D_n$ yields another Dirac spinor

$$\mathbf{D}\psi \quad \text{column spinor} . \quad (15.162)$$

This derivative is a fundamental ingredient in the Lagrangian for a Dirac field, and in the resulting Dirac equations of motion.

The covariant spacetime derivative of a row spinor $\psi \cdot$ is defined to equal the row spinor corresponding to the covariant spacetime derivative of the column spinor ψ , that is, $\mathbf{D}\psi \cdot \equiv (\mathbf{D}\psi) \cdot$. The following manipulation

shows that the spacetime derivative of the row spinor is minus the spacetime derivative acting on the row spinor to the left:

$$\mathbf{D}\psi \cdot \equiv (\mathbf{D}\psi) \cdot = (\mathbf{D}\psi)^\top \varepsilon = \psi^\top \overleftarrow{\mathbf{D}}^\top \varepsilon = -\psi^\top \varepsilon \overleftarrow{\mathbf{D}} = -\psi \cdot \overleftarrow{\mathbf{D}} . \quad (15.163)$$

The penultimate step is true because $\gamma^{n\top} \varepsilon = -\varepsilon \gamma^n$, equation (41.40).

15.20 Gauss' theorem for spinors

In practical applications to spinor Lagrangians, Gauss' theorem occurs in the form

$$\int (\chi \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}\chi) d^4x = \int \mathring{D}^n (\chi \cdot \gamma_n \psi) d^4x = \oint \chi \cdot \gamma_n \psi d^3x^n , \quad (15.164)$$

where ψ and χ are spinors, and $\mathbf{D}$ is the torsion-full covariant spacetime derivative.

Equation (15.164) is proved as follows:

$$\begin{aligned} \chi \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}\chi &= \chi \cdot \mathbf{D}\psi - (\mathbf{D}\chi) \cdot \psi \\ &= \chi \cdot \gamma^n D_n \psi - (\gamma^n D_n \chi) \cdot \psi \\ &= \chi \cdot \gamma^n D_n \psi + (D_n \chi) \cdot \gamma^n \psi \\ &= \chi \cdot \gamma^n (\mathring{D}_n + \tfrac{1}{2} \mathbf{K}_n) \psi + \left((\mathring{D}_n + \tfrac{1}{2} \mathbf{K}_n) \chi \right) \cdot \gamma^n \psi \\ &= \chi \cdot \gamma^n \mathring{D}_n \psi + (\mathring{D}_n \chi) \cdot \gamma^n \psi + \tfrac{1}{2} \chi \cdot \gamma^n \mathbf{K}_n \psi - \tfrac{1}{2} \chi \cdot \mathbf{K}_n \gamma^n \psi \\ &= \mathring{D}_n (\chi \cdot \gamma^n \psi) - \chi \cdot \left(\mathring{D}_n \gamma^n + \tfrac{1}{2} [\mathbf{K}_n, \gamma^n] \right) \psi \\ &= \mathring{D}_n (\chi \cdot \gamma^n \psi) , \end{aligned} \quad (15.165)$$

where $\mathbf{K}_n \equiv K_{kl n} \gamma^k \wedge \gamma^l$ (Implicit sum over distinct antisymmetric indices kl) is the contortion, equation (15.47). The sign flip on the first line comes from the anticommutation of Dirac spinors, equation (41.122). The sign flip is cancelled on the third line from commuting the basis vectors γ^n through the spinor metric, equation (41.40). The sign flip in the last term on the fifth line is from $\mathbf{K}_n^\top = -\mathbf{K}_n$, valid for any bivector. The last term on the penultimate line of equation (15.165) vanishes because the torsion-full covariant derivative of the basis vectors γ^n vanishes, $\mathring{D}_n \gamma^n + \tfrac{1}{2} [\mathbf{K}_n, \gamma^n] = D_n \gamma^n = 0$.

Action principle for electromagnetism and gravity

One of the profound realisations of physics in the second half of the twentieth century was that the four forces of the Standard Model of physics — the electromagnetic, weak, strong (or colour), and gravitational forces — all emerge from an action that is invariant with respect to local symmetries called gauge transformations. Gauge transformations rotate internal degrees of freedom of fields at each point of spacetime.

The simplest of the forces is the electromagnetic force, which is based on the 1-dimensional unitary group $U(1)_Q$ of rotations about a circle. Since the mid 1970s, the electromagnetic group has been understood to be the unbroken remnant of a larger electroweak group $U(1)_Y \times SU(2)_L$, which through interactions with a scalar field called the Higgs field breaks down to the electromagnetic group $U(1)_Q$ at collision energies less than the electroweak scale of about 1 TeV (the $U(1)_Y$ electroweak hypercharge group is not the same as the $U(1)_Q$ electromagnetic group). The group $SU(N)$ is the special unitary group in N dimensions, the group of N -dimensional unitary matrices of unit determinant. The colour group is $SU(3)_c$.

The gravitational force is likewise a gauge force. The gravitational group is the group of spacetime transformations, also known as the Poincaré group, which is the product of the 6-dimensional Lorentz group of rotations and the 4-dimensional group of translations.¹

It is quite remarkable that so much of physics is captured by so simple a mathematical structure as a group of symmetries. During the 1980s there was hope that perhaps all of physics might be described by some theory-of-everything group, and all that was left to do was to discover that group and figure out its consequences. That hope was not realised.

Gravity has been at the heart of the problem. Whereas the three other forces are successfully described by renormalizable quantum field theories, albeit equipped with a large number of seemingly arbitrary parameters, gravity has resisted quantization. Currently the most successful (some would dispute that adjective)

¹ Technically the Poincaré group refers to the global symmetries of Minkowski space, where rotations do not commute with translations (rotation followed by translation yields a different result from translation followed by rotation). The Poincaré group is said to be a semi-direct product of rotations and translations. In general relativity, the translation group of Minkowski space is replaced by general coordinate transformations, which commute with local Lorentz transformations. In general relativity, coordinate transformations should be thought of as simply relabelling coordinates while leaving the underlying physical spacetime unchanged; and similarly local Lorentz transformation should be thought of as changing the tetrad axes with respect to which the locally inertial frame is measured, again while leaving the underlying spacetime unchanged.

theory of quantum gravity is string theory, or more specifically superstring theory (which includes spin- $\frac{1}{2}$ particles), or more specifically some enveloping theory that contains not only strings, 1-dimensional objects sweeping out 2-dimensional worldsheets, but also fundamental objects, branes, of other dimensions. The topic of string theory is beyond the scope of this book. Suffice to say that string theory is apparently a much larger and richer theory than a putative theory of just our Universe. The good thing is that string theory (probably) contains the laws of physics of our Universe. The embarrassing thing (embarras de richesses, advocates would say) is that string theory (probably) contains many other possible laws of physics. This has led to the conjecture that our Universe is just one of a multiverse of universes with different sets of laws of physics. Such ideas are fascinating, but at present distanced from experimental or observational reality. String theory remains work in progress.

This Chapter starts by applying the action principle to the simple example of an unspecified template field φ , deriving the Euler-Lagrange equations in §16.1, and Hamilton's equations in §16.2.

The Chapter goes on to apply the action principle to the simplest example of a gauge field, the electromagnetic field, first in index notation, §16.5, and then in the more difficult but powerful language of differential forms, §16.6. The electromagnetic example brings out features that appear in more complicated form in the gravitational field. Notably, the covariant equations of motion for the electromagnetic field resolve into genuine equations of motion for physical degrees of freedom, constraint equations whose ongoing satisfaction is guaranteed by conservation laws arising from gauge symmetries, and identities that define auxiliary fields that arise in a covariant treatment.

The Chapter then proceeds to apply the action principle to the Hilbert (1915) Lagrangian to derive the equations of motion of gravity, namely the Einstein equations, along with equations for the connection coefficients. The tetrad-frame approach followed in this Chapter makes manifest the dependence of Hilbert's Lagrangian on the two distinct symmetries of general relativity, namely symmetry with respect to local Lorentz transformations, and symmetry with respect to general coordinate transformations.

The Chapter treats the gravitational action using three different mathematical languages, progressing from the more explicit to the more abstract. The first approach, starting at §16.7, lays out all indices explicitly. The second approach, §16.13, uses multivectors. The final approach, §16.14, uses multivector-valued differential forms. The multivector forms notation provides an elegant formulation of the definitions of curvature and torsion, equations (16.210) and (16.214), first formulated by Cartan (1904), and elegant versions of the equations of motion (16.252) that govern them. The dense, abstract notation can be hard to unravel (which is why more explicit approaches are helpful), but offers the clearest picture of the structure of the gravitational equations. A clear picture is essential both from the practical perspective of numerical relativity, and from the esoteric perspective of aspiring to a deeper understanding of the unsolved mysteries of (quantum) gravity.

As expounded in Chapter 15, d^4x denotes the invariant scalar 4-volume element, equation (15.80), while d^4x^{0123} denotes the pseudoscalar coordinate 4-volume element, the indices 0123 serving as a reminder that the coordinate 4-volume element is a totally antisymmetric coordinate tensor of rank 4. The two are related by a factor of the determinant e of the vierbein, $d^4x = e d^4x^{0123}$, equation (15.88).

16.1 Euler-Lagrange equations for a generic field

Let $\varphi(x^\mu)$ denote some unspecified classical continuous field defined throughout spacetime. The least action principle asserts that the equations of motion governing the field can be obtained by minimizing an action S , which is asserted to be an integral over spacetime of a certain scalar Lagrangian. The scalar Lagrangian is asserted to be a function $L(\varphi, \mathring{D}_\mu \varphi)$ of “coordinates” which are the values of the field $\varphi(x^\mu)$ at each point of spacetime, and of “velocities” which are the torsion-free covariant derivatives $\mathring{D}_\mu \varphi$ of the field. The torsion-free covariant derivatives are prescribed because application of the least action principle involves integration by parts, and, as established in Chapter 15, equations (15.91) or (15.99), it is precisely the torsion-free covariant derivative that can be integrated to yield surface terms.

It should be commented that in the case of spinors ψ , the Lagrangian can be considered to be a function $L(\psi, D_\mu \psi)$ of the spinor field ψ and its torsion-full covariant derivative $D_\mu \psi$, since Gauss’ theorem occurs in a form (15.164) where the contortion contribution vanishes on integration by parts. The action principle for spinor fields is deferred to Chapter 44.

The Lagrangian $L(\varphi, \mathring{D}_\mu \varphi)$ is actually a function of functions. Mathematicians refer to such a thing as a **functional**. Derivatives of a functional with respect to the functions it depends on are called functional derivatives, or variational derivatives, and are denoted with a δ symbol. For example, the derivative of the functional L with respect to the function φ is denoted $\delta L / \delta \varphi$.

Least action postulates that the evolution of the field is such that the action

$$S = \int_{\lambda_i}^{\lambda_f} L(\varphi, \mathring{D}_\mu \varphi) d^4x \quad (16.1)$$

takes a minimum value with respect to arbitrary variations of the field, subject to the constraint that the field is fixed on its boundary, the initial and final surfaces. The integral in equation (16.1) is over 4-dimensional spacetime between a fixed initial 3-dimensional hypersurface and a fixed final 3-dimensional hypersurface, labelled respectively λ_i and λ_f . The variation δS of the action with respect to the field and its derivatives is

$$\delta S = \int_{\lambda_i}^{\lambda_f} \left(\frac{\delta L}{\delta \varphi} \delta \varphi + \frac{\delta L}{\delta (\mathring{D}_\mu \varphi)} \delta (\mathring{D}_\mu \varphi) \right) d^4x = 0 . \quad (16.2)$$

Linearity of the covariant derivative,

$$\mathring{D}_\mu (\varphi + \delta \varphi) = \mathring{D}_\mu \varphi + \mathring{D}_\mu (\delta \varphi) , \quad (16.3)$$

implies that the variation of the derivative equals the derivative of the variation, $\delta (\mathring{D}_\mu \varphi) = \mathring{D}_\mu (\delta \varphi)$. Define the canonical momentum π^μ conjugate to the field φ to be

$$\pi^\mu \equiv \frac{\delta L}{\delta (\mathring{D}_\mu \varphi)} . \quad (16.4)$$

The second term in the integrand of equation (16.2) can be written

$$\pi^\mu \delta (\mathring{D}_\mu \varphi) = \mathring{D}_\mu (\pi^\mu \delta \varphi) - (\mathring{D}_\mu \pi^\mu) \delta \varphi , \quad (16.5)$$

The first term on the right hand side of equation (16.5) is a torsion-free covariant divergence, which integrates to a surface term. With the second term in the integrand of equation (16.2) thus integrated by parts, the variation of the action is

$$\delta S = \left[\oint \pi_\mu \delta \varphi d^3 x^\mu \right]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left(\frac{\delta L}{\delta \varphi} - \dot{D}_\mu \pi^\mu \right) \delta \varphi d^4 x = 0 . \quad (16.6)$$

The surface term in equation (16.6), which is an integral over each of the three-dimensional initial and final hypersurfaces, vanishes since by hypothesis the fields are fixed on the initial and final hypersurfaces, $\delta \varphi_i = \delta \varphi_f = 0$. Consequently the integral term must also vanish. Least action demands that the integral vanish for all possible variations $\delta \varphi$ of the field. The only way this can happen is that the integrand must be identically zero. The result is the Euler-Lagrange equations of motion for the field,

$$\boxed{\dot{D}_\mu \pi^\mu = \frac{\delta L}{\delta \varphi}} . \quad (16.7)$$

All of the above derivations carry through with the field φ replaced by a set of fields φ_i , with conjugate momenta $\pi^{\mu i} \equiv \delta L / \delta (\dot{D}_\mu \varphi_i)$. The index i could simply enumerate a list of fields, or it could signify the components of a set of fields that transform into each other under some group of symmetries.

16.2 Super-Hamiltonian formalism

The Lagrangians L of the fields that Nature fields turn out to be writable in super-Hamiltonian form

$$L = \pi^\mu \dot{D}_\mu \varphi - H , \quad (16.8)$$

in which the super-Hamiltonian $H(\varphi, \pi^\mu)$ is a scalar function of the field φ and its conjugate momenta π^μ , defined in terms of the Lagrangian by equation (16.4).

Varying the action with Lagrangian (16.8) with respect to the field φ and its conjugate momenta π^μ gives

$$\delta S = \int_{\lambda_i}^{\lambda_f} \left(\pi^\mu \dot{D}_\mu \delta \varphi + \delta \pi^\mu \dot{D}_\mu \varphi - \frac{\delta H}{\delta \varphi} \delta \varphi - \delta \pi^\mu \frac{\delta H}{\delta \pi^\mu} \right) d^4 x . \quad (16.9)$$

Integrating the first term in the integrand by parts brings the variation of the action to

$$\delta S = \left[\oint \pi^\mu \delta \varphi d^3 x_\mu \right]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left[- \left(\dot{D}_\mu \pi^\mu + \frac{\delta H}{\delta \varphi} \right) \delta \varphi + \delta \pi^\mu \left(\dot{D}_\mu \varphi - \frac{\delta H}{\delta \pi^\mu} \right) \right] d^4 x . \quad (16.10)$$

The principle of least action requires that the variation vanish with respect to arbitrary variations $\delta \varphi$ and $\delta \pi^\mu$ of the field and its conjugate momenta, subject to the condition that the field is held fixed on the initial and final hypersurfaces. The result is **Hamilton's equations** of motion,

$$\boxed{\dot{D}_\mu \pi^\mu = - \frac{\delta H}{\delta \varphi} , \quad \dot{D}_\mu \varphi = \frac{\delta H}{\delta \pi^\mu}} . \quad (16.11)$$

Hamilton's equations (16.11) for the field φ can be compared to Hamilton's equations (4.72) for particles.

16.3 Conventional Hamiltonian formalism

The conventional Hamiltonian is not the same as the super-Hamiltonian. In the conventional Hamiltonian formalism, the coordinates x^μ are split into a time coordinate t and spatial coordinates x^α . The momentum π conjugate to the field φ is defined to be

$$\pi \equiv \frac{\delta L}{\delta(\dot{D}_t \varphi)} . \quad (16.12)$$

The conventional Hamiltonian H is defined in terms of the Lagrangian L by

$$H \equiv \pi \dot{D}_t \varphi - L . \quad (16.13)$$

In the context of general relativity, the covariant super-Hamiltonian approach to fields is, as in the case of point particles, §4.10, simpler and more natural than the non-covariant conventional Hamiltonian approach. Indeed, the most straightforward way to implement the conventional Hamiltonian approach is to use the super-Hamiltonian approach, and then carry out a 3+1 split into space and time coordinates at the end, rather than doing a 3+1 split at the outset.

16.4 Symmetries and conservation laws

Associated with every symmetry is a conserved quantity. The relation between symmetries and conserved quantities is called **Noether's theorem** (Noether, 1918), equations (16.17) and (16.18). Examples of Noether's theorem include local electromagnetic gauge symmetry implying conservation of electric charge (§16.5.6), local Lorentz symmetry implying conservation of angular momentum (§16.11.1), and general coordinate transformations implying conservation of energy-momentum (§16.11.2).

All four of the known forces of Nature, including gravity, arise from local symmetries, in which the Lagrangian is invariant under symmetry transformations that are allowed to vary arbitrarily over spacetime. Commonly, such transformations change not just one field, but multiple fields at the same time. However, the Lagrangian of an individual field may by itself be symmetric, to the extent that the field does not interact with other fields. For example, the local gauge symmetry of electromagnetism changes simultaneously the electromagnetic field and all charged fields, and that symmetry implies the law of conservation of total electric charge. However, an individual field, such as an electron field or a proton field, may individually conserve charge, to the extent that the field does not interact with other fields.

Consider varying the template field $\varphi(x)$ by a transformation with a prescribed shape $\delta\varphi(x)$ as a function of spacetime,

$$\varphi(x) \rightarrow \varphi(x) + \epsilon \delta\varphi(x) , \quad (16.14)$$

where ϵ is an infinitesimal constant parameter. The torsion-free covariant derivatives $\dot{D}_m \varphi$ of the field transform correspondingly as

$$\dot{D}_m \varphi \rightarrow \dot{D}_m \varphi + \epsilon \dot{D}_m (\delta\varphi) . \quad (16.15)$$

The torsion-free covariant derivative is prescribed for the reason explained at the beginning of §16.1. The variation (16.14) is a symmetry of the field if the Lagrangian $L(\varphi, \dot{D}_m \varphi)$ is unchanged by it. The vanishing of the variation of the Lagrangian implies

$$\begin{aligned}
 0 &= \frac{\delta L}{\delta \epsilon} \\
 &= \frac{\delta L}{\delta \varphi} \delta \varphi + \frac{\delta L}{\delta(\dot{D}_m \varphi)} \dot{D}_m(\delta \varphi) \\
 &= \dot{D}_m \left(\frac{\delta L}{\delta(\dot{D}_m \varphi)} \delta \varphi \right) + \left(\frac{\delta L}{\delta \varphi} - \dot{D}_m \frac{\delta L}{\delta(\dot{D}_m \varphi)} \right) \delta \varphi \\
 &= \dot{D}_m (\pi^m \delta \varphi) + \left(\frac{\delta L}{\delta \varphi} - \dot{D}_m \pi^m \right) \delta \varphi ,
 \end{aligned} \tag{16.16}$$

with π^m the momentum conjugate to the field, equation (16.4). The Euler-Lagrange equation of motion (16.7) for the field implies that the second term on the last line of (16.16) vanishes. Consequently the current j^m defined by

$$\boxed{j^m \equiv \pi^m \delta \varphi} \tag{16.17}$$

is covariantly conserved,

$$\boxed{\dot{D}_m j^m = 0} . \tag{16.18}$$

The result (16.18) is Noether's theorem.

16.5 Electromagnetic action

Electromagnetism is a gauge field based on the simplest of all continuous groups, the 1-dimensional unitary group $U(1)_Q$ of rotations about a circle.

16.5.1 Electromagnetic gauge transformations

Under an electromagnetic gauge transformation, a field φ of charge e transforms as

$$\varphi \rightarrow e^{-ie\theta} \varphi , \tag{16.19}$$

where the phase $\theta(x)$ is some arbitrary function of spacetime. The charge e is dimensionless (in units $c = \hbar = 1$). The Lagrangian of the charged field φ involves the torsion-free derivative $\dot{D}_\mu \varphi$ of the field. The torsion-free covariant derivative is prescribed for the reason explained at the beginning of §16.1. To ensure that the Lagrangian remains invariant also under an electromagnetic gauge transformation (16.19), the derivative $\dot{D}_\mu$ must be augmented by an electromagnetic connection A_μ , which equals the thing historically known as the **electromagnetic potential**. The result is an electromagnetic gauge-covariant derivative $\dot{D}_\mu + ieA_\mu$ with

the defining property that, when acting on the charged field φ , it transforms under the electromagnetic gauge transformation (16.19) as

$$(\mathring{D}_\mu + ieA_\mu)\varphi \rightarrow e^{-ie\theta}(\mathring{D}_\mu + ieA_\mu)\varphi . \quad (16.20)$$

In other words, the gauge-covariant derivative of the field φ is required to transform under electromagnetic gauge transformations in the same way as the field φ . The gauge-covariant derivative $\mathring{D}_\mu + ieA_\mu$ transforms correctly provided that the gauge field A_μ transforms under the electromagnetic gauge transformation (16.19) as

$$A_\mu \rightarrow A_\mu + \mathring{D}_\mu \theta . \quad (16.21)$$

Since θ is a scalar phase, its covariant derivative reduces to its partial derivative, $\mathring{D}_\mu \theta = \partial\theta/\partial x^\mu$.

16.5.2 Electromagnetic field tensor

The commutator of the gauge-covariant derivative $\mathring{D}_\mu + ieA_\mu$ defines the **electromagnetic field** tensor $F_{\mu\nu}$,

$$[\mathring{D}_\mu + ieA_\mu, \mathring{D}_\nu + ieA_\nu] \equiv ieF_{\mu\nu} . \quad (16.22)$$

The electromagnetic field $F_{\mu\nu}$ has the key property that it is invariant under an electromagnetic gauge transformation (16.19), in contrast to the electromagnetic potential A_ν itself. Explicitly, the electromagnetic field $F_{\mu\nu}$ is, from equation (16.22),

$$\begin{aligned} F_{\mu\nu} &\equiv \mathring{D}_\mu A_\nu - \mathring{D}_\nu A_\mu \\ &= \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} , \end{aligned} \quad (16.23)$$

the second line of which follows because the coordinate connections cancel in a torsion-free covariant coordinate curl, equation (2.72). The expression on the second line of equations (16.23) is invariant under an electromagnetic gauge transformation (16.21) thanks to the commutation of coordinate derivatives, $\partial^2\theta/\partial x^\mu\partial x^\nu - \partial^2\theta/\partial x^\nu\partial x^\mu = 0$, so the electromagnetic field $F_{\mu\nu}$ is electromagnetic gauge-invariant as claimed. If the torsion-free derivative $\mathring{D}_\mu$ in equation (16.23) were replaced by the torsion-full derivative D_μ , then the electromagnetic field $F_{\mu\nu}$ would not be electromagnetic gauge-invariant.

16.5.3 Source-free Maxwell's equations

For brevity, denote the electromagnetic gauge-covariant derivative by $\mathcal{D}_\mu \equiv \mathring{D}_\mu + ieA_\mu$. The gauge-covariant derivative satisfies the Jacobi identity

$$[\mathcal{D}_\lambda, [\mathcal{D}_\mu, \mathcal{D}_\nu]] = 0 . \quad (16.24)$$

The electromagnetic Jacobi identity (16.24) implies that

$$\mathring{D}_\lambda F_{\mu\nu} + \mathring{D}_\mu F_{\nu\lambda} + \mathring{D}_\nu F_{\lambda\mu} = 0 . \quad (16.25)$$

Since the torsion-free coordinate connections cancel in such an antisymmetrized expression, equation (16.25) can also be written

$$\frac{\partial F_{\mu\nu}}{\partial x^\lambda} + \frac{\partial F_{\nu\lambda}}{\partial x^\mu} + \frac{\partial F_{\lambda\mu}}{\partial x^\nu} = 0 . \quad (16.26)$$

Equations (16.26) constitute a set of 4 equations comprising the source-free Maxwell's equations.

16.5.4 Electromagnetic Lagrangian

The electromagnetic action S_e is

$$S_e = \int_{\lambda_i}^{\lambda_f} L_e d^4x , \quad (16.27)$$

with electromagnetic Lagrangian

$$L_e \equiv -\frac{1}{16\pi} F^{\mu\nu} F_{\mu\nu} , \quad (16.28)$$

where $F_{\mu\nu}$ is the electromagnetic field tensor defined by equation (16.23). The electromagnetic Lagrangian L_e , equation (16.28) is, as required, a scalar with respect to electromagnetic gauge transformations (16.21), as well as with respect to coordinate and tetrad transformations. The justification for the choice (16.28) is that it reproduces Maxwell's equations, which have ample experimental verification. The Lagrangian (16.28) is normalized to Gaussian units. High-energy physicists commonly used Heaviside units (SI units with $\varepsilon_0 = \mu_0 = 1$), for which the normalization factor is 1/4 instead of 1/(16 π). High-energy theorists prefer Heaviside units because then the charge e measures the strength with which an electron couple to the electromagnetic field, equation (44.2).

The momenta conjugate to the electromagnetic coordinates A_ν are, modulo a factor, the electromagnetic field components $F^{\mu\nu}$,

$$\frac{\delta L_e}{\delta(\dot{D}_\mu A_\nu)} = -\frac{1}{4\pi} F^{\mu\nu} . \quad (16.29)$$

In Heaviside instead of Gaussian units, the factor is 1 instead of 4 π .

In the presence of electrically charged matter, the matter action generically contains an interaction term S_q

$$S_q = \int_{\lambda_i}^{\lambda_f} L_q d^4x , \quad (16.30)$$

with interaction Lagrangian L_q taking the form

$$L_q = A_\nu j^\nu , \quad (16.31)$$

where j^ν is the electric current vector.

The combined electromagnetic and charged matter action $S = S_e + S_q$ is, with the Lagrangian expressed as required in terms of the electromagnetic coordinates A_ν and their velocities $\dot{D}_\mu A_\nu$,

$$S = \int_{\lambda_i}^{\lambda_f} \left[-\frac{1}{16\pi} \left(\dot{D}^\mu A^\nu - \dot{D}^\nu A^\mu \right) \left(\dot{D}_\mu A_\nu - \dot{D}_\nu A_\mu \right) + j^\nu A_\nu \right] d^4x . \quad (16.32)$$

Varying the action (16.32) with respect to the electromagnetic coordinates A_ν and their velocities $\dot{D}_\mu A_\nu$, along the same lines as equations (16.2)–(16.6) for the template field φ , yields

$$\delta S = -\frac{1}{4\pi} \left[\oint F^{\mu\nu} \delta A_\nu d^3x_\mu \right]_{\lambda_i}^{\lambda_f} + \frac{1}{4\pi} \int_{\lambda_i}^{\lambda_f} \left(\dot{D}_\mu F^{\mu\nu} + 4\pi j^\nu \right) \delta A_\nu d^4x . \quad (16.33)$$

Least action requires that the variation of the action with respect to arbitrary variations δA_ν be zero, subject to the constraint that the field is fixed on the boundary of integration, $\delta A_\nu = 0$. The resulting Euler-Lagrange equations (16.7) are

$$\dot{D}_\mu F^{\nu\mu} = 4\pi j^\nu . \quad (16.34)$$

The factor 4π disappears if Heaviside units are used in place of Gaussian units. The Euler-Lagrange equations (16.34) constitute 4 equations comprising the source-full Maxwell's equations.

Concept question 16.1 What are the units of the electromagnetic action? **Answer.** Regardless of the choice of unit for charge e , the product eA has units of 1/length. and the product eF has units 1/length². The SI unit of charge is the Coulomb, but that definition stems from ancient experiment. The most natural unit for charge is dimensionless, consistent with its appearance in the gauge transformation (16.19). High-energy physicists define $e = \sqrt{4\pi\alpha}$ where α is the fine-structure constant; e is a dimensionless measure of the coupling strength of the electromagnetic field. In those units, the electromagnetic action (16.27) is dimensionless.

16.5.5 Electromagnetic super-Hamiltonian

The electromagnetic Lagrangian (16.28), coupled with the charged matter interaction Lagrangian (16.31), is in super-Hamiltonian form $L_e + L_q = p^\mu \partial_\mu q - H$ with coordinates $q = A_\nu$ and momenta $p^\mu = -F^{\mu\nu}/4\pi$,

$$L_e = -\frac{1}{4\pi} F^{\mu\nu} \dot{D}_\mu A_\nu - H , \quad (16.35)$$

and super-Hamiltonian H

$$H \equiv -\frac{1}{16\pi} F^{\mu\nu} F_{\mu\nu} - A_\nu j^\nu . \quad (16.36)$$

The Hamiltonian (16.36) looks like the Lagrangian but with a flip of the sign of the interaction term $A_\nu j^\nu$. The electromagnetic Hamiltonian (16.36) is expressed as required in terms of the coordinates A_ν and the momenta $F^{\mu\nu}$.

Varying the action with Lagrangian (16.35) with respect to the coordinates A_ν and momenta $F^{\mu\nu}$ gives

$$\delta S = \frac{1}{4\pi} \int (-F^{\mu\nu} \dot{D}_\mu \delta A_\nu - \delta F^{\mu\nu} \dot{D}_\mu A_\nu + \frac{1}{2} \delta F^{\mu\nu} F_{\mu\nu} + 4\pi j^\nu \delta A_\nu) d^4x . \quad (16.37)$$

Integrating the first term in the integrand of equation (16.37) by parts yields

$$\delta S = -\frac{1}{4\pi} \oint F^{\mu\nu} \delta A_\nu d^3x_\mu + \frac{1}{4\pi} \int \left[(\dot{D}_\mu F^{\mu\nu} + 4\pi j^\nu) \delta A_\nu - \frac{1}{2} (\dot{D}_\mu A_\nu - \dot{D}_\nu A_\mu + F_{\mu\nu}) \delta F^{\mu\nu} \right] d^4x . \quad (16.38)$$

The surface term vanishes provided that the electromagnetic coordinates A_ν are held fixed on the boundary. Requiring that the variation of the action vanish with respect to arbitrary variations δA_ν and $\delta F^{\mu\nu}$ of the coordinates and momenta then yields Hamilton's equations,

$$\dot{D}_\mu F^{\nu\mu} = 4\pi j^\nu , \quad (16.39a)$$

$$\dot{D}_\mu A_\nu - \dot{D}_\nu A_\mu = F_{\mu\nu} . \quad (16.39b)$$

The first Hamilton equation (16.39a) reproduces the Euler-Lagrange equation (16.34) obtained in the Lagrangian approach. The second Hamilton equation (16.39b) implies, as an equation of motion, the relation (16.23) between the field $F_{\mu\nu}$ and the derivatives of A_ν that was simply assumed in the Lagrangian approach.

16.5.6 Electric charge conservation

Maxwell's source-full equations (16.34) enforce covariant conservation of electric charge j^ν ,

$$\dot{D}_\nu j^\nu = 0 . \quad (16.40)$$

At a more profound level, the conservation of electric charge is a consequence of symmetry with respect to electromagnetic gauge transformations. Under an electromagnetic gauge transformation, the field A_ν varies as, equation (16.21),

$$\delta A_\nu = \dot{D}_\nu \theta . \quad (16.41)$$

There are many distinct electrically charged fields in nature (for example, electrons and protons), and the action for each distinct charged field is electromagnetic gauge-invariant (absent interactions that create or destroy charged fields). The variation of a charged matter field under an electromagnetic gauge transformation (16.19) is

$$\delta S_q = \int j^\nu \dot{D}_\nu \theta d^4x . \quad (16.42)$$

Integrating equation (16.42) by parts gives

$$\delta S_q = \oint j^\nu \theta d^3x_\nu - \int (\dot{D}_\nu j^\nu) \theta d^4x . \quad (16.43)$$

Electromagnetic gauge-invariance requires that the variation vanish with respect to arbitrary choices of the gauge parameter θ , subject to the condition that θ is fixed on the boundary. Covariant conservation of electric charge follows,

$$\mathring{D}_\nu j^\nu = 0 . \quad (16.44)$$

The charge conservation law (16.44) is an example of Noether's theorem (Noether, 1918), which relates symmetries and conservation laws.

16.5.7 Electromagnetic wave equation

Eliminating $F_{\mu\nu}$ from Hamilton's equations (16.39) yields a second order differential equation for the electromagnetic potential A_ν ,

$$-\mathring{\square}A_\nu + \mathring{R}_{\nu\lambda}A^\lambda + \mathring{D}_\nu\mathring{D}_\mu A^\mu = 4\pi j_\nu , \quad (16.45)$$

where $\mathring{\square} \equiv \mathring{D}_\mu\mathring{D}^\mu$ is the torsion-free d'Alembertian. The last term $\mathring{D}_\nu\mathring{D}_\mu A^\mu$ on the left hand side of equation (16.45) may be eliminated by imposing the Lorenz (not Lorentz!) gauge condition $\mathring{D}_\mu A^\mu = 0$. Equation (16.45) is a wave equation with the torsion-free Ricci tensor $\mathring{R}_{\nu\lambda}$ acting as an effective potential, and the electromagnetic current j_ν acting as a source.

16.5.8 Space+time (3+1) split of the electromagnetic equations

In Chapter 4 it was found that, applied to point particles, the action principle yielded equal numbers of coordinates and momenta, and Hamilton's equations supplied first order differential equations determining the evolution of each and every one of the coordinates and momenta. This was true in both the super-Hamiltonian and conventional Hamiltonian approaches, where Hamilton's equations were respectively equations (4.72) and (4.75).

Applied to fields, the super-Hamiltonian approach does not yield equal numbers of coordinates and momenta, and Hamilton's equations cannot be interpreted straightforwardly as equations of motion for each and every one of the coordinates and momenta. For example, in the electromagnetic case, the first set of Hamilton's equations (16.39a) apparently constitute 4 equations for 6 momenta $F^{\nu\mu}$, while the second set of Hamilton's equations (16.39b) apparently constitute 6 equations for 4 coordinates A_ν . The mismatch of numbers of equations is not a practical barrier to solving Hamilton's equations of motion. Hamilton's equations (16.39) comprise 10 equations for 10 unknowns. If, for example, the 6 equations (16.39b) are interpreted not as first order differential equations of motion for the coordinates A_ν , but rather as defining the 6 momenta $F_{\mu\nu}$, then eliminating the momenta yields a set of 4 second order differential wave equations for the 4 coordinates A_ν , equation (16.45) (see §27.6 for further exposition). Treating the 6 equations (16.39b) as identities is the same as reverting to the Lagrangian, or second order, approach.

It is nevertheless desirable to attain a better understanding of the first order Hamiltonian formalism for fields, partly so as to understand how to integrate the field equations numerically, and partly because

quantization of fields, as usually implemented, requires identifying the physical degrees of freedom in a matching number of fields and their conjugate momenta.

The problem of mismatching numbers of coordinates and momenta in the super-Hamiltonian formalism arises because symmetry under general coordinate transformations means that different configurations of fields are symmetrically equivalent. The covariant super-Hamiltonian description contains more fields than there are physical degrees of freedom.

Dirac's (1964) solution to the mismatch of numbers of equations is to break general covariance by splitting spacetime into space and time coordinates, and to interpret only the equations involving time derivatives of the fields as genuine equations of motion, while the remainder of the equations, those not involving time derivatives, are "constraints," relations between the fields that serve to remove the redundant degrees of freedom. In the relativist community, the term "constraint" is commonly used to describe an equation which must be arranged to be satisfied in the initial conditions, but which is guaranteed thereafter by some conservation law. Some of Dirac's constraint equations, which Dirac calls "first-class constraints," are of this character, but others, which Dirac calls "second-class constraints," are identities that effectively define some fields in terms of others. This book follows the relativists' convention that a constraint is an equation whose ongoing satisfaction is guaranteed by a conservation law, a first-class constraint. Dirac's second-class constraint equations will be called identities.

Suppose then that the coordinates are split into time and space components, $x^\mu = \{t, x^\alpha\}$. In electromagnetism, the Hamilton's equations (16.39) involving time derivatives of the coordinates and momenta are

$$3 \text{ equations of motion: } \dot{D}_t F^{\alpha t} + \dot{D}_\beta F^{\alpha\beta} = 4\pi j^\alpha, \quad (16.46a)$$

$$3 \text{ equations of motion: } \dot{D}_t A_\alpha - \dot{D}_\alpha A_t = F_{t\alpha}. \quad (16.46b)$$

Equation (16.46a) comprises 3 equations of motion for the 3 momenta $F^{\alpha t}$, while equation (16.46b) comprises 3 equations of motion for the 3 coordinates A_α . The physical degrees of freedom are thus identified as the 3 spatial coordinates A_α and their 3 conjugate momenta $F^{\alpha t}$, which comprise the 3 components $E^\alpha \equiv F^{t\alpha}$ of the electric field. The remaining electromagnetic Hamilton's equations (16.39), those not involving time derivatives of the coordinates and momenta, are

$$1 \text{ constraint: } \dot{D}_\beta F^{t\beta} = 4\pi j^t, \quad (16.47a)$$

$$3 \text{ identities: } \dot{D}_\alpha A_\beta - \dot{D}_\beta A_\alpha = F_{\alpha\beta}. \quad (16.47b)$$

The first equation (16.47a) has the property that, as long as the equation is satisfied on the initial spatial hypersurface, then conservation of electric charge ensures that the equation continues to be satisfied thereafter. Of course, in numerical computations charge is conserved only so long as the equations of motion of charged matter are chosen such as to conserve electric charge, as they should be. If the matter equations conserve charge, then the constraint equation (16.47a) is redundant, but provides a numerical check that electric charge is being conserved.

The second set of equations (16.47b) are identities relating the 3 purely spatial components $F_{\alpha\beta}$, which comprise the 3 components $B^\alpha \equiv \varepsilon^{t\alpha\beta\gamma} F_{\beta\gamma}$ of the magnetic field, to the spatial curl of the spatial coordinates

A_α . Since the equations of motion (16.46b) already determine completely the spatial coordinates A_α , the identities (16.47b) cannot be independent equations, but must be interpreted as defining the magnetic field as an auxiliary field that does not represent additional physical degrees of freedom. The magnetic field is needed as part of the equations of motion, the second term on the left hand side of the equation of motion (16.46a). The magnetic field could be discarded after having been replaced by the curl of A_α in accordance with the identity (16.47b); but the magnetic field is part of the covariant 4-dimensional electromagnetic field tensor $F_{\mu\nu}$, and discarding the magnetic field would obscure the covariant structure of the electromagnetic equations.

16.6 Electromagnetic action in forms notation

Especially in the mathematical literature, actions are often written in the compact notation of differential forms, §15.6. The advantage of forms notation is not that it makes calculations any easier, but rather that it reveals the structure of the action unburdened by indices. Once one gets over the language barrier, forms notation can be a powerful clarifier.

In this section 16.6, implicit sums are over distinct antisymmetric sequences of indices, since this removes the ubiquitous factorial factors that would otherwise appear.

16.6.1 Electromagnetic potential and field forms

The electromagnetic potential 1-form $\mathbf{A}$ and field 2-form $\mathbf{F}$ are defined by

$$\mathbf{A} \equiv A_\nu dx^\nu , \quad (16.48a)$$

$$\mathbf{F} \equiv F_{\mu\nu} d^2x^{\mu\nu} , \quad (16.48b)$$

where in the case of $\mathbf{F}$ the implicit summation is over distinct antisymmetric pairs $\mu\nu$ of indices. With the electromagnetic gauge-covariant derivative 1-form denoted $\mathcal{D} \equiv (\overset{\circ}{D}_\mu + ieA_\mu)dx^\mu$ for brevity, the field 2-form $\mathbf{F}$ is defined by the commutator of the gauge-covariant derivative,

$$[\mathcal{D}, \mathcal{D}] \equiv ie\mathbf{F} . \quad (16.49)$$

Equation (16.49) implies that the field 2-form $\mathbf{F}$ is the exterior derivative of the potential 1-form $\mathbf{A}$,

$$\mathbf{F} = d\mathbf{A} = \left(\frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} \right) d^2x^{\mu\nu} , \quad (16.50)$$

implicitly summed over distinct antisymmetric pairs $\mu\nu$ of indices.

16.6.2 Electromagnetic potential and field multivectors

When working with forms, it is often easier to do calculations in multivector language. In multivector language, the electromagnetic potential is a vector $\mathbf{A}$, while the electromagnetic field is a bivector $\mathbf{F}$,

$$\mathbf{A} \equiv A_n \boldsymbol{\gamma}^n , \quad (16.51a)$$

$$\mathbf{F} \equiv F_{mn} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n , \quad (16.51b)$$

with in the case of $\mathbf{F}$ implicit summation over distinct antisymmetric pairs mn of indices. The field $\mathbf{F}$, equation (16.50), is in multivector language the torsion-free covariant curl of the potential $\mathbf{A}$,

$$\mathbf{F} = \tilde{D} \wedge \mathbf{A} . \quad (16.52)$$

In multivector language, the combined electromagnetic (16.28) and charged interaction (16.31) Lagrangian is the scalar

$$L_e + L_q = \frac{1}{8\pi} \mathbf{F} \cdot \mathbf{F} + \mathbf{A} \cdot \mathbf{j} , \quad (16.53)$$

where $\mathbf{j} \equiv j_n \boldsymbol{\gamma}^n$ is the electric current vector. The action is

$$S = \int (L_e + L_q) d^4x . \quad (16.54)$$

Recall that the scalar volume element d^4x that goes into the action (16.54) is really the dual scalar 4-volume $*d^4x$, equation (15.80). To convert to forms language, the Hodge dual must be transferred from the volume element to the integrand. In multivector language, the required result is

$$(\mathbf{a} \cdot \mathbf{b}) \cdot *d^4x \equiv (\mathbf{a} \cdot \mathbf{b}) \cdot (I d^4x) = ((\mathbf{a} \cdot \mathbf{b})I) \cdot d^4x = (I(\mathbf{a} \cdot \mathbf{b})) \cdot d^4x = ((I\mathbf{a}) \wedge \mathbf{b}) \cdot d^4x , \quad (16.55)$$

where the second expression is the definition (15.80) of the dual volume element, the third expression is an application of the multivector triple-product relation (13.39), the fourth holds because $\mathbf{a} \cdot \mathbf{b}$ is a scalar and therefore commutes with the pseudoscalar I , and the last expression is another application of the triple-product relation (13.39). The action (16.54) is thus, in multivector language,

$$S = \int \left(\frac{1}{8\pi} (I\mathbf{F}) \wedge \mathbf{F} + (I\mathbf{j}) \wedge \mathbf{A} \right) \cdot d^4x . \quad (16.56)$$

16.6.3 Electromagnetic Lagrangian 4-form

In forms notation, the action (16.56) is

$$S = \int L_e + L_q , \quad (16.57)$$

with Lagrangian 4-form

$$L_e + L_q = \frac{1}{8\pi} *F \wedge F + *j \wedge A . \quad (16.58)$$

Here $\mathbf{A}$ and $\mathbf{F}$ are the potential 1-form and field 2-form defined by equations (16.48). The symbol $*$ denotes the form dual, equation (15.79). The dual $*\mathbf{F}$ is a 2-form, while the dual $*\mathbf{j}$ is the 3-form dual of the 1-form electric current $\mathbf{j} \equiv j_\nu dx^\nu$.

16.6.4 Electromagnetic super-Hamiltonian 4-form

The Lagrangian 4-form (16.58) is in super-Hamiltonian form $\mathbf{p} \wedge d\mathbf{q} - H$ with coordinates $\mathbf{q} = \mathbf{A}$ and momenta $\mathbf{p} = *\mathbf{F}/4\pi$,

$$L_e + L_q = \frac{1}{4\pi} *\mathbf{F} \wedge d\mathbf{A} - H, \quad (16.59)$$

and super-Hamiltonian 4-form

$$H = \frac{1}{8\pi} *\mathbf{F} \wedge \mathbf{F} - *\mathbf{j} \wedge \mathbf{A}. \quad (16.60)$$

The variation of the action with Lagrangian (16.59) with respect to the coordinates $\mathbf{A}$ and momenta $*\mathbf{F}$ is

$$\delta S = \frac{1}{4\pi} \int *\mathbf{F} \wedge d\delta\mathbf{A} + \delta *\mathbf{F} \wedge d\mathbf{A} - \delta *\mathbf{F} \wedge \mathbf{F} + 4\pi *\mathbf{j} \wedge \delta\mathbf{A}. \quad (16.61)$$

Integrating the $*\mathbf{F} \wedge d\delta\mathbf{A}$ term in equation (16.61) by parts brings the variation of the action to

$$\delta S = \frac{1}{4\pi} \oint *\mathbf{F} \wedge \delta\mathbf{A} + \frac{1}{4\pi} \int -(d*\mathbf{F} - 4\pi *\mathbf{j}) \wedge \delta\mathbf{A} + \delta *\mathbf{F} \wedge (d\mathbf{A} - \mathbf{F}). \quad (16.62)$$

Requiring that the variation of the action vanish with respect to arbitrary variations $\delta\mathbf{A}$ and $\delta *\mathbf{F}$ of the electromagnetic coordinates and momenta, subject to the condition that $\mathbf{A}$ is fixed on the boundary, yields Hamilton's equations,

$$\boxed{d*\mathbf{F} = 4\pi *\mathbf{j}}, \quad (16.63a)$$

$$\boxed{d\mathbf{A} = \mathbf{F}}. \quad (16.63b)$$

The first Hamilton equation (16.63a) is a 3-form with 4 components comprising Maxwell's source-full equations. The second Hamilton equation (16.63b) is a 2-form with 6 components that enforce the relation (16.50) between the electromagnetic field $\mathbf{F}$ and the electromagnetic potential $\mathbf{A}$ that is assumed in the Lagrangian formalism.

Taking the exterior derivative of the first Hamilton equation (16.63a) yields, since $d^2 = 0$, the electric current conservation law

$$\boxed{d*\mathbf{j} = 0}. \quad (16.64)$$

Taking the exterior derivative of the second Hamilton equation (16.63b) yields

$$d\mathbf{F} = 0, \quad (16.65)$$

which comprises Maxwell's source-free equations.

16.6.5 Electromagnetic wave equation in forms notation

As is common, it is easier to manipulate form equations by translating them into multivector language. In multivector language, the electromagnetic Hamilton's equations (16.63) are

$$\mathring{D} \cdot \mathbf{F} = -4\pi \mathbf{j} , \quad (16.66a)$$

$$\mathring{D} \wedge \mathbf{A} = \mathbf{F} . \quad (16.66b)$$

Applying the multivector triple-product relation (13.40) gives the multivector identities (the torsion-free squared curl of $\mathbf{A}$ vanishes, equation (15.43), so $\mathring{D}\mathring{D}\mathbf{A}$ has only a vector part, no trivector part)

$$\begin{aligned} \mathring{D}\mathring{D}\mathbf{A} &= \mathring{D}(\mathring{D}\mathbf{A}) = \mathring{D} \cdot (\mathring{D} \wedge \mathbf{A}) + \mathring{D} \wedge (\mathring{D} \cdot \mathbf{A}) \\ &= (\mathring{D}\mathring{D})\mathbf{A} = (\mathring{D} \cdot \mathring{D})\mathbf{A} + (\mathring{D} \wedge \mathring{D}) \cdot \mathbf{A} . \end{aligned} \quad (16.67)$$

Eliminating $\mathbf{F}$ from Hamilton's equations (16.66) then yields a second order differential equation for the electromagnetic potential $\mathbf{A}$,

$$-\mathring{\square}\mathbf{A} - (\mathring{D} \wedge \mathring{D}) \cdot \mathbf{A} + \mathring{D}(\mathring{D} \cdot \mathbf{A}) = 4\pi \mathbf{j} , \quad (16.68)$$

where $\mathring{\square} \equiv \mathring{D} \cdot \mathring{D}$ is the torsion-free d'Alembertian operator. Equation (16.68) is equation (16.45) expressed in multivector language. The last term on the left hand side of equation (16.68) can be made to vanish by imposing the Lorenz gauge condition $\mathring{D} \cdot \mathbf{A} = 0$, in which case equation (16.68) reduces to

$$-\mathring{\square}\mathbf{A} - (\mathring{D} \wedge \mathring{D}) \cdot \mathbf{A} = 4\pi \mathbf{j} , \quad (16.69)$$

or more simply

$$-(\mathring{D}\mathring{D})\mathbf{A} = 4\pi \mathbf{j} . \quad (16.70)$$

Equation (16.70) is a wave equation for the electromagnetic potential $\mathbf{A}$, with source the electric current $\mathbf{j}$.

16.6.6 Space+time (3+1) split of the electromagnetic equations in forms notation

As discussed in §16.5.8, the super-Hamiltonian approach yields different numbers of coordinates and momenta, and the resulting Hamilton's equations are unbalanced. Hamilton's equations (16.63) have the appearance of first order differential equations of motion for the momenta and coordinates, but the first equation (16.63a) is 4 equations for the 6 components of the momenta ${}^*\mathbf{F}$, while the second equation (16.63b) is 6 equations for the 4 components of the coordinates $\mathbf{A}$.

The solution to the problem is, as in §16.5.8, to break general covariance by splitting spacetime into time and space coordinates, $x^\mu = \{t, x^\alpha\}$, and to interpret only those Hamilton's equations involving time t derivatives as genuine equations of motion, while the remaining equations are either constraint equations or identities.

In splitting a form $\mathbf{a}$ into time and space components, it is convenient to adopt a notation in which the form $\mathbf{a}_{\bar{t}}$ (subscripted $\bar{t}$) represents all the temporal parts of the form, while the form $\mathbf{a}_{\bar{\alpha}}$ (subscripted $\bar{\alpha}$) represents the remaining all-spatial components. The bars on the time and spatial indices $\bar{t}$ and $\bar{\alpha}$ serves

to distinguish the forms $\mathbf{a}_{\bar{t}} \equiv \mathbf{a}_{tA} dx^{tA}$ and $\mathbf{a}_{\bar{\alpha}} \equiv \mathbf{a}_{\alpha A} dx^{\alpha A}$ from their components $\mathbf{a}_{tA}$ and $\mathbf{a}_{\alpha A}$. Thus a 1-form $\mathbf{a} \equiv a_{\kappa} dx^{\kappa}$ splits into

$$\mathbf{a} = \mathbf{a}_{\bar{t}} + \mathbf{a}_{\bar{\alpha}} \equiv a_t dt + a_{\alpha} dx^{\alpha}, \quad (16.71)$$

while a 2-form $\mathbf{a} \equiv a_{\kappa\lambda} dx^{\kappa\lambda}$ splits into

$$\mathbf{a} = \mathbf{a}_{\bar{t}} + \mathbf{a}_{\bar{\alpha}} \equiv a_{t\alpha} dx^{t\alpha} + a_{\alpha\beta} dx^{\alpha\beta}, \quad (16.72)$$

implicitly summed over distinct sequences of indices. The time component of the exterior product of two forms $\mathbf{a}$ and $\mathbf{b}$ is

$$(\mathbf{a} \wedge \mathbf{b})_{\bar{t}} = \mathbf{a}_{\bar{t}} \wedge \mathbf{b}_{\bar{\alpha}} + \mathbf{a}_{\bar{\alpha}} \wedge \mathbf{b}_{\bar{t}}, \quad (16.73)$$

with no minus signs, the minus signs from the antisymmetry of indices cancelling the minus signs from commuting dt through a spatial form.

The electromagnetic field 2-form $\mathbf{F}$ splits as

$$\mathbf{F} = \mathbf{F}_{\bar{t}} + \mathbf{F}_{\bar{\alpha}} = F_{t\alpha} dx^{t\alpha} + F_{\beta\gamma} dx^{\beta\gamma}, \quad (16.74)$$

whose time and space parts encode the electric and magnetic fields. The dual electromagnetic field 2-form ${}^*\mathbf{F}$ splits as

$${}^*\mathbf{F} = {}^*\mathbf{F}_{\bar{t}} + {}^*\mathbf{F}_{\bar{\alpha}} = \varepsilon_{t\alpha\beta\gamma} F^{\beta\gamma} dx^{t\alpha} + \varepsilon_{t\alpha\beta\gamma} F^{t\alpha} dx^{\beta\gamma}, \quad (16.75)$$

whose time and space parts conversely encode the magnetic and electric fields. With the definitions $E^{\alpha} \equiv F^{t\alpha}$ and $B^{\alpha} \equiv \varepsilon^{t\alpha\beta\gamma} F_{\beta\gamma}$ of electric and magnetic field components, the form expression (16.75) agrees with the equivalent multivector expression (14.63).

The time components of Hamilton's equations (16.63) comprise 3 equations of motion for the 3 spatial components ${}^*\mathbf{F}_{\bar{\alpha}}$ of the momenta, which is the electric field, and 3 equations of motion for the 3 spatial components $\mathbf{A}_{\bar{\alpha}}$ of the coordinates,

$$3 \text{ equations of motion: } (d {}^*\mathbf{F})_{\bar{t}} \equiv d_t {}^*\mathbf{F}_{\bar{\alpha}} + d_{\alpha} {}^*\mathbf{F}_{\bar{t}} = 4\pi {}^*\mathbf{j}_{\bar{t}}, \quad (16.76a)$$

$$3 \text{ equations of motion: } (d \mathbf{A})_{\bar{t}} \equiv d_t \mathbf{A}_{\bar{\alpha}} + d_{\alpha} \mathbf{A}_{\bar{t}} = \mathbf{F}_{\bar{t}}. \quad (16.76b)$$

The exterior time and space derivatives here are the 1-forms $d_t = dt \partial / \partial t$, and $d_{\alpha} = dx^{\alpha} \partial / \partial x^{\alpha}$. Equations (16.76) are the same as equations (16.46), but in forms notation in place of index notation. In translating the forms equations (16.76) into indexed equations (16.46), note minus signs that come from commuting dt through a spatial form, for example $d_{\alpha} \mathbf{A}_{\bar{t}} = dx^{\alpha} \partial / \partial x^{\alpha} dt A_t = -\partial A_t / \partial x^{\alpha} dx^{\alpha}$. The remaining Hamilton's equations (16.63), those not involving any time derivatives, are

$$1 \text{ constraint: } d_{\alpha} {}^*\mathbf{F}_{\bar{\alpha}} = 4\pi {}^*\mathbf{j}_{\bar{\alpha}}, \quad (16.77a)$$

$$3 \text{ identities: } d_{\alpha} \mathbf{A}_{\bar{\alpha}} = \mathbf{F}_{\bar{\alpha}}. \quad (16.77b)$$

In accordance with the relativists' convention, an equation is a constraint if it must be arranged to be satisfied on the initial hypersurface t_i of constant time, but is guaranteed thereafter by some conservation

law. Equation (16.77a) is an example of such a constraint equation, in this case guaranteed by conservation electric charge. The 4-dimensional equation representing conservation of charge,

$$d(d^*F - 4\pi^*j) = -4\pi d^*j = 0, \quad (16.78)$$

becomes in a 3+1 split

$$d_t(d^*F - 4\pi^*j)_{\bar{\alpha}} + d_{\alpha}(d^*F - 4\pi^*j)_{\bar{t}} = 0. \quad (16.79)$$

The second term on the left hand side of equation (16.79) vanishes on the equation of motion (16.76a), so equation (16.79) reduces to

$$d_t(d^*F - 4\pi^*j)_{\bar{\alpha}} = 0. \quad (16.80)$$

If the spatial components $(d^*F - 4\pi^*j)_{\bar{\alpha}}$ are arranged to vanish on the initial spatial hypersurface of constant time, then the equation of motion (16.80) guarantees that those spatial components vanish thereafter. Provided, of course, that the equations governing the charged matter are arranged to satisfy charge conservation, as they should.

Equation (16.77b) on the other hand, which expresses the magnetic field $F_{\bar{\alpha}}$ as the spatial curl of the spatial potential $A_{\bar{\alpha}}$, is a constraint in Dirac's (1964) sense, but not in the relativists' sense, since it is not guaranteed by any conservation law. As in §16.5.8, this book follows the relativists' convention that a constraint is an equation whose ongoing satisfaction is guaranteed by a conservation law, a first-class constraint. Dirac's second-class constraint equations are called identities.

16.6.7 3+1 split of the variation of the electromagnetic action

The equations of motion (16.76) and constraint and identities (16.77) follow directly from splitting Hamilton's equations (16.46) into time and space parts; but they can also be derived more fundamentally from splitting the variation (16.62) of the action into time and space parts,

$$\begin{aligned} \delta S = & \left[\frac{1}{4\pi} \oint {}^*F \wedge \delta A \right]_{t_i}^{t_f} \\ & + \frac{1}{4\pi} \int_{t_i}^{t_f} -(d^*F - 4\pi^*j)_{\bar{t}} \wedge \delta A_{\bar{\alpha}} - (d^*F - 4\pi^*j)_{\bar{\alpha}} \wedge \delta A_{\bar{t}} + \delta^*F_{\bar{\alpha}} \wedge (dA - F)_{\bar{t}} + \delta^*F_{\bar{t}} \wedge (dA - F)_{\bar{\alpha}}. \end{aligned} \quad (16.81)$$

From this variation it can be seen that the equations of motion (16.76) arise from minimizing the action with respect to the 3 spatial coordinates $A_{\bar{\alpha}}$ and 3 spatial momenta ${}^*F_{\bar{\alpha}}$. The 1 constraint (16.77a) arises from minimizing the action with respect to the 1 time component $A_{\bar{t}}$ of the coordinates, and the 3 identities (16.77b) from minimizing with respect to the 3 time components ${}^*F_{\bar{t}}$ of the momenta. Now $A_{\bar{t}}$ is a gauge variable: it can be adjusted arbitrarily by an electromagnetic gauge transformation,

$$A_{\bar{t}} \rightarrow A_{\bar{t}} + d_t \theta. \quad (16.82)$$

Minimizing the action with respect to the gauge variable $A_{\bar{t}}$ yields the constraint equation (16.77a) that effectively expresses current conservation.

The mere fact that $\mathbf{A}_{\bar{t}}$ can be treated as a gauge variable does not mean that it *must* be treated as a gauge variable. Other gauge-fixing choices can be made; see §27.6 for further discussion of this issue.

The time components ${}^*\mathbf{F}_{\bar{t}}$ of the momenta constitute the magnetic field. The dual of ${}^*\mathbf{F}_{\bar{t}}$ constitutes the spatial components of $\mathbf{F}_{\bar{\alpha}}$. The magnetic field ${}^*\mathbf{F}_{\bar{t}}$, or equivalently its dual $\mathbf{F}_{\bar{\alpha}}$, is not a gauge field (that is, it cannot be adjusted by a gauge transformation), but rather an auxiliary field that arises when the electromagnetic field is treated as a generally covariant 4-dimensional object. Minimizing the action (16.81) with respect to the magnetic field ${}^*\mathbf{F}_{\bar{t}}$ determines its own components, the identities (16.77b).

16.6.8 Conventional electromagnetic Hamiltonian

The conventional Hamiltonian H is defined by

$$H \equiv \frac{1}{4\pi} {}^*\mathbf{F}_{\bar{\alpha}} \wedge d_t \mathbf{A}_{\bar{\alpha}} - L . \quad (16.83)$$

The combined electromagnetic and charged interaction Lagrangian (16.59) can be written

$$\begin{aligned} L = \frac{1}{4\pi} [& d_{\alpha} ({}^*\mathbf{F}_{\bar{\alpha}} \wedge \mathbf{A}_{\bar{t}}) + {}^*\mathbf{F}_{\bar{\alpha}} \wedge d_t \mathbf{A} \\ & - (d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j})_{\bar{\alpha}} \wedge \mathbf{A}_{\bar{t}} + {}^*\mathbf{F}_{\bar{t}} \wedge (d\mathbf{A} - \mathbf{F})_{\bar{\alpha}} - \frac{1}{2} {}^*\mathbf{F}_{\bar{\alpha}} \wedge \mathbf{F}_{\bar{t}} + \frac{1}{2} {}^*\mathbf{F}_{\bar{t}} \wedge \mathbf{F}_{\bar{\alpha}} + 4\pi {}^*\mathbf{j}_{\bar{t}} \wedge \mathbf{A}_{\bar{\alpha}}] . \end{aligned} \quad (16.84)$$

Dropping the total derivative term $d_{\alpha} ({}^*\mathbf{F}_{\bar{\alpha}} \wedge \mathbf{A}_{\bar{t}})$ from the Lagrangian (16.84), and inserting the rest into the defining equation (16.83) yields the conventional Hamiltonian

$$H = \frac{1}{4\pi} [(d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j})_{\bar{\alpha}} \wedge \mathbf{A}_{\bar{t}} - {}^*\mathbf{F}_{\bar{t}} \wedge (d\mathbf{A} - \mathbf{F})_{\bar{\alpha}} + \frac{1}{2} {}^*\mathbf{F}_{\bar{\alpha}} \wedge \mathbf{F}_{\bar{t}} - \frac{1}{2} {}^*\mathbf{F}_{\bar{t}} \wedge \mathbf{F}_{\bar{\alpha}} - 4\pi {}^*\mathbf{j}_{\bar{t}} \wedge \mathbf{A}_{\bar{\alpha}}] . \quad (16.85)$$

The first term in the Hamiltonian (16.85) is the constraint (16.77a) wedged with the gauge variable $\mathbf{A}_{\bar{t}}$, while the second term is the identity (16.77b) wedged with the auxiliary field ${}^*\mathbf{F}_{\bar{t}}$, the magnetic field. Both terms vanish on the equations of motion. The third and fourth terms $({}^*\mathbf{F}_{\bar{\alpha}} \wedge \mathbf{F}_{\bar{t}} - {}^*\mathbf{F}_{\bar{t}} \wedge \mathbf{F}_{\bar{\alpha}})/(8\pi)$ go over to $(E^2 + B^2)/(8\pi) d^4x$ in flat space, and comprise the energy density of the electromagnetic field. The final term $- \mathbf{j} \cdot \mathbf{A} d^4x$ is an interaction term.

The conventional Hamiltonian (16.85) is a function of spatial coordinates $\mathbf{A}_{\bar{\alpha}}$ and their conjugate spatial momenta ${}^*\mathbf{F}_{\bar{\alpha}}$, and also a function of the time components $\mathbf{A}_{\bar{t}}$ and ${}^*\mathbf{F}_{\bar{t}}$ of the coordinates and momenta. The spatial derivatives $d_{\bar{\alpha}} \mathbf{A}_{\bar{\alpha}}$ and $d_{\bar{\alpha}} {}^*\mathbf{F}_{\bar{\alpha}}$ in the conventional Hamiltonian are to be interpreted as functions of the coordinates and momenta, not as separate degrees of freedom. One should think of $\mathbf{A}_{\beta}(x^{\alpha})$ and ${}^*\mathbf{F}_{\beta\gamma}(x^{\alpha})$ as being infinite collections of fields indexed by the spatial position x^{α} ; the spatial derivatives of the fields are then effectively linear combinations of those fields.

Varying the conventional Hamiltonian (16.85) with respect to $\mathbf{A}_{\bar{\alpha}}$, $\mathbf{A}_{\bar{t}}$, ${}^*\mathbf{F}_{\bar{\alpha}}$, and ${}^*\mathbf{F}_{\bar{t}}$ recovers Hamilton's equations (16.76) and (16.77). In executing the variation, the terms involving the varied derivatives $\delta(d_{\alpha} \mathbf{A}_{\bar{\alpha}}) = d_{\alpha} \delta \mathbf{A}_{\bar{\alpha}}$ and $\delta(d_{\bar{\alpha}} {}^*\mathbf{F}_{\bar{\alpha}}) = d_{\bar{\alpha}} \delta {}^*\mathbf{F}_{\bar{\alpha}}$ can be integrated by parts.

16.7 Gravitational action

As shown by Hilbert (1915) contemporaneously with Einstein's discovery of the final, successful version of general relativity, Einstein's equations can be derived by the principle of least action applied to the action

$$S_g = \int L_g d^4x, \quad (16.86)$$

with scalar **Hilbert Lagrangian**

$$L_g \equiv \frac{1}{16\pi G} R, \quad (16.87)$$

where R is the Ricci scalar, and G is Newton's gravitational constant. The motivation for the Hilbert action (16.86) is that the Ricci scalar R is the only non-vanishing scalar that can be constructed linearly from the Riemann curvature tensor R_{klmn} .

Least action requires the Lagrangian to be written as a function of the “coordinates” and “velocities” of the gravitational field. The traditional approach, following Hilbert, is to take the coordinates to be the 10 components $g_{\mu\nu}$ of the metric tensor. The gravitational Lagrangian L_g is then a function not only of the coordinates $g_{\mu\nu}$ and their velocities $\partial g_{\mu\nu}/\partial x^\kappa$, but also of their second derivatives $\partial^2 g_{\mu\nu}/\partial x^\kappa \partial x^\lambda$. The presence of the second derivatives (“accelerations”) might seem problematic, but they can be removed into a surface term by integration by parts, leaving a Lagrangian that contains only first derivatives.

A modified approach, with a different choice of “coordinates” for the gravitational field, brings out the Hamiltonian structure of the Hilbert Lagrangian, and makes transparent the dependence of the Hilbert Lagrangian on the two distinct symmetries underlying general relativity, namely general coordinate transformations, and local Lorentz transformations. In terms of the Riemann tensor (11.77) (valid with or without torsion) written in a mixed coordinate-tetrad basis, the Hilbert Lagrangian (16.87) is (units $c = G = 1$)

$$L_g = \frac{1}{16\pi} e^{m\kappa} e^{n\lambda} R_{\kappa\lambda mn} = \frac{1}{16\pi} e^{m\kappa} e^{n\lambda} \left(\frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{mn\kappa}}{\partial x^\lambda} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} - \Gamma_{m\kappa}^p \Gamma_{pn\lambda} \right). \quad (16.88)$$

As usual in this book, greek (brown) indices are coordinate indices, while latin (black) indices are tetrad indices in a tetrad with prescribed constant metric γ_{mn} . If the tetrad is orthonormal, then the tetrad metric is Minkowski, $\gamma_{mn} = \eta_{mn}$, but any tetrad with constant metric γ_{mn} , such as Newman-Penrose, will do. The Lagrangian (16.88) manifests the dependence of the gravitational Lagrangian on coordinate transformations, encoded in the 16 components of the inverse vierbein $e^{m\kappa}$, and on Lorentz transformations, encoded in the 24 connections $\Gamma_{mn\kappa}$. The connections $\Gamma_{mn\kappa}$ form a coordinate vector (index κ) of generators of Lorentz transformations (antisymmetric indices mn), and they constitute the connection associated with a local gauge group of Lorentz transformations. The Lorentz connections $\Gamma_{mn\kappa}$ are sometimes called “spin connections” in the literature. In a local gauge theory such as electromagnetism or Yang-Mills, the connections $\Gamma_{mn\kappa}$ would be interpreted as the “coordinates” of the field.

The mixed coordinate-tetrad expression for the Riemann tensor $R_{\kappa\lambda mn}$ on the right hand side of equation (16.88) is not the same as the coordinate expression (2.112), despite the resemblance of the two expressions. There are 24 Lorentz connections $\Gamma_{mn\kappa}$, but 40 (without torsion, or 64 with torsion) coordinate

connections $\Gamma_{\mu\nu\kappa}$. It is possible — indeed, this is the traditional Hilbert approach — to work entirely with coordinate-frame expressions, the coordinate metric and the coordinate connections, without introducing tetrads. The advantage of the mixed coordinate-tetrad approach is that it makes manifest the fact that the Hilbert Lagrangian is invariant with respect to two distinct symmetries, coordinate transformations encoded in the tetrad, and local Lorentz transformations encoded in the Lorentz connections. Extremization of the Hilbert action with respect to the tetrad yields Einstein's equations, with source the energy-momentum of matter. Extremization of the Hilbert action with respect to the Lorentz connections yields expressions for those connections in terms of the tetrad and its derivatives, with source the spin angular-momentum of matter.

Whereas a purely coordinate approach to extremizing the Hilbert action is possible, a purely tetrad approach is not. In general relativity, tetrad axes $\gamma_m(x^\mu)$ are defined at each point x^μ of spacetime. The coordinates x^μ of the spacetime manifold provide the canvas upon which tetrads can be erected, and through which tetrads can be transported. It is possible to do without tetrads by working with coordinate tangent axes e_μ and the associated coordinate connections, but it is not possible to do without coordinates.

If the Lorentz connections $\Gamma_{mn\kappa}$ are taken to be the coordinates of the gravitational field, then the corresponding canonical momenta are (a factor of 8π is inserted for convenience; or one could use units where $8\pi G = 1$ in place of the units $G = 1$ adopted here)

$$e^{mn\kappa\lambda} \equiv \frac{8\pi \delta L_g}{\delta(\partial\Gamma_{mn\lambda}/\partial x^\kappa)} = \frac{1}{2}(e^{m\kappa}e^{n\lambda} - e^{m\lambda}e^{n\kappa}) . \quad (16.89)$$

The momentum tensor $e^{mn\kappa\lambda}$ is antisymmetric in mn and in $\kappa\lambda$, and as such apparently has $6 \times 6 = 36$ components, but the requirement that it be expressible in terms of the vierbein in accordance with the right hand side of equation (16.89) means that the momentum tensor has only 16 independent degrees of freedom. The approach followed below, §16.8, is to treat the 16 components of the vierbein $e^{m\kappa}$ as the independent degrees of freedom. (A possible approach, not followed here, is to work with the 36-component momentum tensor $e^{mn\kappa\lambda}$ instead of the 16-component vierbein, subjecting the momentum to the identities (constraints, in Dirac's terminology)

$$\varepsilon_{\kappa\lambda\mu\nu} e^{k\kappa\lambda} e^{mn\mu\nu} = \varepsilon^{klmn} , \quad (16.90)$$

which is a symmetric 6×6 matrix of conditions, or 21 conditions, except that the normalization of $\varepsilon_{\kappa\lambda\mu\nu} = -e[\kappa\lambda\mu\nu]$, where e is the vierbein determinant, is arbitrary, so equations (16.90) constitute a set of 20 distinct identities.)

The gravitational Lagrangian (16.88) can be written

$$L_g = \frac{1}{8\pi} e^{mn\kappa\lambda} \left(\frac{\partial\Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) . \quad (16.91)$$

The Lagrangian (16.91) is in (super)-Hamiltonian form $L_g = p^\kappa \partial_\kappa q - H_g$ with coordinates $q = \Gamma_{mn\lambda}$ and momenta $p^\kappa = e^{mn\kappa\lambda}/8\pi$,

$$L_g = \frac{1}{8\pi} e^{mn\kappa\lambda} \frac{\partial\Gamma_{mn\lambda}}{\partial x^\kappa} - H_g , \quad (16.92)$$

and (super-)Hamiltonian $H_g(\Gamma_{mn\lambda}, e^{mn\kappa\lambda})$

$$H_g = -\frac{1}{8\pi} e^{mn\kappa\lambda} \gamma^{pq} \Gamma_{pm\lambda} \Gamma_{qn\kappa} . \quad (16.93)$$

Since a coordinate curl is a torsion-free covariant curl, equation (2.72), the coordinate partial derivatives $\partial/\partial x^\kappa$ in the Lagrangian (16.91) or in the definition (16.89) of momenta could be replaced by torsion-free covariant derivatives $\hat{D}_\kappa$, as was done earlier in the case of the electromagnetic field, equation (16.35). The development below works with coordinate derivatives, but one could equally well choose to work with torsion-free covariant derivatives.

Concept question 16.2 What are the units of the gravitational action? **Answer.** The Riemann tensor and its contractions have units $1/\text{length}^2$. The gravitational action (16.86) has units $\text{length}^2/(8\pi G)$, which in fundamental units $c = \hbar = 1$ is dimensionless as it should be, the gravitational coupling parameter $8\pi G$ being equal to the reduced Planck length l_p squared; with c and $\hbar$ units restored, $8\pi G\hbar/c^3 = l_p^2$. The dimensional nature of the gravitational coupling $8\pi G$ is problematic for a quantum theory of gravity, where renormalizability requires dimensionless coupling parameters. Quantum gravity would seem to require the existence of a fundamental length scale, which might be of the order of the Planck length. In string theory, there is a fundamental length scale, the string scale.

16.7.1 The Lorentz connection is not a tetrad tensor, but any variation of it is

The Lorentz connection $\Gamma_{mn\lambda} \equiv e^l{}_\lambda \Gamma_{mnl}$ is a coordinate vector but not a tetrad tensor. Although the Lorentz connection is not a tetrad tensor, any variation of it with respect to an infinitesimal local Lorentz transformation of the tetrad is a tetrad tensor. Generators of Lorentz transformations are antisymmetric tetrad tensors, Exercise 11.2. Under a local Lorentz transformation generated by the infinitesimal antisymmetric tensor ϵ_{nm} , a tetrad vector a_n varies as

$$a_n \rightarrow a'_n = a_n + \delta a_n = a_n + \epsilon_n{}^m a_m . \quad (16.94)$$

The variation δa_n of the tetrad vector,

$$\delta a_n = \epsilon_n{}^m a_m , \quad (16.95)$$

is thus also a tetrad vector. The Lorentz connection is defined by $\Gamma_{mn\lambda} \equiv \gamma_m \cdot \partial \gamma_n / \partial x^\lambda$, equation (11.37). Its variation under an infinitesimal Lorentz transformation generated by the antisymmetric tensor ϵ_{nm} is

$$\begin{aligned} \delta \Gamma_{mn\lambda} &= \delta \left(\gamma_m \cdot \frac{\partial \gamma_n}{\partial x^\lambda} \right) = \gamma_m \cdot \frac{\partial (\epsilon_n{}^p \gamma_p)}{\partial x^\lambda} + \epsilon_n{}^p \gamma_p \cdot \frac{\partial \gamma_m}{\partial x^\lambda} = \frac{\partial \epsilon_{nm}}{\partial x^\lambda} + \epsilon_n{}^p \Gamma_{mp\lambda} + \epsilon_m{}^p \Gamma_{pn\lambda} \\ &= D_\lambda \epsilon_{nm} \quad \text{a coordinate and tetrad tensor} . \end{aligned} \quad (16.96)$$

Equation (16.96) shows that the variation $\delta \Gamma_{mn\lambda}$ is a covariant derivative of a tetrad tensor, therefore a coordinate and tetrad tensor. The variation of the Lorentz connection by a derivative under an infinitesimal

Lorentz transformation is analogous to the variation $\delta A_\lambda = \partial_\lambda \theta$ of the electromagnetic potential A_λ by the gradient of a scalar θ under a gauge transformation of an electromagnetic field.

As a corollary, it follows that although the Hamiltonian H_g , equation (16.93), is not a tetrad scalar, any variation of it with respect to an infinitesimal local Lorentz transformation is a scalar.

16.8 Variation of the gravitational action

The gravitational action S_g with the Lagrangian (16.91) is

$$S_g = \frac{1}{8\pi} \int e^{mn\kappa\lambda} \left(\frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) e d^4x^{0123}. \quad (16.97)$$

Equations of motion governing the 16 vierbein $e^{m\kappa}$ and the 24 Lorentz connections $\Gamma_{mn\kappa}$ are obtained by varying the action (16.97) with respect to these fields. As shown below, variation with respect to the vierbein $e^{m\kappa}$ yields Einstein's equations in vacuo, equation (16.105), while variation with respect to the Lorentz connections $\Gamma_{mn\kappa}$ recovers the torsion-free expression (11.54) for the tetrad-frame connections Γ_{mnk} , equation (16.110).

The approach of treating the vierbein and connections as independent fields to be varied is the Hamiltonian (as opposed to Lagrangian) approach. In the context of the Hilbert action, the Hamiltonian approach is commonly called the Palatini approach, after Palatini (1919), who first treated the 10 components of the coordinate metric $g_{\mu\nu}$ and the 40 coordinate connections $\Gamma_{\mu\nu\kappa}$ as independent fields.

Before the gravitational action is varied, the spacetime is a manifold equipped with coordinates x^μ , but there is no prior coordinate metric $g_{\mu\nu}$, since the metric is determined by the vierbein, which remain unspecified until determined by the variation itself. Therefore, in varying the action, it is necessary to take the coordinate volume element d^4x^{0123} , which is a pseudoscalar, as the primitive measure of volume. The scalar volume element d^4x is related to the pseudoscalar coordinate volume element by a factor of the determinant e of the vierbein, $d^4x = e d^4x^{0123}$, equation (15.88), and this determinant e must be varied when the vierbein are varied.

Varying the action (16.97) with respect to the vierbein $e^{m\kappa}$ and the Lorentz connections $\Gamma_{mn\kappa}$ yields

$$\delta S_g = \frac{1}{8\pi} \int \left[e^{mn\kappa\lambda} \frac{\partial \delta \Gamma_{mn\lambda}}{\partial x^\kappa} + e^{mn\kappa\lambda} \delta(\Gamma_{m\lambda}^p \Gamma_{pn\kappa}) + \left(\frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) e^{-1} \delta(e e^{mn\kappa\lambda}) \right] e d^4x^{0123}. \quad (16.98)$$

To arrive at Hamilton's equations, the first term of the integrand on the right hand side of equation (16.98) (the $p^\kappa \partial_\kappa(\delta q)$ term) must be integrated by parts, which is accomplished by

$$e^{mn\kappa\lambda} \frac{\partial \delta \Gamma_{mn\lambda}}{\partial x^\kappa} = \frac{e^{-1} \partial(e e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda})}{\partial x^\kappa} - \frac{e^{-1} \partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} \delta \Gamma_{mn\lambda}. \quad (16.99)$$

Since $e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda}$ is a coordinate tensor (and also a tetrad tensor, equation (16.96)), the first term on the right hand side of equation (16.99) is a torsion-free covariant divergence in accordance with equation (2.74)

(the $\circ$ atop $\mathring{D}_\kappa$ is a reminder that it is torsion-free),

$$\frac{e^{-1}\partial(e e^{mn\kappa\lambda}\delta\Gamma_{mn\lambda})}{\partial x^\kappa} = \mathring{D}_\kappa (e^{mn\kappa\lambda}\delta\Gamma_{mn\lambda}) , \quad (16.100)$$

and therefore integrates to a surface term in accordance with Gauss' theorem (15.102). The remaining terms in the integrand of equation (16.98) must be expressed in terms of the variations $\delta\Gamma_{pn\kappa}$ and $\delta e^{m\kappa}$ of the connections and vierbein. The second term in the integrand on the right hand side of equation (16.98) is

$$e^{mn\kappa\lambda}\delta(\Gamma_{m\lambda}^p\Gamma_{pn\kappa}) = 2e^{mn\kappa\lambda}\Gamma_{m\lambda}^p\delta\Gamma_{pn\kappa} . \quad (16.101)$$

The variation $\delta \ln e$ of the vierbein determinant e in may be written in terms of the variation $\delta e^{m\kappa}$ of the vierbein, equation (2.77),

$$\delta \ln e = -e_{m\kappa}\delta e^{m\kappa} . \quad (16.102)$$

The last term in the integrand on the right hand side of equation (16.98) is then

$$\left(\frac{\partial\Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p\Gamma_{pn\kappa}\right)e^{-1}\delta(e e^{mn\kappa\lambda}) = \frac{1}{2}R_{\kappa\lambda mn}e^{-1}\delta(e e^{m\kappa}e^{n\lambda}) = (R_{\kappa m} - \frac{1}{2}e_{m\kappa}R)\delta e^{m\kappa} = G_{km}e^k{}_\kappa\delta e^{m\kappa} , \quad (16.103)$$

where $G_{km} \equiv R_{km} - \frac{1}{2}\gamma_{km}R$ is the tetrad-frame Einstein tensor. The $\frac{1}{2}\gamma_{km}R$ part of the Einstein tensor comes from variation of the vierbein determinant, equation (16.102).

The substitutions (16.99)–(16.103) bring the variation (16.98) of the gravitational action to

$$\begin{aligned} 8\pi\delta S_g = & \oint e^{mn\kappa\lambda}\delta\Gamma_{mn\lambda}d^3x_\kappa \\ & + \int \left[\left(-\frac{e^{-1}\partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} + 2\Gamma_{p\kappa}^{[m}e^{n]p\kappa\lambda} \right) \delta\Gamma_{mn\lambda} + G_{\kappa m}\delta e^{m\kappa} \right] d^4x . \end{aligned} \quad (16.104)$$

The surface term vanishes provided that the connections $\Gamma_{mn\lambda}$ are held fixed on the boundary of integration, so that their variation $\delta\Gamma_{mn\lambda}$ vanishes on the boundary. Hamilton's equations follow from extremizing the remaining integral. Extremizing the action (16.104) with respect to the variation $\delta e^{m\kappa}$ of the vierbein yields Einstein's equations in vacuo,

$$G_{km} = 0 . \quad (16.105)$$

Extremizing the action with respect to the variation $\delta\Gamma_{mn\lambda}$ of the Lorentz connections gives

$$\frac{e^{-1}\partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} = 2\Gamma_{p\kappa}^{[m}e^{n]p\kappa\lambda} . \quad (16.106)$$

Abbreviate the left hand side of equation (16.106) by

$$f^{lmn} \equiv e^l{}_\lambda \frac{e^{-1}\partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} , \quad (16.107)$$

which is antisymmetric in its last two indices, $f_{lmn} = f_{l[mn]}$. In terms of the vierbein derivatives d_{lmn} defined by equation (11.33), the quantities f_{lmn} defined by equation (16.107) are

$$f_{lmn} = d_{l[mn]} - \gamma_{lm}d^k{}_{[kn]} + \gamma_{ln}d^k{}_{[km]} . \quad (16.108)$$

Inverting equation (16.106) yields the tetrad-frame connections Γ_{mnl} in terms of f_{lmn} ,

$$\Gamma_{mnl} = 2f_{lmn} - 3f_{[lmn]} + 2\gamma_{l[m}f_{n]p}^p. \quad (16.109)$$

Inserting the expression (16.108) into equations (16.109) yields the standard torsion-free expression (11.54) for the tetrad-frame connection Γ_{mnl} in terms of vierbein derivatives d_{mnl} ,

$$\Gamma_{mnl} = \dot{\Gamma}_{mnl} = 2d_{l[mn]} - 3d_{[lmn]}. \quad (16.110)$$

The expression for the Ricci scalar in the Hilbert Lagrangian (16.88) is valid with or without torsion, but extremization of the action in vacuo has yielded the torsion-free connection. There remains the possibility that torsion could be generated by matter, §16.11.

16.9 Trading coordinates and momenta

In the Hamiltonian approach, the coordinates and momenta appear on an equal footing. A Lagrangian in Hamiltonian form $L = p \partial q - H$ can be replaced by an alternative Lagrangian $L' = -q \partial p - H$ which differs from the original by a total derivative, $L' = L - \partial(pq)$, and thus yields identical equations of motion. The alternative Lagrangian L' is in Hamiltonian form with $q \rightarrow p$ and $p \rightarrow -q$.

Consider integrating the first term of the gravitational Lagrangian (16.91) by parts (this is essentially the same integration by parts as (16.99), but with the connection $\Gamma_{mn\lambda}$ itself instead of the varied connection $\delta\Gamma_{mn\lambda}$),

$$e^{mn\kappa\lambda} \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} = \frac{e^{-1} \partial(e^{mn\kappa\lambda} \Gamma_{mn\lambda})}{\partial x^\kappa} - \frac{e^{-1} \partial(e^{mn\kappa\lambda})}{\partial x^\kappa} \Gamma_{mn\lambda}. \quad (16.111)$$

Now $e^{mn\kappa\lambda} \Gamma_{mn\lambda}$ is a coordinate tensor but not a tetrad tensor. However, its variation with respect to any infinitesimal Lorentz transformation is a tetrad tensor, §16.7.1. Therefore the variation of the first term on the right hand side of equation (16.111) is a torsion-free covariant divergence $\dot{D}_\kappa \delta(e^{mn\kappa\lambda} \Gamma_{mn\lambda})$, which can be discarded from the Lagrangian without changing the equations of motion. The resulting alternative gravitational Lagrangian is

$$8\pi L'_g = -\frac{e^{-1} \partial(e^{mn\kappa\lambda})}{\partial x^\kappa} \Gamma_{mn\lambda} + e^{mn\kappa\lambda} \Gamma_{m\lambda}^p \Gamma_{pn\kappa}. \quad (16.112)$$

Again, this alternative Lagrangian is a coordinate scalar but not a tetrad scalar, but any variation of it is a tetrad scalar, so is a satisfactory Lagrangian.

In this alternative Lagrangian (16.112), the coordinates are the vierbein $e^{n\lambda}$, and the corresponding canonically conjugate momenta are

$$\pi_n{}^\kappa{}_\lambda \equiv \frac{8\pi \delta L'_g}{\partial(\partial e^{n\lambda}/\partial x^\kappa)} = e^{m\kappa} \pi_{nm\lambda}, \quad (16.113)$$

where $\pi_{nm\lambda}$ and $\Gamma_{nm\lambda}$ are related by

$$\pi_{nm\lambda} \equiv \Gamma_{nm\lambda} - e_{n\lambda} \Gamma_{mp}^p + e_{m\lambda} \Gamma_{np}^p, \quad \Gamma_{nm\lambda} = \pi_{nm\lambda} - \frac{1}{2} e_{n\lambda} \pi_{mp}^p + \frac{1}{2} e_{m\lambda} \pi_{np}^p. \quad (16.114)$$

Like the tetrad connection $\Gamma_{nm\lambda}$, the covariant momentum $\pi_{nm\lambda}$ is antisymmetric in its first two indices nm , and therefore has $6 \times 4 = 24$ independent components. The traces are related by $\pi_{mp}^p = -2\Gamma_{mp}^p$. The alternative Lagrangian (16.112) is in Hamiltonian form $L'_g = p^\kappa \partial_\kappa q - H_g$ with coordinates $q = e^{n\lambda}$ and momenta $p^\kappa = \pi_n{}^\kappa{}_\lambda / 8\pi$,

$$L'_g = \frac{1}{8\pi} \pi_n{}^\kappa{}_\lambda \frac{\partial e^{n\lambda}}{\partial x^\kappa} - H_g , \quad (16.115)$$

and the same (super-)Hamiltonian (16.93) as before.

Equations of motion come from varying the alternative action $\delta S'_g$ with respect to the coordinates $e^{m\kappa}$ and momenta $\pi_{mn\lambda}$. The coefficients of the variations $\delta e^{m\kappa}$ and $\delta \pi_{mn\lambda}$ are linear combinations of the coefficients of $\delta e^{m\kappa}$ and $\delta \Gamma_{mn\lambda}$ in the varied action of equation (16.104). The end result is the same equations of motion as before, equations (16.105) and (16.110). The only difference is that variation of the alternative action gives a revised surface term,

$$8\pi \delta S'_g = \oint \pi_{nm\lambda} \delta e^{n\lambda} d^3x^m + \int \text{as eq. (16.104)} . \quad (16.116)$$

The surface term vanishes provided that the vierbein $e^{n\lambda}$ is held fixed on the boundary.

16.10 Matter energy-momentum and the Einstein equations with matter

Einstein's equations in vacuo, equation (16.105), emerged from varying the gravitational action with respect to the vierbein. Einstein's equations including matter are obtained by including the variation of the matter action with respect to the vierbein. The variation of the matter action S_m with respect to the vierbein defines the energy-momentum tensor $T_{\kappa m}$ of matter,

$$\delta S_m = - \int T_{\kappa m} \delta e^{m\kappa} d^4x . \quad (16.117)$$

Adding the variation (16.104) of the gravitational action and the variation (16.117) of the matter action gives

$$8\pi (\delta S_g + \delta S_m) = \int (G_{\kappa m} - 8\pi T_{\kappa m}) \delta e^{m\kappa} d^4x , \quad (16.118)$$

extremization of which implies Einstein's equations in the presence of matter

$$\boxed{G_{\kappa m} = 8\pi T_{\kappa m}} . \quad (16.119)$$

The Einstein equations (16.119) constitute a set of 16 equations. Conditions on the energy-momentum imposed by the invariance of the matter action under local Lorentz transformations and under coordinate transformations are discussed in §§16.11.1 and 16.11.2 below.

If the matter action is $S_m = \int L_m d^4x$, then the matter energy-momentum is the sum of a part from the

variation of the matter Lagrangian L_m , and a part from the variation of the vierbein determinant e in the scalar volume element $d^4x \equiv e d^4x^{0123}$,

$$T_{\kappa m} = -\frac{\delta L_m}{\delta e^{m\kappa}} + L_m e_{m\kappa} . \quad (16.120)$$

16.11 Spin angular-momentum

In the standard $U(1)_Y \times SU(2)_L \times SU(3)_c$ model of physics, the connections associated with the gauge groups are dynamical fields, the gauge bosons, which include (after electroweak symmetry breaking) photons, weak gauge bosons, and gluons. As has been seen above, the gauge symmetries of general relativity include not only coordinate transformations, encoded in the vierbein $e^{m\kappa}$, but also Lorentz transformations, encoded in the Lorentz connection $\Gamma_{mn\lambda}$. Treating the vierbein as a dynamical field leads to Einstein's equations (16.119) and standard general relativity. If the Lorentz connection is treated similarly as a dynamical field, as it surely should be, then the inevitable consequence is the extension of general relativity to include torsion, which is called **Einstein-Cartan theory**.

Einstein-Cartan theory follows general relativity in taking the Lagrangian to be the Hilbert Lagrangian, the only difference being that the Lorentz connections $\Gamma_{mn\lambda}$ in the Riemann tensor are allowed to have torsion. The Riemann tensor with torsion equals the torsion-free Riemann tensor plus extra terms depending on the contortion, equation (15.49). Since torsion is a tensor, it is possible to include additional torsion-dependent terms in the Lagrangian (Hammond, 2002; Hehl, 2012; Blagojević and Hehl, 2013), but the various possible extensions go beyond the scope of this book.

As shown below, in Einstein-Cartan theory, torsion vanishes in empty space, and it does not propagate as a wave, unlike the (trace-free, Weyl part of the) Riemann curvature. Consequently conventional experimental tests of gravity do not rule out torsion. The gravitational force is intrinsically much weaker than the other three forces of the standard model. It makes itself felt only because gravity is long-ranged, and cumulative with mass. Since torsion in Einstein-Cartan theory is local, it is hard to detect.

Just as the variation of the matter action with respect to the vierbein $e^{m\kappa}$ defines the energy-momentum tensor T_{km} , so also the variation of the matter action with respect to the Lorentz connections $\Gamma_{mn\lambda}$ defines the spin angular-momentum tensor $\Sigma^{\lambda mn}$,

$$\delta S_m = \frac{1}{2} \int \Sigma^{\lambda mn} \delta \Gamma_{mn\lambda} d^4x \quad (16.121)$$

(implicitly summed over both indices m and n). The spin angular-momentum tensor $\Sigma^{\lambda mn}$ is so-called because it is sourced by the spin of fermionic fields such as Dirac fields, Exercise 16.7. The spin angular-momentum vanishes for gauge fields such as the electromagnetic field, Exercise 16.6. Like the torsion tensor $S_{\lambda mn}$, the spin angular-momentum $\Sigma_{\lambda mn}$ is antisymmetric in its last two indices mn . Adding the variation (16.104) of the gravitational action and the variation (16.121) of the matter action gives

$$8\pi (\delta S_g + \delta S_m) = \int \left(-\frac{e^{-1} \partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} + 2\Gamma_{p\kappa}^{[m} e^{n]p\kappa\lambda} + 4\pi \Sigma^{\lambda mn} \right) \delta \Gamma_{mn\lambda} d^4x , \quad (16.122)$$

extremization of which implies

$$\frac{e^{-1}\partial(e e^{mn\kappa\lambda})}{\partial x^\kappa} = 2\Gamma_{p\kappa}^{[m}e^{n]p\kappa\lambda} + 4\pi\Sigma^{\lambda mn} . \quad (16.123)$$

Inverting equation (16.123) along the lines of equations (16.106)–(16.110) recovers the usual expression (11.55) for the torsion-full tetrad connection $\Gamma_{mn\lambda}$ as a sum of the torsion-free connection $\mathring{\Gamma}_{mn\lambda}$ given by equation (16.110), and a contortion tensor $K_{mn\lambda}$,

$$\Gamma_{mn\lambda} = \mathring{\Gamma}_{mn\lambda} + K_{mn\lambda} , \quad (16.124)$$

with the contortion tensor K_{mnl} being related to the spin angular-momentum Σ_{lmn} by

$$K_{mnl} = 8\pi \left(-\Sigma_{lmn} + \frac{3}{2}\Sigma_{[lmn]} - \gamma_{[m}\Sigma^p_{n]p} \right) . \quad (16.125)$$

The contortion K_{mnl} is related to the torsion S_{mnl} by equations (11.56). Equation (16.125) implies that the torsion $S_{\lambda mn}$ is related to the spin angular-momentum $\Sigma_{\lambda mn}$ by

$$S_{mn}^\lambda = 8\pi \left(\Sigma_{mn}^\lambda + e_{[m}^\lambda \Sigma_{n]k}^k \right) . \quad (16.126)$$

Equation (16.126) inverts to

$$\boxed{S_{mn}^\lambda + 2e_{[m}^\lambda S_{n]k}^k = 8\pi\Sigma_{mn}^\lambda} . \quad (16.127)$$

Equation (16.127) relating the torsion to the spin angular-momentum is the analogue of Einstein's equations (16.119) relating the Einstein tensor to the matter energy-momentum. Whereas the Einstein equations (16.119) determine only 10 of the 20 components of the Riemann tensor (for vanishing torsion) leaving 10 components (the Weyl tensor) to describe tidal forces and gravitational waves, the torsion equations (16.127) determine all 24 components of the torsion tensor in terms of the 24 components of the spin angular-momentum. Thus, at least in this vanilla version of general relativity with torsion, torsion vanishes in empty space, and it cannot propagate as a wave.

An equivalent spin angular-momentum tensor $\tilde{\Sigma}^{\lambda mn}$ is obtained by varying the matter action with respect to $\pi_{mn\lambda}$ in place of $\Gamma_{mn\lambda}$,

$$\delta S_m = \frac{1}{2} \int \tilde{\Sigma}^{\lambda mn} \delta\pi_{mn\lambda} d^4x . \quad (16.128)$$

The relation between the torsion S_{mn}^λ and the modified spin angular-momentum $\tilde{\Sigma}_{mn}^\lambda$ is

$$S_{mn}^\lambda = 8\pi \tilde{\Sigma}_{mn}^\lambda . \quad (16.129)$$

Comparing equation (16.129) to equation (16.126) shows that the modified and original spin angular-momenta $\tilde{\Sigma}_{mn}^\lambda$ and Σ_{mn}^λ differ by a trace term,

$$\tilde{\Sigma}_{mn}^\lambda = \Sigma_{mn}^\lambda + e_{[m}^\lambda \Sigma_{n]k}^k , \quad \Sigma_{mn}^\lambda = \tilde{\Sigma}_{mn}^\lambda + 2e_{[m}^\lambda \tilde{\Sigma}_{n]k}^k . \quad (16.130)$$

As seen above, the torsion, contortion, and spin angular-momentum tensors are all invertibly related to each other. The relations between them are conceptually clearer when decomposed into irreducible parts. Each is

a 24-component tensor that decomposes into a 4-component trace part, a 4-component totally antisymmetric part, and a remaining 16-component trace-free antisymmetry-free part. The torsion S_{lmn} , contortion K_{lmn} , and spin angular-momentum Σ_{lmn} package these parts with different weights. The three parts are related by

$$S_{nk}^k = K_{nk}^k = -4\pi\Sigma_{nk}^k = 8\pi\tilde{\Sigma}_{nk}^k \quad \text{trace part ,} \quad (16.131a)$$

$$S_{[lmn]} = 2K_{[mnl]} = 8\pi\Sigma_{[lmn]} = 8\pi\tilde{\Sigma}_{[lmn]} \quad \text{totally antisymmetric part ,} \quad (16.131b)$$

$$S_{lmn} = -K_{mnl} = 8\pi\Sigma_{lmn} = 8\pi\tilde{\Sigma}_{lmn} \quad \text{trace-free, antisymmetry-free part .} \quad (16.131c)$$

16.11.1 Conservation of angular momentum and the symmetry of the energy-momentum tensor

The action S_m of any matter field is invariant under Lorentz transformations. Symmetry under Lorentz transformations implies a conservation law (16.136) of angular momentum. If torsion vanishes, the conservation law (16.136) implies that the energy-momentum tensor T^{mn} of the field is symmetric, equation (16.137). I thank Prof. Fred Hehl for pointing out that the antisymmetric part of the energy-momentum tensor can be interpreted consistently as half the divergence of orbital angular momentum, §19(c) of Corson (1953), so that equation (16.136) can be interpreted as a conservation law of total angular momentum, spin plus orbital.

Equation (16.95) gives the variation of a tetrad vector under a local Lorentz transformation generated by the infinitesimal antisymmetric tensor ϵ_{mn} . Under such an infinitesimal Lorentz transformation, the vierbein tensor $e^{m\kappa}$ varies as

$$\delta e^{m\kappa} = \epsilon^m_n e^{n\kappa} = \epsilon^{m\kappa} . \quad (16.132)$$

Equation (16.96) gives the variation of the Lorentz connection under an infinitesimal Lorentz transformation generated by ϵ_{mn} ,

$$\delta\Gamma_{mn\lambda} = -D_\lambda\epsilon_{mn} . \quad (16.133)$$

The coefficients of the variation δS_m of the matter action with respect to $\delta e^{m\kappa}$ and $\delta\Gamma_{mn\lambda}$ are by definition the energy-momentum and spin angular-momentum of the matter, equations (16.121) and (16.117). Inserting the variations (16.132) and (16.133) with respect to Lorentz transformations yields the variation of the matter action under a Lorentz transformation,

$$\delta S_m = - \int \left(\frac{1}{2} \Sigma^{\lambda mn} D_\lambda \epsilon_{mn} + T_{\kappa m} \epsilon^{m\kappa} \right) d^4x . \quad (16.134)$$

An integration by parts brings the variation to

$$\delta S_m = - \oint \frac{1}{2} \Sigma_\lambda{}^{mn} \epsilon_{mn} d^3x^\lambda + \int \left(\frac{1}{2} D_\lambda \Sigma^{\lambda mn} + T^{[mn]} \right) \epsilon_{mn} d^4x . \quad (16.135)$$

Requiring that the matter action be invariant under Lorentz transformations imposes that the variation (16.135) must vanish under arbitrary variations of the antisymmetric Lorentz generators ϵ_{mn} , subject to the

generators being fixed on the initial and final hypersurfaces of integration. Therefore the integrand of the rightmost integral in equation (16.135) must vanish, implying the conservation law

$$\left[\frac{1}{2} D_\lambda \Sigma^{\lambda mn} + T^{[mn]} = 0 \right]. \quad (16.136)$$

If the spin angular-momentum of the matter component vanishes, $\Sigma^{\lambda mn} = 0$, then the energy-momentum tensor of the matter component is symmetric,

$$T^{mn} = T^{nm}. \quad (16.137)$$

16.11.2 Conservation of energy-momentum

The action S_m of any matter field is also invariant under coordinate transformations. Symmetry under coordinate transformations implies a conservation law (16.146) for the energy-momentum T^{mn} of the field.

Under a coordinate transformation generated by the coordinate shift

$$\delta x^\mu = \epsilon^\mu, \quad (16.138)$$

the variation of any quantity is given by minus its Lie derivative $\mathcal{L}_\epsilon$ with respect to the coordinate shift ϵ^μ , equation (7.131). The Lie derivative of a coordinate tensor is given by equation (7.159), and this equation continues to hold for tensors that are tetrad as well as coordinate tensors, the tetrad components being treated as coordinate scalars (because tetrad components are unchanged under a coordinate transformation). However, a difficulty arises because the Lie derivative of a tetrad tensor is not a tetrad tensor (see Concept question 26.3). Consequently, although the vierbein is a coordinate and tetrad tensor, its Lie derivative is a coordinate tensor but not a tetrad tensor. The solution to the difficulty is pointed out at the beginning of §5.2.1 of Hehl et al. (1995): the Lagrangian is a Lorentz scalar, so its coordinate derivative is also its Lorentz-covariant derivative. Thus in varying the Lagrangian, the coordinate derivative of any tetrad tensor can be replaced by its Lorentz-covariant derivative. The Lorentz-covariant Lie derivative $\mathcal{L}_{\Gamma\epsilon}$ of the vierbein is

$$\begin{aligned} \mathcal{L}_{\Gamma\epsilon} e^{m\kappa} &= -e^{m\lambda} \frac{\partial \epsilon^\kappa}{\partial x^\lambda} + \epsilon^\lambda \left(\frac{\partial e^{m\kappa}}{\partial x^\lambda} + \Gamma_{n\lambda}^m e^{n\kappa} \right) \\ &= -\mathring{D}^m \epsilon^\kappa - \epsilon_\lambda K^{m\kappa\lambda} \\ &= -D^m \epsilon^\kappa - \epsilon_l S^{\kappa ml}, \end{aligned} \quad (16.139)$$

which differs from equation (26.18) in that the derivative of $e^{m\kappa}$ on the right hand side of the first line is covariant with respect to the tetrad index m . The expressions on the second and third lines of equations (16.139) are equivalent; the second line is in terms of the torsion-free covariant derivative $\mathring{D}$, while the third line is in terms of the torsion-full covariant derivative D . Thus the vierbein tensor $e^{m\kappa}$ varies under a coordinate transformation as, equation (26.18),

$$\delta e^{m\kappa} = -\mathcal{L}_{\Gamma\epsilon} e^{m\kappa} = \mathring{D}^m \epsilon^\kappa + \epsilon_\lambda K^{m\kappa\lambda}. \quad (16.140)$$

The Lorentz connection $\Gamma_{mn\lambda}$ is not a tetrad-frame tensor, so the usual formula for the Lie derivative

does not apply. Rather, the variation $\delta\Gamma_{mn\lambda}$ of the Lorentz connection follows from a difference of covariant derivatives,

$$\delta D_\lambda a_n - D_\lambda \delta a_n = \delta(\partial_\lambda a_n - \Gamma_{n\lambda}^m a_m) - (\partial_\lambda \delta a_n - \Gamma_{n\lambda}^m \delta a_m) = -(\delta\Gamma_{n\lambda}^m) a_m . \quad (16.141)$$

Thus the variation of the Lorentz connection under a coordinate transformation by ϵ^κ satisfies

$$\begin{aligned} (\delta\Gamma_{mn\lambda}) a^m &= \mathcal{L}_\Gamma \epsilon (D_\lambda a_n) - D_\lambda \mathcal{L}_\Gamma \epsilon a_n = \epsilon^\kappa \left(\frac{\partial}{\partial x^\kappa} D_\lambda a_n - \Gamma_{n\kappa}^m D_\lambda a_m \right) + (D_\kappa a_n) \frac{\partial \epsilon^\kappa}{\partial x^\lambda} - D_\lambda (\epsilon^\kappa D_\kappa a_n) \\ &= \epsilon^\kappa a^m R_{\lambda\kappa mn} . \end{aligned} \quad (16.142)$$

Equation (16.142) is true for arbitrary a^m , so

$$\delta\Gamma_{mn\lambda} = \epsilon^\kappa R_{\lambda\kappa mn} . \quad (16.143)$$

Inserting the variations (16.140) and (16.143) of the vierbein and Lorentz connection into the variations (16.117) and (16.121) of the matter action yields the variation of the matter action under a coordinate transformation by ϵ^κ ,

$$\delta S_m = \int \left[-T_{\kappa m} \left(\dot{D}^m \epsilon^\kappa + \epsilon_\lambda K^{m\kappa\lambda} \right) + \frac{1}{2} \Sigma^{\lambda mn} \epsilon^\kappa R_{\lambda\kappa mn} \right] d^4 x . \quad (16.144)$$

An integration by parts brings the variation of the matter action to

$$\delta S_m = - \oint T_{\kappa\lambda} \epsilon^\kappa d^3 x^\lambda + \int \left(\dot{D}^m T_{\kappa m} + T^{mn} K_{mn\kappa} + \frac{1}{2} \Sigma^{\lambda mn} R_{\lambda\kappa mn} \right) \epsilon^\kappa d^4 x . \quad (16.145)$$

Invariance of the action under coordinate transformations requires that the variation (16.145) vanish for arbitrary coordinate shifts ϵ^κ that vanish on the boundary. Therefore the integrand of the rightmost integral in equation (16.145) must vanish, implying the law of conservation of energy-momentum,

$$\boxed{\dot{D}^m T_{\kappa m} + T^{mn} K_{mn\kappa} + \frac{1}{2} \Sigma^{\lambda mn} R_{\lambda\kappa mn} = 0} . \quad (16.146)$$

Since the contortion $K_{mn\kappa}$ is antisymmetric in its first two indices mn , the second term of the conservation law (16.146) depends on the antisymmetric part $T^{[mn]}$ of the energy-momentum tensor.

Equation (16.146) splits into equations for the symmetric and antisymmetric parts of the energy-momentum of the matter component. The equation for the symmetric part of the energy-momentum is just

$$\dot{D}_m T^{(nm)} = 0 . \quad (16.147)$$

Concept question 16.3 Can the coordinate metric be Minkowski in the presence of torsion?

Can the coordinate metric be the Minkowski metric $g_{\mu\nu} = \eta_{\mu\nu}$ over a finite region of spacetime where torsion does not vanish? **Answer.** As discussed in Concept question 2.6, yes, torsion could technically be finite even in flat (Minkowski) space. In practice, no, because torsion at any point of spacetime is determined by the spin angular-momentum of matter there, which contributes energy-momentum that ensures that the metric is not Minkowski over the finite region (of course, the metric can always be made locally Minkowski).

Concept question 16.4 What kinds of metric or vierbein admit torsion? **Answer.** Any kind. Coordinate derivatives of the metric or vierbein determine torsion-free connections, placing no constraint on torsion.

Concept question 16.5 Why the names matter energy-momentum and spin angular-momentum? What is the justification for calling $T_{\kappa m}$ the matter energy-momentum and $\Sigma^{\lambda mn}$ the spin angular-momentum?

Answer. In flat spacetime, conservation of energy and momentum are associated with translation symmetry with respect to time and space. Conservation of angular momentum is associated with rotational symmetry of space. In general relativity, these global symmetries are replaced by local symmetries. Translation symmetry is replaced by symmetry under coordinate transformations; rotational symmetry is replaced by symmetry under local Lorentz transformations (which include Lorentz boosts as well as spatial rotations). The matter energy-momentum tensor $T_{\kappa m}$ satisfies a conservation law (16.146) that arises as a result of symmetry under coordinate transformations. The spin angular-momentum tensor $\Sigma^{\lambda mn}$ satisfies a conservation law (16.136) that arises as a result of symmetry under local Lorentz transformations. The reason for the adjective “spin” is that, as seen in Exercises 16.6 and 16.7, spin angular-momentum vanishes for bosonic fields such as electromagnetism, but is non-vanishing for fermionic (half-integral spin) fields.

Exercise 16.6 Energy-momentum and spin angular-momentum of the electromagnetic field. Derive the energy-momentum and spin angular-momentum of the electromagnetic field. The energy-momentum and spin angular-momentum of a field are defined by equations (16.117) and (16.121).

Solution.

1. **Energy-momentum of the electromagnetic field.** The Lagrangian of the electromagnetic field is, equation (16.28),

$$L \equiv -\frac{1}{16\pi} g^{\kappa\mu} g^{\lambda\nu} F_{\kappa\lambda} F_{\mu\nu} , \quad (16.148)$$

where the inverse metric is in terms of the vierbein,

$$g^{\kappa\mu} = \eta_{km} e^{k\kappa} e^{m\mu} . \quad (16.149)$$

The fact that the Lagrangian depends on the vierbein only in the symmetrized combination constituting the inverse metric guarantees that the energy-momentum tensor is symmetric. The variation of the electromagnetic Lagrangian (16.148) with respect to the vierbein is

$$\delta L = -\frac{1}{4\pi} F_{\kappa\lambda} F_k{}^\lambda \delta e^{k\kappa} . \quad (16.150)$$

An additional contribution to the energy-momentum comes from variation of the vierbein determinant in the volume element, equation (16.120). The resulting tetrad-frame energy-momentum tensor T_{kl} of the electromagnetic field is the symmetric tensor

$$T_{kl} = \frac{1}{4\pi} \left(F_{km} F_l{}^m - \frac{1}{4} \gamma_{kl} F_{mn} F^{mn} \right) . \quad (16.151)$$

The factor $1/4\pi$ factor is for Gaussian units, and is not present in Heaviside units.

2. **Spin angular-momentum of the electromagnetic field.** The Lagrangian of the electromagnetic field depends on the torsion-free curl of the electromagnetic potential, so does not involve any Lorentz connections. Therefore the spin angular-momentum of the electromagnetic field is zero,

$$\Sigma_{lmn} = 0 . \quad (16.152)$$

Exercise 16.7 Energy-momentum and spin angular-momentum of a Dirac field. Find the energy-momentum and spin angular-momentum of a Dirac spinor field.

Solution.

1. **Energy-momentum of a Dirac field.** The Lagrangian of a Dirac field is, equation (44.5),

$$L = \frac{1}{2} \bar{\psi} \cdot (e^{k\lambda} \gamma_k D_\lambda + m) \psi - \frac{1}{2} \psi \cdot (e^{k\lambda} \gamma_k D_\lambda + m) \bar{\psi} , \quad (16.153)$$

where the (torsion-full) covariant derivative is $D_\lambda = \partial_\lambda + \frac{1}{4} \Gamma_{mn\lambda} \gamma^m \wedge \gamma^n$ (implicit sum over both indices m and n). The two terms in the Lagrangian (16.153) are complex conjugates of each other, ensuring that the Lagrangian is real. Variation with respect to the vierbein $e^{k\lambda}$ yields the energy-momentum tensor $T_{lk} = e_l^\lambda T_{\lambda k}$, which is *not* symmetric in lk ,

$$T_{lk} = \frac{1}{2} \bar{\psi} \cdot \gamma_k D_l \psi - \frac{1}{2} \psi \cdot \gamma_k D_l \bar{\psi} . \quad (16.154)$$

The fact that energy-momentum tensor of a Dirac field is not symmetric is associated with the fact that the spin-angular momentum of the field does not vanish, §16.11.1. The Dirac Lagrangian L vanishes on the equations of motion, so the contribution to the energy-momentum, equation (16.120), arising from variation of the vierbein determinant in the scalar volume element $d^4x \equiv e d^4x^{0123}$ vanishes. Again, the two terms in the energy-momentum (16.154) are complex conjugates of each other, ensuring that the energy-momentum is real.

2. **Spin angular-momentum of a Dirac field.** Variation with respect to the connection $\Gamma_{mn\lambda}$ yields the spin angular-momentum $\Sigma_{lmn} \equiv e_l^\lambda \Sigma_{\lambda mn}$, which is a trivector current totally antisymmetric in lmn ,

$$\Sigma_{lmn} = \frac{1}{2} \bar{\psi} \cdot \gamma_l \wedge \gamma_m \wedge \gamma_n \psi . \quad (16.155)$$

The possible vector current contribution cancels between the two terms on the right hand side of equation (16.153).

Exercise 16.8 Electromagnetic field in the presence of torsion. Does torsion affect the propagation of the electromagnetic field?

Solution. No. The electromagnetic field equations involve only torsion-free derivatives, so the propagation of the electromagnetic field is unaffected by torsion.

Exercise 16.9 Dirac spinor field in the presence of torsion. How does torsion affect the propagation of a massive Dirac spin- $\frac{1}{2}$ field? Assume for simplicity that the background metric is Minkowski, that the spinor field is uniform (a plane wave) and at rest, and that the spin angular-momentum Σ_{mnk} is uniform.

Solution. The torsion-free part of the connection vanishes for a Minkowski metric, so the only non-vanishing

part of the connection is the contortion K_{mnk} . If the spin angular-momentum is uniform, then so is the contortion. The equation of motion of a Dirac spinor field of rest mass m is

$$[\boldsymbol{\gamma}^k(\partial_k + \frac{1}{4}K_{mnk}\boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n) + m]\psi = 0. \quad (16.156)$$

For simplicity, go to the rest frame of the spinor field, where the particle is in a time-up and spin-up eigenstate $\psi \propto \epsilon_{\uparrow\uparrow}$, equation (14.108), which means that the particle is a particle, not an antiparticle, and its spin is along the positive 3-direction. The only Dirac γ -matrices that are non-vanishing when acting on a spinor ψ in this state are $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$. Thus the equation of motion in the rest frame is

$$(\partial_0 + \frac{3}{2}K_{[012]} + K_{0a}^a + m)\psi = 0. \quad (16.157)$$

The solutions are

$$\psi \propto e^{-i(m+\delta m)t}, \quad (16.158)$$

where the mass change δm is

$$\delta m = \frac{3}{2}K_{[012]} + K_{0a}^a = 4\pi G(\Sigma_{[012]} - \Sigma_{0a}^a), \quad (16.159)$$

the contortion being related to the spin angular-momentum Σ_{mnk} by equations (16.131). Thus the effect of torsion is to change the effective mass m of the spinor particle. The trace part of the spin angular-momentum produces a mass change that has opposite signs for particles and antiparticles, but is independent of the direction of the spin of the particle, while the totally antisymmetric part of the spin angular-momentum produces a mass change that depends on the direction of the spin of the particle. As seen in Exercise 16.7, a Dirac spinor field produces only a totally antisymmetric spin angular-momentum $\Sigma_{[mnk]}$. This antisymmetric component is directional, so tends to cancel if the spins of the background system of spinor particles are pointed in random directions. The antisymmetric spin angular-momentum is significant only if the spins of the background particles are aligned. Whatever the case, since the gravitational coupling G is so weak compared to typical electromagnetic couplings, the resulting change in the mass of a spinor is typically tiny.

16.12 Lagrangian as opposed to Hamiltonian formulation

In the Lagrangian approach to the least action principle, as opposed to the Hamiltonian approach followed above, the Lagrangian is required to be a function of the coordinates and velocities, as opposed to the momenta. For gravity, the coordinates are the vierbein $e^{n\lambda}$, and the velocities are their coordinate derivatives $\partial e^{n\lambda}/\partial x^\kappa$. In the Lagrangian approach, the Lorentz connections $\Gamma_{mn\lambda}$ are not independent coordinates, but rather are taken to be given in terms of the coordinates and velocities $e^{n\lambda}$ and $\partial e^{n\lambda}/\partial x^\kappa$. In other words, the Lorentz connections are assumed to satisfy the equations of motion that in the Hamiltonian approach are derived by varying the action with respect to the connections.

The Hilbert Lagrangian depends not only on the vierbein and its first derivatives, but also on its second derivatives. To bring the Hilbert Lagrangian to a form that depends only on the first, not second, derivatives

of the vierbein, the Hilbert action must be integrated by parts. This is precisely the integration by parts that was carried out in the previous section §16.9. In the Lagrangian approach, the alternative Lagrangian L'_g given by equation (16.112) provides a satisfactory Lagrangian, once the connections $\Gamma_{mn\kappa}$ are expressed in terms of the vierbein $e^{m\kappa}$ and its first derivatives.

16.12.1 Quadratic gravitational Lagrangian

The derivative term on the right hand side of the expression (16.112) for the Lagrangian L'_g was previously determined by Hamilton's equations to be given by equation (16.106), in which the connection proved to be the torsion-free connection. Substituting equation (16.106) (with torsion-free connection $\overset{\circ}{\Gamma}_{p\kappa}^m$) brings the alternative Lagrangian (16.112) to

$$8\pi L'_g = e^{mn\kappa\lambda} \left(-2 \overset{\circ}{\Gamma}_{m\lambda}^p \Gamma_{pn\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) = e^{mn\kappa\lambda} \left(-\overset{\circ}{\Gamma}_{m\lambda}^p \overset{\circ}{\Gamma}_{pn\kappa} + K_{m\lambda}^p K_{pn\kappa} \right), \quad (16.160)$$

the last step of which follows from expanding the torsion-full connection as a sum of the torsion-free connection and the contortion tensor, $\Gamma_{mn\kappa} = \overset{\circ}{\Gamma}_{mn\kappa} + K_{mn\kappa}$, equation (11.55). The torsion-free connections $\overset{\circ}{\Gamma}_{pn\kappa} \equiv e^k_{\kappa} \overset{\circ}{\Gamma}_{pnk}$ here are given by expression (16.110) (same as equation (11.54)), which are functions of the vierbein, linear in its first derivatives. The Lagrangian (16.160) is quadratic in the torsion-free connections, and therefore quadratic in the first derivatives of the vierbein, but independent of any second derivatives.

If torsion vanishes, as general relativity assumes, then

$$8\pi L'_g = -e^{mn\kappa\lambda} \Gamma_{m\lambda}^p \Gamma_{pn\kappa}. \quad (16.161)$$

Thus, for vanishing torsion, the first (“surface”) term in the original alternative Lagrangian (16.112) equals minus twice the second (“quadratic”) term. Padmanabhan (2010) has termed this property of the Hilbert Lagrangian “holographic,” and has suggested that it points to profound consequences.

16.12.2 A quick way to derive the quadratic gravitational Lagrangian

There is a quick way to derive the quadratic gravitational Lagrangian (16.160) that seems like it should not work, but it does. Suppose, incorrectly, that the Lorentz connections $\Gamma_{mn\lambda}$ formed a coordinate and tetrad tensor. Then contracting the Riemann tensor would give the Ricci scalar in the form

$$R = 2 \overset{\circ}{D}_{\kappa} (e^{mn\kappa\lambda} \Gamma_{mn\lambda}) + 2 e^{mn\kappa\lambda} (-\overset{\circ}{\Gamma}_{m\lambda}^p \overset{\circ}{\Gamma}_{pn\kappa} + K_{m\lambda}^p K_{pn\kappa}). \quad (16.162)$$

Discarding the torsion-free covariant divergence recovers the quadratic gravitational Lagrangian (16.160). Why does this work? The answer is that, as discussed in §16.12, although $\Gamma_{mn\lambda}$ is not a tetrad tensor, it is a tetrad tensor with respect to infinitesimal tetrad transformations about the value that satisfies the equations of motion. In the Lagrangian formalism, the connections are assumed to satisfy their equations of motion. Since least action invokes only infinitesimal variations of the coordinates and tetrad, for the purposes of applying least action, the argument $e^{mn\kappa\lambda} \Gamma_{mn\lambda}$ of the covariant divergence can be treated as a tensor, and the covariant divergence thus discarded legitimately.

16.13 Gravitational action in multivector notation

The derivation of the gravitational equations of motion from the Hilbert action can be translated into multivector language. Translating into multivector language does not make calculations any easier, but, by removing some of the blizzard of indices, it makes the structure of the gravitational Lagrangian more manifest. The multivector approach followed in this section 16.13 is a stepping stone to the even more compact, abstract, and powerful notation of multivector-valued differential forms, dealt with starting from §16.14.

16.13.1 Multivector gravitational Lagrangian

In multivector notation, the Hilbert Lagrangian (16.88) is

$$L_g \equiv \frac{1}{16\pi} (e^\lambda \wedge e^\kappa) \cdot R_{\kappa\lambda} = \frac{1}{16\pi} (e^\lambda \wedge e^\kappa) \cdot \left(\frac{\partial \Gamma_\lambda}{\partial x^\kappa} - \frac{\partial \Gamma_\kappa}{\partial x^\lambda} + \frac{1}{2} [\Gamma_\kappa, \Gamma_\lambda] \right), \quad (16.163)$$

implicitly summed over both indices κ and λ . In equation (16.163), $e^\kappa = e_m{}^\kappa \gamma^m$ are the usual coordinate (co)tangent vectors, equation (11.6), and the bivectors Γ_κ and $R_{\kappa\lambda}$ are given by equations (15.20) and (15.25). The dot in equation (16.163) signifies the multivector dot product, equation (13.35), which here is a scalar product of bivectors. The order of $e^\lambda \wedge e^\kappa$ is flipped to cancel a minus sign from taking a scalar product of bivectors.

Applying the multivector triple-product relation (13.39) to the derivative term in the rightmost expression of equation (16.163) brings the Hilbert Lagrangian to

$$L_g = \frac{1}{16\pi} (2 e^\lambda \cdot (\partial \cdot \Gamma_\lambda) + \frac{1}{2} (e^\lambda \wedge e^\kappa) \cdot [\Gamma_\kappa, \Gamma_\lambda]) , \quad (16.164)$$

where $\partial \equiv e^\kappa \partial / \partial x^\kappa$. The form of the Lagrangian (16.164) indicates that the “velocities” corresponding to the “coordinates” Γ_λ are $\partial \cdot \Gamma_\lambda$. The Lagrangian (16.164) is in (super-)Hamiltonian form with bivector coordinates Γ_λ , vector velocities $\partial \cdot \Gamma_\lambda$, and vector momenta $e^\lambda / 8\pi$,

$$L_g = \frac{1}{8\pi} e^\lambda \cdot (\partial \cdot \Gamma_\lambda) - H_g , \quad (16.165)$$

and (super-)Hamiltonian $H_g(\Gamma_\lambda, e^\lambda)$ (compare (16.93))

$$H_g = -\frac{1}{32\pi} (e^\lambda \wedge e^\kappa) \cdot [\Gamma_\kappa, \Gamma_\lambda] . \quad (16.166)$$

Whereas in tensor notation the gravitational coordinates and momenta appeared to be objects of different types, with different numbers of indices, in multivector notation the coordinates and momenta are all multivectors, albeit of different grades. In multivector notation, the number of coordinates Γ_λ and momenta e^λ is the same, 4.

16.13.2 Variation of the multivector gravitational Lagrangian

In multivector notation, the fields to be varied in the gravitational Lagrangian are the Lorentz connection bivectors $\mathbf{\Gamma}_\lambda$ and the coordinate vectors e^κ . In multivector notation, when the fields are varied, it is the coefficients $\Gamma_{kl\lambda}$ and $e_k{}^\kappa$ that are varied, the tetrad basis vectors γ_k being considered fixed. Thus the variation $\delta\mathbf{\Gamma}_\lambda$ of the Lorentz connections is

$$\delta\mathbf{\Gamma}_\lambda \equiv \frac{1}{2}(\delta\Gamma_{kl\lambda})\gamma^k \wedge \gamma^l \quad (16.167)$$

(implicitly summed over all indices; the factor of $\frac{1}{2}$ would disappear if the sum were over distinct antisymmetric indices kl). The variation δe^κ of the coordinate vectors is

$$\delta e^\kappa \equiv (\delta e_k{}^\kappa)\gamma^k. \quad (16.168)$$

As remarked in §16.8, when the vierbein are varied, the variation of the determinant e of the vierbein that goes into the scalar volume element $d^4x = e d^4x^{0123}$ must be taken into account. The variation of the vierbein determinant is related to the variation δe^κ of the coordinate vectors by

$$\delta \ln e = -e^k{}_\kappa \delta e_k{}^\kappa = -e_\kappa \cdot \delta e^\kappa. \quad (16.169)$$

The variation of the gravitational action with the multivector Lagrangian (16.170) is

$$\delta S_g = \frac{1}{16\pi} \int \left[(e^\lambda \wedge e^\kappa) \cdot \left(2 \frac{\partial \delta \mathbf{\Gamma}_\lambda}{\partial x^\kappa} + \frac{1}{2} \delta[\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] \right) + e^{-1} \delta(e e^\lambda \wedge e^\kappa) \cdot \mathbf{R}_{\kappa\lambda} \right] d^4x. \quad (16.170)$$

The first term in the integrand of equation (16.170) integrates by parts to

$$(e^\lambda \wedge e^\kappa) \cdot \frac{\partial \delta \mathbf{\Gamma}_\lambda}{\partial x^\kappa} = \overset{\circ}{D}_\kappa ((e^\lambda \wedge e^\kappa) \cdot \delta \mathbf{\Gamma}_\lambda) - \frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} \cdot \delta \mathbf{\Gamma}_\lambda. \quad (16.171)$$

The second term in the integrand of equation (16.170) is

$$\frac{1}{2} (e^\lambda \wedge e^\kappa) \cdot \delta[\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] = (e^\lambda \wedge e^\kappa) \cdot [\mathbf{\Gamma}_\kappa, \delta \mathbf{\Gamma}_\lambda] = [e^\lambda \wedge e^\kappa, \mathbf{\Gamma}_\kappa] \cdot \delta \mathbf{\Gamma}_\lambda, \quad (16.172)$$

the last step of which follows from the multivector triple-product relation (13.39) and the fact that (half) the anticommutator of two bivectors is the bivector part of their geometric product. The third term in the integrand of equation (16.170) is

$$\begin{aligned} \mathbf{R}_{\kappa\lambda} \cdot e^{-1} \delta(e e^\lambda \wedge e^\kappa) &= 2 \mathbf{R}_{\kappa\lambda} \cdot (e^\lambda \wedge \delta e^\kappa) - (\mathbf{R}_{\kappa\lambda} \cdot (e^\lambda \wedge e^\kappa)) e_\mu \cdot \delta e^\mu \\ &= (2 \mathbf{R}_{\kappa\lambda} \cdot e^\lambda - R e_\kappa) \cdot \delta e^\kappa \\ &= 2 \mathbf{G}_\kappa \cdot \delta e^\kappa, \end{aligned} \quad (16.173)$$

where the second line again follows from the multivector triple-product relation (13.39), and $\mathbf{G}_\kappa$ is the Einstein vector

$$\mathbf{G}_\kappa \equiv \mathbf{R}_{\kappa\lambda} \cdot e^\lambda - \frac{1}{2} R e_\kappa = (R_{\kappa m} - \frac{1}{2} R e_{m\kappa}) \gamma^m. \quad (16.174)$$

The manipulations (16.171)–(16.173) bring the variation (16.170) of the action to

$$\delta S_g = \frac{1}{8\pi} \oint (e_\lambda \wedge e_\kappa) \cdot \delta \Gamma^\lambda d^3 x^\kappa + \frac{1}{8\pi} \int \left[\left(-\frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} + \frac{1}{2} [e^\lambda \wedge e^\kappa, \Gamma_\kappa] \right) \cdot \delta \Gamma_\lambda + \mathbf{G}_\kappa \cdot \delta e^\kappa \right] d^4 x . \quad (16.175)$$

The surface term vanishes provided that Γ^λ is held fixed on the boundaries of integration. Extremizing the action (16.175) with respect to the variation δe^κ of coordinate vectors yields the Einstein equations in vacuo,

$$\mathbf{G}_\kappa = 0 . \quad (16.176)$$

Extremizing the action (16.175) with respect to the variation $\delta \Gamma_\lambda$ of the Lorentz connections yields the multivector equivalent of equation (16.106),

$$\frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} = \frac{1}{2} [e^\lambda \wedge e^\kappa, \Gamma_\kappa] . \quad (16.177)$$

The left hand side of equation (16.177) is

$$\frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} = \partial \wedge e^\lambda - e^\lambda \wedge (\partial \wedge e^\mu) . \quad (16.178)$$

The “velocities” of the coordinate vectors are their curls $\partial \wedge e^\lambda$,

$$\partial \wedge e^\lambda \equiv e^\kappa \wedge \frac{\partial e^\lambda}{\partial x^\kappa} = d_{[mn]}^\lambda \gamma^m \wedge \gamma^n \quad (16.179)$$

implicitly summed over both indices m and n . The $d_{mn}^\lambda \equiv e^{l\lambda} d_{lmn}$ in equation (16.179) are the vierbein derivatives defined by equation (11.33). Equation (16.177) solves to yield the torsion-free relation between the connections Γ^λ and the velocities $\partial \wedge e^\lambda$ of the coordinate vectors,

$$\Gamma^\lambda = \mathring{\Gamma}^\lambda = \partial \wedge e^\lambda - e^\lambda \cdot (e_\mu \wedge (\partial \wedge e^\mu)) , \quad \partial \wedge e^\lambda = \mathring{\Gamma}^\lambda - 2 e^\lambda \cdot (e_\mu \wedge \mathring{\Gamma}^\mu) . \quad (16.180)$$

16.13.3 Alternative multivector gravitational action

As in §16.9, since the Lagrangian (16.163) is in Hamiltonian form, the coordinates Γ_λ and momenta e^λ can be traded without changing the equations of motion. Integrating the Lagrangian (16.163) by parts gives

$$8\pi L_g = \frac{e^{-1} \partial((e^\lambda \wedge e^\kappa) \cdot \Gamma_\lambda)}{\partial x^\kappa} - \frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} \cdot \Gamma_\lambda + \frac{1}{4} (e^\lambda \wedge e^\kappa) \cdot [\Gamma_\kappa, \Gamma_\lambda] . \quad (16.181)$$

As in §16.9, the connection Γ_λ is not a tetrad tensor, but any infinitesimal variation of it is, §16.7.1, so the variation of the first term on the right hand side of equation (16.181) is a covariant divergence $\mathring{D}_\kappa \delta((e^\lambda \wedge e^\kappa) \cdot \Gamma_\lambda)$, which can be discarded from the Lagrangian without changing the equations of motion.

The middle term on the right hand side of equation (16.181) can be written

$$-\Gamma_\lambda \cdot \frac{e^{-1} \partial(e e^\lambda \wedge e^\kappa)}{\partial x^\kappa} = \pi_\lambda \cdot (\partial \wedge e^\lambda) , \quad (16.182)$$

where $\boldsymbol{\partial} \wedge \mathbf{e}^\lambda$ is given by equation (16.179), and $\boldsymbol{\pi}_\lambda$ is the trace-modified Lorentz connection bivector

$$\boldsymbol{\pi}_\lambda = \boldsymbol{\Gamma}_\lambda - \mathbf{e}_\lambda \wedge (\mathbf{e}^\mu \cdot \boldsymbol{\Gamma}_\mu) , \quad \boldsymbol{\Gamma}_\lambda = \boldsymbol{\pi}_\lambda - \frac{1}{2} \mathbf{e}_\lambda \wedge (\mathbf{e}^\mu \cdot \boldsymbol{\pi}_\mu) , \quad (16.183)$$

with components

$$\boldsymbol{\pi}_\lambda = \frac{1}{2} \pi_{mn\lambda} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n . \quad (16.184)$$

The components $\pi_{mn\lambda}$ are as given by equation (16.114).

Discarding the torsion-free divergence from the Lagrangian (16.181) yields the alternative Lagrangian

$$L'_g = \frac{1}{8\pi} \boldsymbol{\pi}_\lambda \cdot (\boldsymbol{\partial} \wedge \mathbf{e}^\lambda) - H_g , \quad (16.185)$$

with the same (super-)Hamiltonian (16.166) as before. The alternative Lagrangian (16.185) is in Hamiltonian form with coordinates $\mathbf{e}^\lambda$, velocities $\boldsymbol{\partial} \wedge \mathbf{e}^\lambda$, and corresponding canonically conjugate momenta $\boldsymbol{\pi}_\lambda/(8\pi)$. As with the alternative Lagrangian (16.112) in index notation, the alternative Lagrangian (16.185) in multivector notation is not a tetrad scalar because the Lorentz connection is not a tetrad tensor, but any infinitesimal variation of it is a (coordinate and) tetrad tensor, so the alternative Lagrangian (16.185) is satisfactory despite not being a tetrad scalar.

16.13.4 Einstein equations with matter, in multivector notation

In multivector notation, Einstein's equations including matter are obtained by including the variation of the matter action with respect to the variation $\delta \mathbf{e}^\kappa$ of the coordinate vectors. The variation defines the matter energy-momentum vector $\mathbf{T}_\kappa$,

$$\delta S_m = - \int \mathbf{T}_\kappa \cdot \delta \mathbf{e}^\kappa d^4x , \quad (16.186)$$

with

$$\mathbf{T}_\kappa = T_{\kappa m} \boldsymbol{\gamma}^m . \quad (16.187)$$

The combined variation of the gravitational and matter actions with respect to $\delta \mathbf{e}^\kappa$ is

$$8\pi(\delta S_g + \delta S_m) = \int (\mathbf{G}_\kappa - 8\pi \mathbf{T}_\kappa) \cdot \delta \mathbf{e}^\kappa d^4x , \quad (16.188)$$

extremization of which yields Einstein's equations with matter,

$$\mathbf{G}_\kappa = 8\pi \mathbf{T}_\kappa . \quad (16.189)$$

16.13.5 Spin angular-momentum in multivector notation

Just as the variation of the matter action with respect to the the variation $\delta \mathbf{e}^\kappa$ of the coordinate vectors defines the matter energy-momentum vector $\mathbf{T}_\kappa$, so also the variation of the matter action with respect to

the variation $\delta\Gamma_\lambda$ of the Lorentz connection bivectors defines the spin angular-momentum bivector Σ^λ ,

$$\delta S_m = \int \Sigma^\lambda \cdot \delta\Gamma_\lambda d^4x , \quad (16.190)$$

with (the minus sign is introduced for the same reason as the minus in equation (15.27))

$$\Sigma^\lambda \equiv -\frac{1}{2} \Sigma_{mn}^\lambda \gamma^m \wedge \gamma^n , \quad (16.191)$$

implicitly summed over both indices m and n . As in §16.11, the usual expression (15.46) for the torsion-full tetrad connections Γ_λ as a sum of the torsion-free connection $\mathring{\Gamma}_\lambda$ and the contortion K_λ is recovered,

$$\Gamma_\lambda = \mathring{\Gamma}_\lambda + K_\lambda , \quad (16.192)$$

provided that the torsion bivector $S^\lambda \equiv -\frac{1}{2} S_{mn}^\lambda \gamma^m \wedge \gamma^n$ is related to the spin angular-momentum bivector Σ^λ by

$$S^\lambda = 8\pi (\Sigma^\lambda - \frac{1}{2} e^\lambda \wedge (e_\mu \cdot \Sigma^\mu)) , \quad S^\lambda - e^\lambda \wedge (e_\mu \cdot S^\mu) = 8\pi \Sigma^\lambda . \quad (16.193)$$

16.14 Gravitational action in multivector forms notation

Especially in the mathematical literature, actions are often written in the even more compact notation of differential forms. The reward, if you can get over the language barrier, is a succinct picture of the structure of the gravitational action and equations of motion. For example, forms notation facilitates the intricate problem of executing a satisfactory 3+1 split of the gravitational equations, §16.15. If you aspire to a deeper understanding of numerical relativity or of quantum gravity, you would do well to understand forms.

As seen in §16.7, the Hilbert action is most insightful when the local Lorentz symmetry of general relativity, encoded in the tetrad γ_m , is kept distinct from the symmetry with respect to coordinate transformations, encoded in the tangent vectors e_μ . The distinction can be retained in forms language by considering multivector-valued forms. Local Lorentz transformations transform the multivectors while keeping the forms unchanged, while coordinate transformations transform the forms while keeping the multivectors unchanged.

To avoid conflict between multivector and form notations, it is convenient to reserve the wedge sign $\wedge$ to signify a wedge product of multivectors, not of forms. No ambiguity results from omitting the wedge sign for forms, since there is only one way to multiply forms, the exterior product². Similarly, it is convenient to reserve the Hodge duality symbol $*$ to signify the dual of a form, equation (15.79), not the dual of a multivector, and to write $I\mathbf{a}$ for the Hodge dual of a multivector $\mathbf{a}$, equation (13.24). The form dual of a p -form $\mathbf{a} = \mathbf{a}_\Lambda d^p x^\Lambda$ with multivector coefficients $\mathbf{a}_\Lambda$ is the multivector q -form $*\mathbf{a}$ given by (this is equation (15.79) generalized to allow multivector coefficients)

$$*\mathbf{a} \equiv (*\mathbf{a})_\Pi d^q x^\Pi = (-)^{pq} \mathbf{a}_\Lambda *d^q x^\Lambda = \varepsilon_{\Pi\Lambda} \mathbf{a}^\Lambda d^q x^\Pi , \quad (16.194)$$

² This is not true. The entire apparatus of multivectors can be translated into forms language. However, I take the point of view that, since multivectors are easier to manipulate than forms, there is not much to be gained from such a translation. The only occasion I find that necessitates introducing a dot product of forms is in deriving the law of conservation of energy-momentum in multivector forms language, equation (16.301).

implicitly summed over distinct sequences Λ and Π of respectively p and $q \equiv N - p$ (in N dimensional spacetime) coordinate indices. The dual (16.194) is a form dual, not a multivector dual. If $\mathbf{a}$ is a multivector of grade n (not necessarily equal to p or q), the dual form $^*\mathbf{a}$ remains a multivector of the same grade n . The double dual of a multivector form $\mathbf{a}$, both a multivector dual and a form dual, crops up often enough to merit its own notation, a double-asterisk overscript ** ,

$$^{**}\mathbf{a} \equiv I^* \mathbf{a} . \quad (16.195)$$

In this section 16.14 and in the remainder of this Chapter, unless otherwise stated, implicit sums are over distinct antisymmetric sequences of indices, since this removes the ubiquitous factorial factors that otherwise appear. For example, the wedge product of two multivectors $\mathbf{a}$ and $\mathbf{b}$ is

$$\mathbf{a} \wedge \mathbf{b} = (a_k \boldsymbol{\gamma}^k) \wedge (b_l \boldsymbol{\gamma}^l) = 2 a_{[k} b_{l]} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l = 2 a_k b_l \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l , \quad (16.196)$$

implicitly summed over distinct antisymmetric pairs of indices kl . In any expression for a multivector form in components, it can be helpful to think of the multivector and form indices as each carrying an implicit antisymmetrization symbol [...], as in the example $a_{[k} b_{l]}$ of equation (16.196). The antisymmetrization symbol will usually not be made explicit, both for brevity and to avoid a certain awkwardness of notation.

It is convenient to adopt the convention that the commutator of a multivector p -form $\mathbf{a}$ with a multivector q -form $\mathbf{b}$ is commuting if p and q are both odd, anticommuting otherwise,

$$[\mathbf{a}, \mathbf{b}] = \begin{cases} [\mathbf{b}, \mathbf{a}] & p \text{ and } q \text{ odd} , \\ -[\mathbf{b}, \mathbf{a}] & \text{otherwise} , \end{cases} \quad (16.197)$$

the anticommutation of the multivectors being deemed to cancel the anticommutation of forms when p and q are both odd. The advantage of this convention is that the contribution of the Lorentz connection to the covariant derivative of any multivector form $\mathbf{a}$ is always the commutator $\frac{1}{2}[\Gamma, \mathbf{a}]$. For example, the expression (15.26) for the Riemann tensor in terms of the commutator of the covariant derivative carries through to the language of multivector-valued forms, equation (16.210). As an example of the case where p and q are both odd, if $\mathbf{a}$ and $\mathbf{b}$ are two 1-forms, then their commutator is

$$[\mathbf{a}, \mathbf{b}] = [a_\kappa, b_\lambda] d^2 x^{\kappa\lambda} = (a_\kappa b_\lambda - b_\lambda a_\kappa) d^2 x^{\kappa\lambda} = (a_\kappa b_\lambda + b_\kappa a_\lambda) d^2 x^{\kappa\lambda} = [\mathbf{b}, \mathbf{a}] , \quad (16.198)$$

implicitly summed over distinct antisymmetric indices $\kappa\lambda$. As a corollary, the (anti-)commutator of a p -form $\mathbf{a}$ with itself vanishes if p is (odd) even,

$$\{\mathbf{a}, \mathbf{a}\} = 0 \quad p \text{ odd} , \quad (16.199a)$$

$$[\mathbf{a}, \mathbf{a}] = 0 \quad p \text{ even} . \quad (16.199b)$$

The exterior derivative d of a product of a multivector p -form $\mathbf{a}$ and a multivector q -form $\mathbf{b}$ satisfies the usual Leibniz-like rule for the exterior derivative d , equation (15.71),

$$d(\mathbf{a}\mathbf{b}) = (d\mathbf{a})\mathbf{b} + (-)^p \mathbf{a}(d\mathbf{b}) . \quad (16.200)$$

Exercise 16.10 Commutation of multivector forms.

1. Argue that if $\mathbf{a} \equiv \mathbf{a}_{K\Lambda} \gamma^K dx^\Lambda$ is a multivector form of grade k and form index p , and $\mathbf{b} \equiv \mathbf{a}_{K\Lambda} \gamma^K dx^\Lambda$ is a multivector form of grade l and form index q , then the grade $k+l-2n$ component of their product $\mathbf{ab}$ commutes or anticommutes as

$$\langle \mathbf{ab} \rangle_{k+l-2n} = (-)^{kl-n+pq} \langle \mathbf{ba} \rangle_{k+l-2n} . \quad (16.201)$$

As particular cases of equation (16.201), conclude that

$$\mathbf{a} \cdot \mathbf{a} = 0 \quad p \text{ odd} , \quad (16.202a)$$

$$\mathbf{a} \wedge \mathbf{a} = 0 \quad k+p \text{ odd} . \quad (16.202b)$$

2. What is the form index of the product $\mathbf{ab}$ of multivector forms $\mathbf{a}$ and $\mathbf{b}$ of form index p and q ?

Solution.

1. This is a combination of equations (13.28) and (15.61).
2. A product of forms is always their exterior product, so the form index of the product $\mathbf{ab}$ is $p+q$.

16.14.1 Interval, connection

In multivector notation, the gravitational coordinates and momenta proved to be $\mathbf{\Gamma}_\kappa$ and $\mathbf{e}^\kappa$ (or vice versa). In forms notation, the corresponding coordinates and momenta are the Lorentz connection bivector 1-form $\mathbf{\Gamma}$ and the line interval vector 1-form $\mathbf{e}$ defined by

$$\boxed{\mathbf{\Gamma} \equiv \mathbf{\Gamma}_\kappa dx^\kappa = \Gamma_{kl\kappa} \gamma^k \wedge \gamma^l dx^\kappa} , \quad (16.203a)$$

$$\boxed{\mathbf{e} \equiv \mathbf{e}_\kappa dx^\kappa = e_{k\kappa} \gamma^k dx^\kappa} , \quad (16.203b)$$

with, for $\mathbf{\Gamma}$, implicit summation over distinct antisymmetric sets of indices kl . The Lorentz connection 1-form $\mathbf{\Gamma}$ and coordinate interval 1-form $\mathbf{e}$ are abstract coordinate and tetrad gauge-invariant objects, whose components in any coordinate and tetrad frame constitute the Lorentz connection $\Gamma_{kl\kappa}$ and the vierbein $e_{k\kappa}$ in the mixed coordinate-tetrad basis.

The line interval $\mathbf{e}$ is essentially the same as the object $\mathbf{dx}$ first introduced in this book in equation (2.19). I contemplated using the symbol $\mathbf{dx}$ in place of $\mathbf{e}$ everywhere in this Chapter, to emphasize that using forms language does not require switching to a whole new set of symbols. But $\mathbf{e}$ is the symbol for the line-interval form conventionally used in the literature; and the symbol $\mathbf{dx}$ risks being misinterpreted as a composition of $\mathbf{d}$ and $\mathbf{x}$ (for example, an exterior derivative of $\mathbf{x}$), as opposed to the single holistic object $\mathbf{dx}$ that it really is. Moreover, if the dot product $\mathbf{e} \cdot \mathbf{e}$ is defined (as here) to be a form, then that dot product is not the same as the scalar spacetime interval squared $ds^2 = \mathbf{dx} \cdot \mathbf{dx}$, equation (2.25) (see Concept question 16.11).

It is convenient to use the symbol $\mathbf{e}^p$ to denote the normalized p -volume element,

$$\mathbf{e}^p \equiv \frac{1}{p!} \overbrace{\mathbf{e} \wedge \dots \wedge \mathbf{e}}^{p \text{ terms}} , \quad (16.204)$$

which is both a p -form and a grade- p multivector. The factor of $1/p!$ compensates for the multiple counting of distinct indices, and ensures that e^p correctly measures the p -volume element.

Concept question 16.11 Scalar product of the interval form e . In Chapter 2, the scalar product of the line interval with itself defined its scalar length squared, $ds^2 = d\mathbf{x} \cdot d\mathbf{x} = g_{\mu\nu} dx^\mu dx^\nu$, equation (2.25). Is this still true in multivector forms language? **Answer.** No. A differential p -form represents physically a p -volume element, and as such is always a sum of antisymmetrized products of p intervals. The scalar product of the interval 1-form e with itself is

$$e \cdot e = 2g_{\mu\nu} d^2x^{\mu\nu} = 0 \quad (16.205)$$

(implicitly summed over distinct antisymmetric sequences $\mu\nu$, hence the factor 2). The scalar product vanishes because of the symmetry of the metric $g_{\mu\nu}$ and the antisymmetry of the area element $d^2x^{\mu\nu}$.

A different version of a dot product of forms (not much used in this book) can be defined in precise analogy to a dot product of multivectors to yield a form of smaller form index, equation (16.286). This form dot product of the interval 1-form e with itself yields the 0-form

$$e \cdot e = e_m{}^\nu e_{n\nu} \gamma^m \gamma^n = \eta_{mn} \gamma^m \gamma^n = 4, \quad (16.206)$$

which again differs from the scalar product $ds^2 = d\mathbf{x} \cdot d\mathbf{x}$.

16.14.2 Curvature and torsion forms

The Riemann bivector 2-form R is defined by

$$R \equiv R_{\kappa\lambda} d^2x^{\kappa\lambda} = R_{\kappa\lambda mn} \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda}, \quad (16.207)$$

again implicitly summed over distinct antisymmetric indices mn and $\kappa\lambda$. The exterior derivative of the Lorentz connection 1-form is, equation (15.70), the 2-form

$$d\mathbf{\Gamma} = \left(\frac{\partial \mathbf{\Gamma}_\lambda}{\partial x^\kappa} - \frac{\partial \mathbf{\Gamma}_\kappa}{\partial x^\lambda} \right) d^2x^{\kappa\lambda} = \left(\frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{mn\kappa}}{\partial x^\lambda} \right) \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda}, \quad (16.208)$$

implicitly summed over distinct antisymmetric indices mn and $\kappa\lambda$. The commutator $\frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}]$ of the 1-form $\mathbf{\Gamma}$ with itself is the bivector 2-form

$$\frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}] = \frac{1}{2}[\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] d^2x^{\kappa\lambda} = (\Gamma_{m\lambda}^p \Gamma_{pn\kappa} - \Gamma_{m\kappa}^p \Gamma_{pn\lambda}) \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda}, \quad (16.209)$$

again implicitly summed over distinct antisymmetric indices mn and $\kappa\lambda$. The commutator $[\mathbf{\Gamma}, \mathbf{\Gamma}]$ of the bivector 1-form $\mathbf{\Gamma}$ is symmetric, the anticommutation of multivectors cancelling against the anticommutation of 1-forms, equation (16.197). Equations (16.208) and (16.209) imply that the Riemann 2-form R is related to the Lorentz connection 1-form $\mathbf{\Gamma}$ by

$$\boxed{R \equiv d\mathbf{\Gamma} + \frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}]} \quad (16.210)$$

Equation (16.210) is **Cartan's second equation of structure**. It constitutes the definition of Riemann curvature $\mathbf{R}$ in terms of the Lorentz connection $\mathbf{\Gamma}$.

The torsion vector 2-form $\mathbf{S}$ is defined by (the minus sign ensures that Cartan's equation (16.214) takes conventional form, given the definition (11.48) of the components $S_{\kappa\lambda}^m$)

$$\mathbf{S} \equiv -S_{m\kappa\lambda} \gamma^m d^2x^{\kappa\lambda}, \quad (16.211)$$

implicitly summed over distinct antisymmetric indices $\kappa\lambda$. The exterior derivative of the line interval 1-form $\mathbf{e}$ is the 2-form

$$d\mathbf{e} \equiv \left(\frac{\partial e_\lambda}{\partial x^\kappa} - \frac{\partial e_\kappa}{\partial x^\lambda} \right) d^2x^{\kappa\lambda} = \left(\frac{\partial e_{m\lambda}}{\partial x^\kappa} - \frac{\partial e_{m\kappa}}{\partial x^\lambda} \right) \gamma^m d^2x^{\kappa\lambda} = -2d_{m\kappa\lambda} \gamma^m d^2x^{\kappa\lambda}, \quad (16.212)$$

again implicitly summed over distinct antisymmetric indices $\kappa\lambda$. The $d_{m\kappa\lambda}$ in the rightmost expression of equations (16.212) are the vierbein derivatives defined by equation (11.32). The commutator $\frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}]$ of the 1-forms $\mathbf{\Gamma}$ and $\mathbf{e}$ is the vector 2-form

$$\frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}] = [\mathbf{\Gamma}_\kappa, \mathbf{e}_\lambda] d^2x^{\kappa\lambda} = -2\Gamma_{m\kappa\lambda} \gamma^m d^2x^{\kappa\lambda}, \quad (16.213)$$

implicitly summed over distinct antisymmetric indices $\kappa\lambda$. The fundamental relation (11.49), or equivalently (15.29), between the torsion and the vierbein derivatives and Lorentz connections translates in multi-vector forms language to, from equations (16.211)–(16.213),

$$\boxed{\mathbf{S} \equiv d\mathbf{e} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}]} . \quad (16.214)$$

Equation (16.214) is **Cartan's first equation of structure**. Cartan's equations of structure (16.214) and (16.210), introduced by Cartan (1904), are not equations of motion; rather, they are compact and elegant expressions of the definition (11.58) of torsion and curvature. Equations of motion (16.252) for the torsion and curvature are obtained from extremizing the Hilbert action.

16.14.3 Area and volume forms

The other factor in the gravitational Lagrangian (16.163) is $\mathbf{e}^\kappa \wedge \mathbf{e}^\lambda$. The 2-form corresponding to $\mathbf{e}^\kappa \wedge \mathbf{e}^\lambda$ is the element of area $\mathbf{e}^2$ defined by equation (16.204),

$$\mathbf{e}^2 = \frac{1}{2} \mathbf{e} \wedge \mathbf{e} = \mathbf{e}_\kappa \wedge \mathbf{e}_\lambda d^2x^{\kappa\lambda} = (e_{m\kappa} e_{n\lambda} - e_{m\lambda} e_{n\kappa}) \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda}, \quad (16.215)$$

implicitly summed over distinct antisymmetric pairs mn and $\kappa\lambda$ of indices.

The exterior derivative $d\mathbf{e}^p$ of the p -volume element is

$$d\mathbf{e}^p = (-)^{p-1} \mathbf{e}^{p-1} \wedge d\mathbf{e}. \quad (16.216)$$

The $1/p!$ factor in the definition (16.204) of the p -volume element absorbs the factor of p from differentiating p products of $\mathbf{e}$. The $(-)^{p-1}$ sign comes from commuting $d\mathbf{e}$ past $\mathbf{e}^{p-1}$.

The form dual of the p -volume $\mathbf{e}^p$ is the dual q -volume, $*(\mathbf{e}^p) \equiv *\mathbf{e}^q$, which in turn equals the pseudoscalar I times the q -volume, equation (15.80),

$$*(\mathbf{e}^p) = *\mathbf{e}^q = I \mathbf{e}^q . \quad (16.217)$$

The p -volume and its q -form dual are both multivectors of grade p . For example, the form dual, equation (16.194), of the area element $\mathbf{e}^2$ is the dual area element $*\mathbf{e}^2$,

$$*\mathbf{e}^2 = \varepsilon_{\kappa\lambda\mu\nu} \mathbf{e}^\mu \wedge \mathbf{e}^\nu d^2x^{\kappa\lambda} = 2 \varepsilon_{\kappa\lambda\mu\nu} e_m^\mu e_n^\nu \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n d^2x^{\kappa\lambda} = 2 \varepsilon_{klmn} e_{\kappa}^k e_{\lambda}^l \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n d^2x^{\kappa\lambda} , \quad (16.218)$$

implicitly summed over distinct antisymmetric indices $\kappa\lambda$, $\mu\nu$, kl , and mn . The exterior derivative $d*\mathbf{e}^q$ of the dual q -volume element is

$$d*\mathbf{e}^q = (-)^{N-1} I d\mathbf{e}^q = (-)^p I (\mathbf{e}^{q-1} \wedge d\mathbf{e}) = (-)^p (I \mathbf{e}^{q-1}) \cdot d\mathbf{e} = (-)^p *\mathbf{e}^{q-1} \cdot d\mathbf{e} , \quad (16.219)$$

the third equality following from the duality relation (13.41).

Exercise 16.12 Triple products involving products of the interval form $\mathbf{e}$. Let $\mathbf{a}$ be a multivector form of grade n and any form index.

1. Show that

$$\mathbf{e} \wedge (\mathbf{e} \cdot \mathbf{a}) = \mathbf{e} \cdot (\mathbf{e} \wedge \mathbf{a}) . \quad (16.220)$$

2. Conclude that if $n \geq q$ then

$$\mathbf{e}^p \wedge (\mathbf{e}^q \cdot \mathbf{a}) = \mathbf{e}^q \cdot (\mathbf{e}^p \wedge \mathbf{a}) . \quad (16.221)$$

3. Prove that the grade $p+n-q$ part of the multivector form $\mathbf{e}^{p+q}\mathbf{a}$ is

$$\langle \mathbf{e}^{p+q}\mathbf{a} \rangle_{p+n-q} = \mathbf{e}^q \cdot (\mathbf{e}^p \wedge \mathbf{a}) . \quad (16.222)$$

Equation (16.222) is equivalent to

$$\langle \mathbf{e}^{p+q}\mathbf{a} \rangle_{p+n-q} = \langle \mathbf{e}^q \langle \mathbf{e}^p \mathbf{a} \rangle_{p+n} \rangle_{p+n-q} . \quad (16.223)$$

The proof below of equations (16.222) or (16.223) uses the triple-product relation (13.40). The proof demonstrates along the way the triple-product relation

$$\langle \mathbf{e}^p \langle \mathbf{e}^q \mathbf{a} \rangle_{q+n-2l} \rangle_{p+q+n-2k-2l} = \frac{(k+l)! (p+q-k-l)!}{k!l! (p-k)!(q-l)!} \langle \mathbf{e}^{p+q}\mathbf{a} \rangle_{p+q+n-2k-2l} . \quad (16.224)$$

Solution.

1. Equation (16.220) can be proved by expanding the multivector forms $\mathbf{e}$ and $\mathbf{a}$ in components. Equation (16.220) remains true in the special case where $\mathbf{a}$ is a scalar (grade 0), in which case $\mathbf{e} \cdot \mathbf{a} = 0$ by definition (13.36), and $\mathbf{e} \cdot \mathbf{e} = 0$, equation (16.205).
2. Equation (16.221) follows from successive application of equation (16.220).

3. Equation (16.223) can be proved by induction. Certainly equation (16.223) holds for $p = 0$ or $q = 0$, in which case the equation becomes an identity. Recall that, in view of the way that volume elements e^p are normalized, equation (16.204),

$$e^{p+q} = \frac{p!q!}{(p+q)!} e^p \wedge e^q . \quad (16.225)$$

The triple-product relation (13.40), along with the fact that $e \cdot e = 0$, implies

$$\langle e^{p+q} \mathbf{a} \rangle_{p+q+n-2m} = \frac{p!q!}{(p+q)!} \sum_{l=0}^m \langle e^p \langle e^q \mathbf{a} \rangle_{q+n-2l} \rangle_{p+q+n-2m} . \quad (16.226)$$

Assume that equation (16.222) is inductively true up to some p and q . The inductive hypothesis (16.222) implies

$$\langle e^q \mathbf{a} \rangle_{q+n-2l} = e^l \cdot (e^{q-l} \wedge \mathbf{a}) \quad (16.227)$$

subject to the conditions that $q-l$, l , n , and $q+n-2l$ are all non-negative integers. Inserting the hypothesis (16.227), and a similar one for $\langle e^p \dots \rangle$, into equation (16.226) implies that the summand on the right hand side of equation (16.226) is

$$\langle e^p \langle e^q \mathbf{a} \rangle_{q+n-2l} \rangle_{p+q+n-2k-2l} = e^k \cdot (e^{p-k} \wedge (e^l \cdot (e^{q-l} \wedge \mathbf{a}))) . \quad (16.228)$$

The sum in equation (16.226) is over non-negative integers k and l satisfying $k+l=m$, $k \leq p$, and $l \leq q$. By equation (16.221), the summand (16.228) rearranges as

$$\begin{aligned} \langle e^p \langle e^q \mathbf{a} \rangle_{q+n-2l} \rangle_{p+q+n-2k-2l} &= (e^k \wedge e^l) \cdot (e^{p-k} \wedge e^{q-l} \wedge \mathbf{a}) \\ &= \frac{(k+l)!}{k!l!} \frac{(p+q-k-l)!}{(p-k)!(q-l)!} e^{k+l} \cdot (e^{p+q-k-l} \wedge \mathbf{a}) \\ &= \frac{m!}{k!l!} \frac{(p+q-m)!}{(p-k)!(q-l)!} e^m \cdot (e^{p+q-m} \wedge \mathbf{a}) . \end{aligned} \quad (16.229)$$

Equation (16.226) thus reduces to

$$\langle e^{p+q} \mathbf{a} \rangle_{p+q+n-2m} = e^m \cdot (e^{p+q-m} \wedge \mathbf{a}) \frac{p!q!}{(p+q)!} \sum_{k+l=m, k \leq p, l \leq q} \frac{m!(p+q-m)!}{k!l!(p-k)!(q-l)!} . \quad (16.230)$$

The summed term on the right hand side of equation (16.230) equals $(p+q)!/(p!q!)$, cancelling the prefactor $p!q!/(p+q)!$. To prove this, it suffices to restrict to $p = 1$ or $q = 1$, with $m \geq 1$ (the result is trivial for $m = 0$), and then the general result follows by induction. For $p = 1$ the sum is over $k = 0$ and 1 , while for $q = 1$ the sum is over $l = 0$ and 1 . For example, for $q = 1$,

$$\frac{p!1!}{(p+1)!} \sum_{l=0}^1 \frac{m!(p+1-m)!}{(m-l)!l!(p+l-m)!(1-l)!} = \frac{1}{p+1} [(p+1-m) + m] = 1 . \quad (16.231)$$

Thus, at least for $p = 1$ or $q = 1$, equation (16.230) reduces to

$$\langle e^{p+q} \mathbf{a} \rangle_{p+q+n-2m} = e^m \cdot (e^{p+q-m} \wedge \mathbf{a}) , \quad (16.232)$$

reproducing the to-be-proved equation (16.222). The result for $p = 1$ or $q = 1$ establishes the desired result (16.222) inductively for all p and q . Equation (16.229) and (16.232) together imply equation (16.224).

16.14.4 Gravitational Lagrangian 4-form

Recall that the scalar volume element d^4x that goes into the action is really the dual scalar 4-volume $*d^4x$, equation (15.80). To convert to forms language, the Hodge dual must be transferred from the volume element to the integrand. In multivector language, the required result is equation (16.55), invoked previously to convert the electromagnetic Lagrangian to forms language. Translated back into forms language, equation (16.55) says that a “scalar product” of 2-forms $\mathbf{a}$ and $\mathbf{b}$ over a dual scalar volume element is the 4-form equal to the exterior product of the dual form $*\mathbf{a}$ with the form $\mathbf{b}$.

The gravitational action thus becomes

$$S_g = \int L_g , \quad (16.233)$$

where L_g is the gravitational Lagrangian scalar 4-form corresponding to the Lagrangian (16.163),

$$L_g \equiv -\frac{1}{8\pi} *e^2 \cdot \mathbf{R} = -\frac{1}{8\pi} *e^2 \cdot (d\mathbf{\Gamma} + \frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}]) . \quad (16.234)$$

The dot in $*e^2 \cdot \mathbf{R}$ signifies a scalar product of the bivectors $*e^2$ and $\mathbf{R}$. The minus sign comes from taking a scalar product of bivectors. The product $*e^2 \cdot \mathbf{R}$ is an exterior product of two 2-forms, hence a 4-form. As remarked at the beginning of this section 16.14, the wedge sign for the exterior product of forms is suppressed because it conflicts with the wedge sign for a multivector product, and because it is unnecessary, there being only one way to multiply forms. The Lagrangian 4-form (16.234) is in Hamiltonian form $L_g = \mathbf{p} \cdot d\mathbf{q} - H_g$ with coordinates $\mathbf{q} = \mathbf{\Gamma}$ and momenta $\mathbf{p} = -*e^2/(8\pi)$, and (super-)Hamiltonian 4-form

$$H_g = \frac{1}{32\pi} *e^2 \cdot [\mathbf{\Gamma}, \mathbf{\Gamma}] . \quad (16.235)$$

The Lagrangian scalar 4-form (16.234) can be written elegantly, from the expression (16.217) for the dual volume $*e^2$ and the duality relation (13.41),

$$\boxed{L_g \equiv -\frac{I}{8\pi} e^2 \wedge \mathbf{R} = -\frac{I}{8\pi} e^2 \wedge (d\mathbf{\Gamma} + \frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}])} . \quad (16.236)$$

Expanded in components, the gravitational Lagrangian 4-form (16.234) or (16.236) is

$$L_g = -\frac{1}{8\pi} \varepsilon_{\mu\nu\pi\rho} (e^\pi \wedge e^\rho) \cdot \mathbf{R}_{\kappa\lambda} d^4x^{\kappa\lambda\mu\nu} = -\frac{I}{8\pi} e_\mu \wedge e_\nu \wedge \mathbf{R}_{\kappa\lambda} d^4x^{\kappa\lambda\mu\nu} , \quad (16.237)$$

implicitly summed over distinct antisymmetric indices $\kappa\lambda$, $\mu\nu$, and $\pi\rho$. The Lagrangian 4-form (16.236) is in Hamiltonian form $L_g = I(\mathbf{p} \wedge d\mathbf{q}) - H_g$ with coordinates $\mathbf{q} = \mathbf{\Gamma}$ and momenta $\mathbf{p} = -e^2/(8\pi)$, and (super-)Hamiltonian scalar 4-form

$$\boxed{H_g = \frac{I}{32\pi} e^2 \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}]} . \quad (16.238)$$

16.14.5 Variation of the gravitational action in multivector forms notation

Equations of motion for the gravitational field are obtained by varying the action with respect to the Lorentz connection $\mathbf{\Gamma}$ and the line-element $\mathbf{e}$. In forms notation, when the fields are varied, it is the coefficients $\Gamma_{kl\kappa}$ and $e_{k\kappa}$ that are varied, the tetrad γ^k and the line interval dx^κ being considered fixed. Thus the variation $\delta\mathbf{\Gamma}$ of the Lorentz connection is

$$\delta\mathbf{\Gamma} \equiv (\delta\mathbf{\Gamma}_\kappa) dx^\kappa \equiv (\delta\Gamma_{kl\kappa}) \gamma^k \wedge \gamma^l dx^\kappa, \quad (16.239)$$

implicitly summed over distinct antisymmetric indices kl , and the variation $\delta\mathbf{e}$ of the line interval is

$$\delta\mathbf{e} \equiv (\delta\mathbf{e}_\kappa) dx^\kappa \equiv (\delta e_{k\kappa}) \gamma^k dx^\kappa. \quad (16.240)$$

The variation $\delta\mathbf{e}^p$ of the p -volume element defined by equation (16.204) is

$$\delta\mathbf{e}^p = \mathbf{e}^{p-1} \wedge \delta\mathbf{e}. \quad (16.241)$$

The variation $\delta^*\mathbf{e}^q$ of the dual q -volume element is

$$\delta^*\mathbf{e}^q = I \delta\mathbf{e}^q = I(\mathbf{e}^{q-1} \wedge \delta\mathbf{e}) = (I\mathbf{e}^{q-1}) \cdot \delta\mathbf{e} = {}^*\mathbf{e}^{q-1} \cdot \delta\mathbf{e}, \quad (16.242)$$

the third equality following from the duality relation (13.41).

The variation of the action with gravitational Lagrangian (16.236) with respect to the fields $\mathbf{\Gamma}$ and $\mathbf{e}$ is

$$\delta S_g = -\frac{I}{8\pi} \int \mathbf{e}^2 \wedge d(\delta\mathbf{\Gamma}) + \frac{1}{4} \mathbf{e}^2 \wedge \delta[\mathbf{\Gamma}, \mathbf{\Gamma}] + \delta(\mathbf{e}^2) \wedge \mathbf{R}. \quad (16.243)$$

The first term integrates by parts to

$$\mathbf{e}^2 \wedge d(\delta\mathbf{\Gamma}) = d(\mathbf{e}^2 \wedge \delta\mathbf{\Gamma}) - d(\mathbf{e}^2) \wedge \delta\mathbf{\Gamma}. \quad (16.244)$$

The second term in the integrand of (16.243) is

$$\frac{1}{4} \mathbf{e}^2 \wedge \delta[\mathbf{\Gamma}, \mathbf{\Gamma}] = \frac{1}{2} \mathbf{e}^2 \wedge [\mathbf{\Gamma}, \delta\mathbf{\Gamma}] = \frac{1}{2} [\mathbf{e}^2, \mathbf{\Gamma}] \wedge \delta\mathbf{\Gamma} = -\frac{1}{2} [\mathbf{\Gamma}, \mathbf{e}^2] \wedge \delta\mathbf{\Gamma}, \quad (16.245)$$

the second step of which applies the multivector triple-product relation (13.39). The coefficients of the $\wedge \delta\mathbf{\Gamma}$ terms in equations (16.244) and (16.245) combine to

$$-d(\mathbf{e}^2) - \frac{1}{2} [\mathbf{\Gamma}, \mathbf{e}^2] = \mathbf{e} \wedge \mathbf{S}, \quad (16.246)$$

the torsion $\mathbf{S}$ being defined by equation (16.214). To switch between commutators $[\mathbf{\Gamma}, \mathbf{e}]$ and commutators $[\mathbf{\Gamma}, \mathbf{e}^2]$, use the result (16.220) along with the fact that $\mathbf{e} \cdot \mathbf{a} = \frac{1}{2} [\mathbf{e}, \mathbf{a}]$ for any bivector form $\mathbf{a}$. The third term in the integrand of (16.243) is

$$\delta(\mathbf{e}^2) \wedge \mathbf{R} = \delta\mathbf{e} \wedge \mathbf{e} \wedge \mathbf{R} = \delta\mathbf{e} \wedge {}^{**}\mathbf{G}, \quad (16.247)$$

where $\overset{**}{\mathbf{G}} \equiv \mathbf{e} \wedge \mathbf{R}$ is the double dual, equation (16.195), of the Einstein vector 1-form $\mathbf{G} \equiv G_{\nu n} \gamma^n dx^\nu$,

$$\begin{aligned} \mathbf{G} &\equiv I^*(\mathbf{e} \wedge \mathbf{R}) \\ &= \varepsilon_k^{lmn} \varepsilon_\kappa^{\lambda\mu\nu} e_{l\lambda} R_{\mu\nu mn} \gamma^k dx^\kappa \\ &= -3! e^{[k}_{[\kappa} e^m_\mu e^{n]}_{\nu]} R^{\mu\nu}_{mn} \gamma_k dx^\kappa \\ &= (R_{\kappa k} - \tfrac{1}{2} R e_{\kappa\kappa}) \gamma^k dx^\kappa, \end{aligned} \quad (16.248)$$

implicitly summed over distinct antisymmetric sequences mn and $\mu\nu$, and over all k, l, κ , and λ . The factor of $3!$ on the third line of equations (16.248) is the number of permutations of the indices of a 3-form. Combining equations (16.244)–(16.247) brings the variation (16.243) of the gravitational action to

$$\delta S_g = -\frac{I}{8\pi} \oint e^2 \wedge \delta \mathbf{\Gamma} - \frac{I}{8\pi} \int (\mathbf{e} \wedge \mathbf{S}) \wedge \delta \mathbf{\Gamma} + \delta \mathbf{e} \wedge (\mathbf{e} \wedge \mathbf{R}). \quad (16.249)$$

The variation of the matter action S_m with respect to $\delta \mathbf{\Gamma}$ and $\delta \mathbf{e}$ defines the spin angular-momentum $\mathbf{\Sigma}$ (compare equation (16.121)), and the matter energy-momentum $\mathbf{T}$ (compare equation (16.117)),

$$\delta S_m = - \int {}^* \mathbf{\Sigma} \cdot \delta \mathbf{\Gamma} + \delta \mathbf{e} \cdot {}^* \mathbf{T} = I \int \overset{**}{\mathbf{\Sigma}} \wedge \delta \mathbf{\Gamma} + \delta \mathbf{e} \wedge \overset{**}{\mathbf{T}}, \quad (16.250)$$

where $\overset{**}{\mathbf{\Sigma}}$ and $\overset{**}{\mathbf{T}}$ are the double duals, equation (16.195), of the spin angular-momentum $\mathbf{\Sigma}$ and energy-momentum $\mathbf{T}$ of the matter. The components of the spin angular-momentum bivector 1-form $\mathbf{\Sigma}$ and the energy-momentum vector 1-form $\mathbf{T}$ are (the minus sign in the definition of $\mathbf{\Sigma}$ conforms to the convention for the definition of torsion $\mathbf{S}$, equation (16.211))

$$\mathbf{\Sigma} \equiv \Sigma_\kappa dx^\kappa = -\Sigma_{\kappa lm} \gamma^l \wedge \gamma^m dx^\kappa, \quad (16.251a)$$

$$\mathbf{T} \equiv T_\kappa dx^\kappa = T_{\kappa m} \gamma^m dx^\kappa, \quad (16.251b)$$

with, for $\mathbf{\Sigma}$, implicit summation over distinct antisymmetric sets of indices kl . Extremizing the combined gravitational and matter actions with respect to $\delta \mathbf{\Gamma}$ and $\delta \mathbf{e}$ yields the torsion and Einstein equations of motion in the form

$$\boxed{\mathbf{e} \wedge \mathbf{S} = 8\pi \overset{**}{\mathbf{\Sigma}}}, \quad (16.252a)$$

$$\boxed{\mathbf{e} \wedge \mathbf{R} = 8\pi \overset{**}{\mathbf{T}}}. \quad (16.252b)$$

The torsion equation of motion (16.252a) is a bivector 3-form with $6 \times 4 = 24$ components, while the Einstein equation of motion (16.252b) is a pseudovector 3-form with $4 \times 4 = 16$ components. The Einstein equation (16.252b) is equivalent to the traditional expression

$$\mathbf{G} = 8\pi \mathbf{T}. \quad (16.253)$$

The contracted Bianchi identities (16.408) enforce conservation laws for the total spin angular-momentum $\overset{**}{\mathbf{\Sigma}}$ and total matter energy-momentum $\overset{**}{\mathbf{T}}$, §16.14.8 and §16.14.9.

Notice that if the area element e^2 had been taken to be the momentum conjugate to $\mathbf{\Gamma}$, rather than the

line element e , all the components of the area element e^2 being considered to be independent degrees of freedom, then the variation (16.243) of the gravitational action with respect to e^2 would have yielded an equation of motion for the Riemann tensor R rather than for the Einstein tensor G , and the theory would not be general relativity. To recover general relativity, it is necessary to treat the area element as a wedge product $e^2 = \frac{1}{2}e \wedge e$ of the line interval e .

16.14.6 Alternative gravitational action in multivector forms notation

As in §16.9 and §16.13.3, the coordinates and momenta can be traded without changing the equations of motion. Integrating the $-e^2 \wedge d\Gamma$ term in the Lagrangian (16.236) by parts gives

$$\begin{aligned} -e^2 \wedge d\Gamma &= -d(e^2 \wedge \Gamma) + de \wedge (e \wedge \Gamma) \\ &= d\vartheta + \pi \wedge de, \end{aligned} \quad (16.254)$$

where π is the momentum conjugate to e , a trivector 2-form with 24 components,

$$\boxed{\pi \equiv -e \wedge \Gamma}, \quad (16.255)$$

and ϑ is the expansion, the contraction of π , a pseudoscalar 3-form with 4 components,

$$\boxed{\vartheta \equiv \frac{1}{2}e \wedge \pi = -e^2 \wedge \Gamma}. \quad (16.256)$$

The double dual of the expansion is the scalar 1-form

$$\vartheta^{**} = \Gamma_{\kappa m}^m dx^\kappa. \quad (16.257)$$

The transpose (16.273) of the double dual of the expansion is

$$\vartheta^{**\top} = \Gamma_{km}^m \gamma^k, \quad (16.258)$$

whose tetrad time component Γ_{0m}^m is what is commonly called the expansion, §18.3, justifying the nomenclature.

The total exterior derivative term $d\vartheta$ in equation (16.254) is the Gibbons-Hawking-York boundary term (York, 1972; Gibbons and Hawking, 1977). Discarding the boundary term $d\vartheta$ yields the alternative Lagrangian

$$L'_g \equiv L_g - d\vartheta = \frac{I}{8\pi} \pi \wedge de - H_g = \frac{I}{8\pi} \pi \wedge (de + \frac{1}{4}[\Gamma, e]), \quad (16.259)$$

with H_g is the same (super-)Hamiltonian as before, equation (16.238). The alternative Lagrangian (16.259) is in Hamiltonian form with coordinates e and momenta $\pi/(8\pi)$.

The Lorentz connection Γ , which is a bivector 1-form, and the momentum π , which is a pseudovector 2-form, both have the same number of components, 24. The components are invertibly related to each other, the Lorentz connection Γ being given in terms of the momentum π by

$$\Gamma = -\pi^{**\top} + e \wedge \vartheta^{**\top}, \quad (16.260)$$

where $\top$ denotes the transpose operation (16.273). Equation (16.260) is the analogue of equation (16.183) in multivector notation. The bivector 1-form momentum π_λ in the earlier notation is related to the trivector 2-form momentum π in the present notation by $\pi_\lambda dx^\lambda = -\pi^{**\top}$. The difference between the momenta because the earlier Lagrangian (16.185) involved $\pi \cdot de$, whereas the present Lagrangian (16.259) involves $\pi \wedge de$.

Variation of the gravitational action S'_g with the alternative Lagrangian (16.259) yields

$$\delta S'_g = \frac{I}{8\pi} \oint \pi \wedge \delta e + \frac{I}{8\pi} \int \delta \pi \wedge S - \Pi \wedge \delta e, \quad (16.261)$$

where the curvature pseudovector 3-form Π is defined to be

$$\boxed{\Pi \equiv e \wedge R - S \wedge \Gamma = d\pi + \frac{1}{2}[\Gamma, \pi] - \frac{1}{4}e \wedge [\Gamma, \Gamma]}. \quad (16.262)$$

Previously, variation of the matter action S_m with respect to $\delta\Gamma$ and δe defined the (double duals of the) spin angular-momentum Σ and matter energy-momentum T , equation (16.250). Variation of the matter action S_m instead with respect to $\delta\pi$ and δe defines modified versions $\tilde{\Sigma}$ and $\tilde{T}$ of the spin angular-momentum and energy-momentum,

$$\delta S_m = I \int -\delta \pi \wedge \tilde{\Sigma} + \tilde{T} \wedge \delta e. \quad (16.263)$$

where the vector 2-form $\tilde{\Sigma}$ is (the minus sign conforms to the convention for the torsion S and spin angular-momentum Σ , equations (16.211) and (16.251a))

$$\tilde{\Sigma} = -\tilde{\Sigma}_{\lambda mn} \gamma^m \wedge \gamma^n dx^\lambda. \quad (16.264)$$

The original Σ and modified $\tilde{\Sigma}$ spin angular-momenta are invertibly related to each other (similarly to the way that the momentum π and connection Γ are invertibly related, equations (16.255) and (16.260)), while the (double dual of the) original T and modified $\tilde{T}$ energy-momenta differ by a term depending on the spin angular-momentum,

$$\Sigma^{**} = e \wedge \tilde{\Sigma}, \quad \tilde{\Sigma} = \Sigma^\top - e \wedge \sigma^{**\top}, \quad (16.265a)$$

$$T^{**} = \tilde{T} - \tilde{\Sigma} \wedge \Gamma, \quad (16.265b)$$

where σ is the contraction of Σ^{**} , a pseudoscalar 3-form with 4 components,

$$\sigma \equiv \frac{1}{2} e \wedge \Sigma^{**} = e^2 \wedge \tilde{\Sigma}. \quad (16.266)$$

In components, the relation (16.265a) between the original Σ and modified $\tilde{\Sigma}$ spin angular-momenta is equation (16.130). The components of the double dual σ^{**} and its transpose are (compare equations (16.257) and (16.258) for ϑ^{**} and $\vartheta^{**\top}$)

$$\sigma^{**} = -\tilde{\Sigma}_{\nu k}^k dx^\nu = \frac{1}{2} \Sigma_{\nu k}^k dx^\nu, \quad \sigma^{**\top} = -\tilde{\Sigma}_{nk}^k \gamma^n = \frac{1}{2} \Sigma_{nk}^k \gamma^n. \quad (16.267)$$

The equations of motion for the torsion $\mathbf{S}$ and curvature $\mathbf{\Pi}$ are

$$\boxed{\mathbf{S} = 8\pi\tilde{\mathbf{\Sigma}}} , \quad (16.268a)$$

$$\boxed{\mathbf{\Pi} = 8\pi\tilde{\mathbf{T}}} . \quad (16.268b)$$

More explicitly, the equations of motion (16.268) are

$$d\mathbf{e} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}] = 8\pi\tilde{\mathbf{\Sigma}} , \quad (16.269a)$$

$$d\mathbf{\pi} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\pi}] - \frac{1}{4}\mathbf{e} \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}] = 8\pi\tilde{\mathbf{T}} . \quad (16.269b)$$

The expansion $\boldsymbol{\vartheta}$ is a pseudoscalar, so its exterior derivative equals its Lorentz-covariant exterior derivative, $d\boldsymbol{\vartheta} = d\boldsymbol{\vartheta} + \frac{1}{2}[\mathbf{\Gamma}, \boldsymbol{\vartheta}]$, which is

$$d\boldsymbol{\vartheta} = \frac{1}{2} (d\mathbf{e} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}]) \wedge \mathbf{\pi} - \frac{1}{2}\mathbf{e} \wedge (d\mathbf{\pi} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\pi}]) , \quad (16.270)$$

which rearranges to

$$d\boldsymbol{\vartheta} + \frac{1}{4}\mathbf{e}^2 \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}] = -\mathbf{e}^2 \wedge \mathbf{R} + \mathbf{e} \wedge \mathbf{S} \wedge \mathbf{\Gamma} . \quad (16.271)$$

The first term on the right hand side of equation (16.271) is proportional to the double-dual of the trace G of the Einstein tensor, $\mathbf{e}^2 \wedge \mathbf{R} = \mathbf{G}^{**}$. If $\mathbf{e}^2 \wedge \mathbf{R}$ and $\mathbf{e} \wedge \mathbf{S}$ are replaced by their matter energy-momentum and spin angular-momentum sources in accordance with equation (16.252), then the equation of motion (16.271) for the expansion becomes

$$d\boldsymbol{\vartheta} + \frac{1}{4}\mathbf{e}^2 \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}] = 8\pi(-\frac{1}{2}\mathbf{T}^{**} + \mathbf{\Sigma}^{**} \wedge \mathbf{\Gamma}) . \quad (16.272)$$

16.14.7 Transpose of a multivector form

The **transpose** $\mathbf{a}^\top$ of a multivector form $\mathbf{a} \equiv a_{K\Lambda} \boldsymbol{\gamma}^K d^p x^\Lambda$ of grade n and form index p is defined to be the multivector form, of grade p and form index n , with multivector and form indices transposed,

$$\mathbf{a}^\top = (a_{K\Lambda} \boldsymbol{\gamma}^K d^p x^\Lambda)^\top \equiv e^K_{\Pi} e_L^\Lambda a_{K\Lambda} \boldsymbol{\gamma}^L d^p x^\Pi = a_{\Pi L} \boldsymbol{\gamma}^L d^k x^\Pi , \quad (16.273)$$

implicitly summed over distinct sequences K, L, Λ, Π of indices. For example, the transpose of a vector 2-form is the bivector 1-form

$$(a_{k\lambda\mu} \boldsymbol{\gamma}^k d^2 x^{\lambda\mu})^\top \equiv e^k_{\kappa} e_l^\lambda e_m^\mu a_{k\lambda\mu} \boldsymbol{\gamma}^l \wedge \boldsymbol{\gamma}^m dx^\kappa = a_{\kappa lm} \boldsymbol{\gamma}^l \wedge \boldsymbol{\gamma}^m dx^\kappa . \quad (16.274)$$

The transpose of a symmetric tensor $\mathbf{a}$, one satisfying, $a_{k\lambda} \equiv a_{kl} e^l_\lambda = a_{lk} e^l_\lambda \equiv a_{\lambda k}$, is itself,

$$\mathbf{a}^\top = (a_{k\lambda} \boldsymbol{\gamma}^k dx^\lambda)^\top = a_{\lambda k} \boldsymbol{\gamma}^k dx^\lambda = a_{k\lambda} \boldsymbol{\gamma}^k dx^\lambda = \mathbf{a} . \quad (16.275)$$

As a particular example, the vierbein is symmetric in this sense, because the tetrad metric is symmetric, $e_{k\lambda} = \eta_{kl} e^l_\lambda$, so the transpose of the line interval $\mathbf{e}$ is itself,

$$\mathbf{e}^\top = (e_{k\lambda} \boldsymbol{\gamma}^k dx^\lambda)^\top = e^k_{\kappa} e_l^\lambda e_{k\lambda} \boldsymbol{\gamma}^l dx^\kappa = e_{l\kappa} \boldsymbol{\gamma}^l dx^\kappa = \mathbf{e} . \quad (16.276)$$

The transpose of a wedge product of multivector forms $\mathbf{a}$ and $\mathbf{b}$ is the wedge product of their transposes,

$$(\mathbf{a} \wedge \mathbf{b})^\top = \mathbf{a}^\top \wedge \mathbf{b}^\top . \quad (16.277)$$

The transpose of the double dual of a multivector form $\mathbf{a}$ is the double dual of its transpose,

$$\mathbf{a}^{**\top} = \mathbf{a}^{\top**} . \quad (16.278)$$

Equations (16.277) and (16.278) say that the operation of transposition commutes both with taking the wedge product and with taking the double dual. Note however that the operations of taking the wedge product and taking the double dual do not commute.

16.14.8 Conservation of angular momentum in multivector forms language

The action S_m of any individual matter field is Lorentz invariant. Lorentz symmetry implies a conservation law (16.283) of angular momentum.

Under an infinitesimal Lorentz transformation generated by the bivector $\epsilon = \epsilon_{kl} \gamma^k \wedge \gamma^l$, any multivector form $\mathbf{a}$ whose multivector components transform like a tensor varies as, equation (16.95),

$$\delta \mathbf{a} = \frac{1}{2} [\epsilon, \mathbf{a}] . \quad (16.279)$$

In particular, since the vierbein $e_{k\kappa}$ is a tetrad vector, the variation of the line interval $e \equiv e_{k\kappa} \gamma^k dx^\kappa$ under an infinitesimal Lorentz transformation is

$$\delta e = \frac{1}{2} [\epsilon, e] . \quad (16.280)$$

The components of the Lorentz connection $\Gamma \equiv \Gamma_{mn\lambda} \gamma^m \wedge \gamma^n dx^\lambda$ do not constitute a tetrad tensor, so do not transform like equation (16.279). Rather, the Lorentz connection transforms as equation (16.96), which in multivector forms language translates to

$$\delta \Gamma = - (d\epsilon + \frac{1}{2} [\Gamma, \epsilon]) . \quad (16.281)$$

Inserting the variations (16.280) and (16.281) of the line interval e and Lorentz connection Γ into the variation (16.250) of the matter action yields the variation of the matter action under an infinitesimal Lorentz transformation generated by the bivector ϵ ,

$$\begin{aligned} \delta S_m &= I \int -\mathbf{\Sigma} \wedge (d\epsilon + \frac{1}{2} [\Gamma, \epsilon]) + \frac{1}{2} [\epsilon, e] \wedge \mathbf{T} \\ &= I \oint \mathbf{\Sigma} \wedge \epsilon - I \int (d\mathbf{\Sigma} + \frac{1}{2} [\Gamma, \mathbf{\Sigma}] - e \cdot \mathbf{T}) \wedge \epsilon . \end{aligned} \quad (16.282)$$

Invariance of the action under local Lorentz transformations requires that the variation (16.282) must vanish for arbitrary choices of the bivector ϵ vanishing on the initial and final hypersurfaces. Consequently the spin angular-momentum $\mathbf{\Sigma}$ must satisfy the covariant conservation equation

$$\boxed{d\mathbf{\Sigma} + \frac{1}{2} [\Gamma, \mathbf{\Sigma}] - e \cdot \mathbf{T} = 0} . \quad (16.283)$$

Equation (16.283) is the same as the conservation equation (16.136) derived previously in index notation. Equation (16.283) is consistent with the contracted torsion Bianchi identity (16.408a), which enforces the angular-momentum conservation equation (16.283) summed over all species. If the spin angular-momentum of a matter component vanishes identically, $\overset{**}{\Sigma} = 0$, then the conservation equation (16.283) implies that the energy-momentum tensor of that matter component is symmetric,

$$\boldsymbol{e} \cdot \overset{**}{T} = 0 . \quad (16.284)$$

16.14.9 Conservation of energy-momentum in multivector forms language

The action S_m of any individual matter field is invariant under coordinate transformations. Symmetry under coordinate transformations implies a conservation law (16.301) for the energy-momentum of the field.

Any infinitesimal 1-form $\boldsymbol{\epsilon} \equiv \epsilon_\kappa dx^\kappa$ generates an infinitesimal coordinate transformation

$$x^\kappa \rightarrow x^\kappa + \epsilon^\kappa . \quad (16.285)$$

As discussed in §7.34, the variation of any quantity $\boldsymbol{a}$ with respect to an infinitesimal coordinate transformation $\boldsymbol{\epsilon}$ is, by construction, minus its Lie derivative, $-\mathcal{L}_\epsilon \boldsymbol{a}$, with respect to the vector ϵ^κ . The Lie derivative of a form is written most elegantly in terms of a dot product of forms. As usual, algebraic operations with forms are derived most easily by translating from multivector language into forms language. Thus the dot product of a 1-form $\boldsymbol{\epsilon}$ with a p -form $\boldsymbol{a} \equiv a_\Lambda d^p x^\Lambda$ is, mirroring the multivector dot product (13.35) (the form dot $\cdot$ is written slightly larger than the multivector dot $\cdot$ to distinguish the two),

$$\boldsymbol{\epsilon} \cdot \boldsymbol{a} \equiv p \epsilon^\kappa a_{\kappa\Lambda} d^{p-1} x^\Lambda , \quad (16.286)$$

implicitly summed over distinct antisymmetric sets of indices $\kappa\Lambda$. The form dot product of a 1-form $\boldsymbol{\epsilon}$ and a 0-form is zero, consistent with the convention (13.36) for multivectors. A useful result is that for any two multivector forms $\boldsymbol{a}$ and $\boldsymbol{b}$ with product $\boldsymbol{ab}$ (a geometric product of multivectors and an exterior product of forms), the dot product of the 1-form $\boldsymbol{\epsilon}$ with the product $\boldsymbol{ab}$ is

$$\boldsymbol{\epsilon} \cdot (\boldsymbol{ab}) = (\boldsymbol{\epsilon} \cdot \boldsymbol{a})\boldsymbol{b} + (-)^p \boldsymbol{a}(\boldsymbol{\epsilon} \cdot \boldsymbol{b}) , \quad (16.287)$$

where p is the form index of $\boldsymbol{a}$.

From the definition (7.157) of the Lie derivative of a coordinate tensor, it can be shown (Exercise 16.13 asks you to do this) that the Lie derivative of a p -form $\boldsymbol{a}$ with respect to a 1-form $\boldsymbol{\epsilon}$ is given by the elegant expression

$$\mathcal{L}_\epsilon \boldsymbol{a} = \boldsymbol{\epsilon} \cdot (d\boldsymbol{a}) + d(\boldsymbol{\epsilon} \cdot \boldsymbol{a}) , \quad (16.288)$$

which is known as **Cartan's magic formula**. With indices explicit, the term $d(\boldsymbol{\epsilon} \cdot \boldsymbol{a})$ in formula (16.288) is

$$d(\boldsymbol{\epsilon} \cdot \boldsymbol{a}) = \partial_\mu (\epsilon^\kappa a_{\kappa\Lambda}) d^p x^{\mu\Lambda} , \quad (16.289)$$

implicitly summed over distinct antisymmetric sets of indices Λ , and over κ not in Λ , and over μ not in Λ (κ and μ may or may not be the same index). Cartan's magic formula (16.288), along with the vanishing of

the exterior derivative squared, $d^2 = 0$, implies that, acting on forms, the Lie derivative $\mathcal{L}_\epsilon$ commutes with the exterior derivative d ,

$$\mathcal{L}_\epsilon d\mathbf{a} - d\mathcal{L}_\epsilon \mathbf{a} = 0 . \quad (16.290)$$

Cartan's magic formula (16.288) holds also for multivector-valued forms, since multivectors are coordinate scalars (they are unchanged by coordinate transformations). However, the Lie derivative of a tetrad tensor is not a tetrad tensor (see Concept question 26.3). A covariant Lie derivative $\mathcal{L}_\Gamma$ can be defined by replacing the exterior derivative with the Lorentz-covariant exterior derivative, $d\mathbf{a} \rightarrow D_\Gamma \mathbf{a} \equiv d\mathbf{a} + \frac{1}{2}[\Gamma, \mathbf{a}]$ (see §16.17.2),

$$\mathcal{L}_\Gamma \epsilon \mathbf{a} \equiv \epsilon \cdot (D_\Gamma \mathbf{a}) + D_\Gamma (\epsilon \cdot \mathbf{a}) . \quad (16.291)$$

The Lorentz-covariant derivative D_Γ is torsion-free acting on coordinate indices, but torsion-full acting on multivector (Lorentz) indices. An alternative version of the covariant magic formula (16.291) in terms of the torsion-free exterior derivative $\mathring{D}$ and the contortion $\mathbf{K}$ is

$$\mathcal{L}_\Gamma \epsilon \mathbf{a} = \epsilon \cdot (\mathring{D}\mathbf{a}) + \mathring{D}(\epsilon \cdot \mathbf{a}) + \frac{1}{2}[\epsilon \cdot \mathbf{K}, \mathbf{a}] , \quad (16.292)$$

which follows from $\Gamma = \mathring{\Gamma} + \mathbf{K}$ and the relation (16.287). As a particular example of the Lorentz-covariant magic formula (16.291), the variation $\delta\mathbf{e}$ of the line interval under an infinitesimal coordinate transformation (16.285) generated by ϵ is

$$\delta\mathbf{e} = -\mathcal{L}_\Gamma \epsilon \mathbf{e} = -\epsilon \cdot (D_\Gamma \mathbf{e}) - D_\Gamma (\epsilon \cdot \mathbf{e}) = -d(\epsilon \cdot \mathbf{e}) - \frac{1}{2}[\Gamma, \epsilon \cdot \mathbf{e}] - \epsilon \cdot \mathbf{S} . \quad (16.293)$$

Alternatively, in terms of the torsion-free exterior derivative $\mathring{D}$,

$$\delta\mathbf{e} = -\mathcal{L}_\Gamma \epsilon \mathbf{e} = -\mathring{D}(\epsilon \cdot \mathbf{e}) - \frac{1}{2}[\epsilon \cdot \mathbf{K}, \mathbf{e}] = -d(\epsilon \cdot \mathbf{e}) - \frac{1}{2}[\mathring{\Gamma}, \epsilon \cdot \mathbf{e}] - \frac{1}{2}[\epsilon \cdot \mathbf{K}, \mathbf{e}] , \quad (16.294)$$

which is the forms version of equation (16.140) derived earlier in index notation.

The Lorentz connection Γ is a coordinate tensor but not a tetrad tensor, so the covariant magic formula (16.291) does not apply to the Lorentz connection. Rather, the variation $\delta\Gamma$ of the Lorentz connection follows from the difference

$$\delta D_\Gamma \mathbf{a} - D_\Gamma \delta \mathbf{a} = \delta (d\mathbf{a} + \frac{1}{2}[\Gamma, \mathbf{a}]) - (d\delta \mathbf{a} + \frac{1}{2}[\Gamma, \delta \mathbf{a}]) = \frac{1}{2}[\delta\Gamma, \mathbf{a}] . \quad (16.295)$$

Thus the variation $\delta\Gamma$ of the Lorentz connection under an infinitesimal coordinate transformation generated by ϵ satisfies

$$\frac{1}{2}[\delta\Gamma, \mathbf{a}] = -\mathcal{L}_\Gamma \epsilon D_\Gamma \mathbf{a} + D_\Gamma \mathcal{L}_\Gamma \epsilon \mathbf{a} = -\epsilon \cdot (D_\Gamma D_\Gamma \mathbf{a}) + D_\Gamma D_\Gamma (\epsilon \cdot \mathbf{a}) = -\frac{1}{2}\epsilon \cdot [\mathbf{R}, \mathbf{a}] + \frac{1}{2}[\mathbf{R}, \epsilon \cdot \mathbf{a}] = -\frac{1}{2}[\epsilon \cdot \mathbf{R}, \mathbf{a}] , \quad (16.296)$$

where $\mathbf{R}$ is the Riemann curvature bivector 2-form. Equation (16.295) holds for all multivector forms $\mathbf{a}$, so

$$\delta\Gamma = -\mathcal{L}_\Gamma \Gamma = -\epsilon \cdot \mathbf{R} . \quad (16.297)$$

which is the forms version of equation (16.143) derived earlier in index notation.

Inserting the variations (16.293) and (16.297) of the line interval $\mathbf{e}$ and Lorentz connection Γ into the

variation (16.250) of the matter action yields the variation of the action under an infinitesimal coordinate transformation (16.285) generated by the 1-form ϵ ,

$$\delta S_m = -I \int \left(d(\epsilon \cdot e) + \frac{1}{2}[\Gamma, \epsilon \cdot e] + \epsilon \cdot S \right) \wedge \overset{**}{T} + \overset{**}{\Sigma} \wedge (\epsilon \cdot R) . \quad (16.298)$$

Integrating the $d(\epsilon \cdot e) \wedge \overset{**}{T}$ term by parts, and rearranging the $\frac{1}{2}[\Gamma, \epsilon \cdot e] \wedge \overset{**}{T}$ term using the multivector triple-product relation (13.39), yields

$$\delta S_m = -I \oint (\epsilon \cdot e) \wedge \overset{**}{T} + I \int (\epsilon \cdot e) \wedge (d\overset{**}{T} + \frac{1}{2}[\Gamma, \overset{**}{T}]) - (\epsilon \cdot S) \wedge \overset{**}{T} - \overset{**}{\Sigma} \wedge (\epsilon \cdot R) . \quad (16.299)$$

Invariance of the action under coordinate transformations requires that the variation (16.299) must vanish for arbitrary choices of the 1-form ϵ vanishing on the initial and final hypersurfaces. Consequently the matter energy-momentum $\overset{**}{T}$ must satisfy the conservation equation

$$(\epsilon \cdot e) \wedge (d\overset{**}{T} + \frac{1}{2}[\Gamma, \overset{**}{T}]) - (\epsilon \cdot S) \wedge \overset{**}{T} - \overset{**}{\Sigma} \wedge (\epsilon \cdot R) = 0 . \quad (16.300)$$

Equivalently, in terms of the torsion-free connection $\overset{\circ}{\Gamma}$ and the contortion K ,

$$\boxed{(\epsilon \cdot e) \wedge (d\overset{**}{T} + \frac{1}{2}[\overset{\circ}{\Gamma}, \overset{**}{T}]) - \frac{1}{2}[\epsilon \cdot K, e] \wedge \overset{**}{T} - \overset{**}{\Sigma} \wedge (\epsilon \cdot R) = 0} . \quad (16.301)$$

I don't know a way to recast equations (16.300) or (16.301) in multivector forms notation with the arbitrary 1-form ϵ factored out, but in components equation (16.301) reduces to equation (16.146) derived earlier.

If the spin angular-momentum of a matter component vanishes, $\overset{**}{\Sigma} = 0$, then its energy-momentum is symmetric, equation (16.284), and then the energy-momentum conservation law (16.301) for that matter component simplifies to

$$d\overset{**}{T} + \frac{1}{2}[\overset{\circ}{\Gamma}, \overset{**}{T}] = 0 . \quad (16.302)$$

If the energy-momentum conservation law (16.300) is summed over all matter components, and the total spin angular-momentum $\overset{**}{\Sigma}$ and energy-momentum $\overset{**}{T}$ eliminated in favour of torsion S and curvature R using Hamilton's equations (16.252), then the law of conservation of total energy-momentum becomes

$$(\epsilon \cdot e) \wedge (d(e \wedge R) + \frac{1}{2}[\Gamma, e \wedge R]) - (\epsilon \cdot S) \wedge e \wedge R - e \wedge S \wedge (\epsilon \cdot R) = 0 , \quad (16.303)$$

which by the relation (16.287) rearranges to

$$(\epsilon \cdot e) \wedge (d(e \wedge R) + \frac{1}{2}[\Gamma, e \wedge R] - S \wedge R) = 0 . \quad (16.304)$$

Equation (16.304) is true for arbitrary infinitesimal ϵ , so the law of conservation of total energy-momentum is

$$d(e \wedge R) + \frac{1}{2}[\Gamma, e \wedge R] - S \wedge R = 0 , \quad (16.305)$$

which agrees with the contracted Bianchi identity (16.408b).

Exercise 16.13 Lie derivative of a form. Confirm from the definition (7.157) that the Lie derivative of a p -form is indeed given by Cartan's magic formula (16.288).

16.15 Space+time (3+1) split in multivector forms notation

As discussed in §16.5.8, when applied to fields, the super-Hamiltonian approach does not yield equal numbers of coordinates and momenta. The problem arises because symmetry under general coordinate transformations means that different configurations of fields are symmetrically equivalent. To permit manifest covariance, the super-Hamiltonian formalism is forced to admit more fields than there are physical degrees of freedom. As found previously with the electromagnetic field, §16.6.6, the solution to the problem is to break general covariance by splitting spacetime into separate space and time coordinates.

Executing a 3+1 split of the gravitational equations successfully, in the sense of achieving a balanced number of coordinates and momenta with the right number of physical degrees of freedom, is, unsurprisingly, a more complicated challenge than splitting the electromagnetic equations.

In splitting a multivector form $\mathbf{a}$ into time and space components, it is convenient to adopt the notation of §16.6.6, generalized to multivector-valued forms. A multivector p -form $\mathbf{a}$ splits into a component $\mathbf{a}_{\bar{t}}$ (subscripted $\bar{t}$) that represents all the coordinate time t parts of the form, and a component $\mathbf{a}_{\bar{\alpha}}$ (subscripted $\bar{\alpha}$) that represents the remaining spatial-coordinate components. The $\bar{t}$ and $\bar{\alpha}$ subscripts should be interpreted as labels, not indices. Thus a multivector p -form $\mathbf{a}$ splits as

$$\mathbf{a} = \mathbf{a}_{\bar{t}} + \mathbf{a}_{\bar{\alpha}} \equiv a_{A\bar{t}\Lambda} \gamma^A dx^{t\Lambda} + a_{A\bar{\alpha}\Lambda} \gamma^A dx^{A\Lambda}, \quad (16.306)$$

implicitly summed over distinct antisymmetric sequences of indices. Note that only the coordinates are being split: the Lorentz indices are *not* split into time and space parts. The option of also splitting the Lorentz indices is explored further in §16.16, equation (16.330).

The time component of a product (geometric product of multivectors, exterior product of forms) of any two multivector forms $\mathbf{a}$ and $\mathbf{b}$ satisfies

$$(\mathbf{a}\mathbf{b})_{\bar{t}} = \mathbf{a}_{\bar{t}}\mathbf{b}_{\bar{\alpha}} + \mathbf{a}_{\bar{\alpha}}\mathbf{b}_{\bar{t}}, \quad (16.307)$$

with no minus signs (minus signs from the antisymmetry of form indices cancel minus signs from commuting dt through a spatial form). The space component of a product of two multivector forms $\mathbf{a}$ and $\mathbf{b}$ satisfies

$$(\mathbf{a}\mathbf{b})_{\bar{\alpha}} = \mathbf{a}_{\bar{\alpha}}\mathbf{b}_{\bar{\alpha}}. \quad (16.308)$$

16.15.1 3+1 split of the gravitational Lagrangian in multivector forms notation

Consider first a 3+1 split of the standard gravitational Lagrangian (16.236). The gravitational coordinates in this case are the Lorentz connections $\mathbf{\Gamma}$, and their conjugate momentum are the components of the line interval $\mathbf{e}$. Actually, the momentum canonically conjugate to the Lorentz connection $\mathbf{\Gamma}$ in the standard

gravitational Lagrangian (16.236) is the area element e^2 , but as remarked at the end of §16.14.5, if all components of the area element are considered independent, then variation of the action with respect to all those components does not lead to general relativity. The fix is to consider the area element to be a product $e^2 = \frac{1}{2}e \wedge e$, in which the physical degrees of freedom are contained in the line element e .

After the space+time 3+1 split, the coordinates are the spatial components of the Lorentz connection $\Gamma_{\bar{\alpha}}$, which is a bivector 1-form with $6 \times 3 = 18$ components, and the momentum is the spatial line interval $e_{\bar{\alpha}}$, which is a vector 1-form with $4 \times 3 = 12$ components,

$$\Gamma_{\bar{\alpha}} = \Gamma_{\alpha} dx^{\alpha} = \Gamma_{kl\alpha} \gamma^k \wedge \gamma^l dx^{\alpha}, \quad (16.309a)$$

$$e_{\bar{\alpha}} = e_{\alpha} dx^{\alpha} = e_{k\alpha} \gamma^k dx^{\alpha}. \quad (16.309b)$$

The mismatch between the number 18 of components of the spatial connection $\Gamma_{\bar{\alpha}}$ and the number 12 of components of the spatial line interval $e_{\bar{\alpha}}$ is problematic. Despite the mismatch, it is useful to pursue the approach further, because it leads to a set of constraint equations commonly called the Gaussian and Hamiltonian constraints. These constraints are analogous to the electromagnetic constraint (16.77a), which has the property that, if it is satisfied initially, then conservation of electric charge guarantees it thereafter. Conservation of electric charge is a consequence of electromagnetic gauge symmetry. The Gaussian and Hamiltonian constraints are similarly constraint equations which, if satisfied initially, are guaranteed thereafter respectively by the conservation equations for spin angular-momentum and energy-momentum. These conservation equations are in turn a consequence of symmetries under Lorentz transformations and coordinate transformations.

The equations of motion (16.311) and constraint equations (16.312) follow directly from splitting equations (16.252) into time and space parts, but they can be derived at a more fundamental level by splitting the variation δS of the action into time and space parts. Splitting the variation δS_g of the gravitational action, equation (16.249), into time and space parts gives

$$\begin{aligned} \delta S_g = & -\frac{I}{8\pi} \oint_{t_i}^{t_f} (e^2 \wedge \delta \Gamma)_{\bar{t}} - \left[\oint (e^2 \wedge \delta \Gamma)_{\bar{\alpha}} \right]_{t_i}^{t_f} \\ & - \frac{I}{8\pi} \int (e \wedge S)_{\bar{t}} \wedge \delta \Gamma_{\bar{\alpha}} + \delta e_{\bar{\alpha}} \wedge (e \wedge R)_{\bar{t}} + (e \wedge S)_{\bar{\alpha}} \wedge \delta \Gamma_{\bar{t}} + \delta e_{\bar{t}} \wedge (e \wedge R)_{\bar{\alpha}}. \end{aligned} \quad (16.310)$$

The two surface integrals are respectively over the timelike spatial boundary of the 4-volume from t_i to t_f , and over the two spacelike caps of the 4-volume at t_i and t_f . Variation of the combined gravitational and matter actions with respect to the variations $\delta \Gamma_{\bar{\alpha}}$ and $\delta e_{\bar{\alpha}}$ of the spatial coordinates and momenta yields the equations of motion,

$$18 \text{ equations of motion: } (e \wedge S)_{\bar{t}} = 8\pi \Sigma_{\bar{t}}^{**}, \quad (16.311a)$$

$$12 \text{ equations of motion: } (e \wedge R)_{\bar{t}} = 8\pi T_{\bar{t}}^{**}. \quad (16.311b)$$

These are just the coordinate time components of the equations of motion (16.252). Variation with respect to the variations $\delta \Gamma_{\bar{t}}$ and $\delta e_{\bar{t}}$ of the time components of the coordinates and momenta yields the Gaussian

and Hamiltonian constraints,

$$6 \text{ Gaussian constraints: } (e \wedge S)_{\bar{\alpha}} = 8\pi \Sigma_{\bar{\alpha}}^{**}, \quad (16.312a)$$

$$4 \text{ Hamiltonian constraints: } (e \wedge R)_{\bar{\alpha}} = 8\pi T_{\bar{\alpha}}^{**}. \quad (16.312b)$$

These are the purely spatial coordinate components of the equations of motion (16.252). Whereas the equations of motion (16.311) involve derivatives with respect to time t , the constraint equations (16.312) involve no time derivatives. More explicitly, the equations of motion (16.311) are

$$18 \text{ equations of motion: } e_{\bar{\alpha}} \wedge (d_t e_{\bar{\alpha}} + \tfrac{1}{2}[\Gamma_{\bar{t}}, e_{\bar{\alpha}}] + d_{\alpha} e_{\bar{t}} + \tfrac{1}{2}[\Gamma_{\bar{\alpha}}, e_{\bar{t}}]) + e_{\bar{t}} \wedge S_{\bar{\alpha}} = 8\pi \Sigma_{\bar{t}}^{**}, \quad (16.313a)$$

$$12 \text{ equations of motion: } e_{\bar{\alpha}} \wedge (d_t \Gamma_{\bar{\alpha}} + d_{\alpha} \Gamma_{\bar{t}} + \tfrac{1}{2}[\Gamma_{\bar{\alpha}}, \Gamma_{\bar{t}}]) + e_{\bar{t}} \wedge R_{\bar{\alpha}} = 8\pi T_{\bar{t}}^{**}. \quad (16.313b)$$

The exterior time derivative here is the 1-form $d_t \equiv dt \partial / \partial t$. The equations of motion (16.313) are problematic not only because they remain unbalanced despite the 3+1 split, but also because the time derivative is not d_t but rather $e_{\bar{\alpha}} \wedge d_t$.

Both $\Gamma_{\bar{t}}$ and $e_{\bar{t}}$ can be treated as gauge variables: the 6 components of $\Gamma_{\bar{t}}$ can be adjusted arbitrarily by a Lorentz transformation; and the 4 components of $e_{\bar{t}}$ can be adjusted arbitrarily by a coordinate transformation. Thus the Gaussian and Hamiltonian constraint equations (16.312) can be interpreted as representing conserved Noether charges. The spin angular-momentum $\Sigma_{\bar{\alpha}}^{**}$ on the right hand side of the Gaussian constraint equation (16.312a) satisfies the conservation law (16.283). The energy-momentum $T_{\bar{\alpha}}^{**}$ on the right hand side of the Hamiltonian constraint equation (16.312b) satisfies the conservation law (16.301). The left hand sides of the Gaussian and Hamiltonian constraints satisfy corresponding conservation laws enforced by the contracted Bianchi identities (16.408). The Gaussian and Hamiltonian constraints are constraint equations in the sense commonly used by relativists: if the equations are arranged to be satisfied on the initial spatial hypersurface of constant time, then the conservation equations ensure that the equations will continue to be satisfied thereafter.

16.15.2 Conventional gravitational Hamiltonian

The conventional Hamiltonian is not the same as the super-Hamiltonian. Whereas the super-Hamiltonian approach is fully covariant, and the super-Hamiltonian (16.238) is a pseudoscalar 4-form, the conventional Hamiltonian approach picks out the coordinate time dimension as special, and the conventional Hamiltonian is the time component of a different pseudoscalar 4-form, equation (16.320).

Split into time and spatial components, the gravitational Lagrangian 4-form (16.236) is

$$L_g = -\frac{I}{8\pi} (e_{\bar{\alpha}}^2 \wedge (d_t \Gamma_{\bar{\alpha}} + d_{\alpha} \Gamma_{\bar{t}} + \tfrac{1}{2}[\Gamma_{\bar{\alpha}}, \Gamma_{\bar{t}}]) + e_{\bar{t}} \wedge (e \wedge R)_{\bar{\alpha}}) . \quad (16.314)$$

The $e_{\bar{\alpha}}^2 \wedge d_t \Gamma_{\bar{\alpha}}$ term in the Lagrangian (16.314) indicates that the momentum conjugate to the 18-component spatial connection $\Gamma_{\bar{\alpha}}$ is the 18-component spatial area element $e_{\bar{\alpha}}^2$. But, as discussed in the §16.15.1 above, the spatial area element $e_{\bar{\alpha}}^2$ has excess degrees of freedom compared to the 12-component line interval $e_{\bar{\alpha}}$. The fix adopted in §16.15.1 was to regard the spatial line interval $e_{\bar{\alpha}}$ rather than the spatial area element as the

conjugate momentum. Indeed, if all 18 degrees of freedom of the area element were treated as independent, then the Einstein equation (16.311b) would be replaced by an equation for $\mathbf{R}_{\bar{\alpha}}$ in place of $(\mathbf{e} \wedge \mathbf{R})_{\bar{\alpha}}$, and the result would not be general relativity, contradicting observation and experiment. To treat $\mathbf{\Gamma}_{\bar{\alpha}}$ and $\mathbf{e}_{\bar{\alpha}}$ as conjugate variables, the $\mathbf{e}_{\bar{\alpha}}^2 \wedge d_t \mathbf{\Gamma}_{\bar{\alpha}}$ term may be rewritten

$$\mathbf{e}_{\bar{\alpha}}^2 \wedge d_t \mathbf{\Gamma}_{\bar{\alpha}} = \frac{1}{2} \mathbf{e}_{\bar{\alpha}} \wedge (\mathbf{e}_{\bar{\alpha}} \wedge d_t \mathbf{\Gamma}_{\bar{\alpha}}) . \quad (16.315)$$

Equation (16.315) effectively replaces the time derivative d_t with $\mathbf{e}_{\bar{\alpha}} \wedge d_t$, consistent with the time derivative in the equations of motion (16.313). The remaining terms in the gravitational Lagrangian (16.314) rearrange as follows. The $d_{\alpha} \mathbf{\Gamma}_{\bar{t}}$ term integrates by parts to

$$\mathbf{e}_{\bar{\alpha}}^2 \wedge d_{\alpha} \mathbf{\Gamma}_{\bar{t}} = d_{\alpha} (\mathbf{e}_{\bar{\alpha}}^2 \wedge \mathbf{\Gamma}_{\bar{t}}) - (d\mathbf{e}^2)_{\bar{\alpha}} \wedge \mathbf{\Gamma}_{\bar{t}} . \quad (16.316)$$

The $\frac{1}{2}[\mathbf{\Gamma}_{\bar{\alpha}}, \mathbf{\Gamma}_{\bar{t}}]$ term rearranges by the multivector triple-product relation (13.39) to

$$\frac{1}{2} \mathbf{e}_{\bar{\alpha}}^2 \wedge [\mathbf{\Gamma}_{\bar{\alpha}}, \mathbf{\Gamma}_{\bar{t}}] = \frac{1}{2} [\mathbf{e}^2, \mathbf{\Gamma}]_{\bar{\alpha}} \wedge \mathbf{\Gamma}_{\bar{t}} = \frac{1}{2} [\mathbf{\Gamma}, \mathbf{e}^2]_{\bar{\alpha}} \wedge \mathbf{\Gamma}_{\bar{t}} . \quad (16.317)$$

The coefficients of the $\wedge \mathbf{\Gamma}_{\bar{t}}$ terms in equations (16.316) and (16.317) are

$$(-d\mathbf{e}^2 - \frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}^2])_{\bar{\alpha}} = (\mathbf{e} \wedge \mathbf{S})_{\bar{\alpha}} , \quad (16.318)$$

where $\mathbf{S}$ is the torsion defined by equation (16.214). The manipulations (16.315)–(16.318) bring the gravitational action to

$$S_g = -\frac{I}{8\pi} \oint_{t_i}^{t_f} \mathbf{e}_{\bar{\alpha}}^2 \wedge \mathbf{\Gamma}_{\bar{t}} - \frac{I}{8\pi} \int \frac{1}{2} \mathbf{e}_{\bar{\alpha}} \wedge (\mathbf{e}_{\bar{\alpha}} \wedge d_t \mathbf{\Gamma}_{\bar{\alpha}}) + (\mathbf{e} \wedge \mathbf{S})_{\bar{\alpha}} \wedge \mathbf{\Gamma}_{\bar{t}} + \mathbf{e}_{\bar{t}} \wedge (\mathbf{e} \wedge \mathbf{R})_{\bar{\alpha}} . \quad (16.319)$$

With the surface term discarded, the gravitational action (16.319) is in conventional Hamiltonian form $L_g = I(\mathbf{p} \wedge (\mathbf{e} \wedge d_t \mathbf{q})) - H_g$ with coordinates $\mathbf{q} \equiv \mathbf{\Gamma}_{\bar{\alpha}}$ and momenta $\mathbf{p} \equiv -\mathbf{e}_{\bar{\alpha}}/(16\pi)$, and a somewhat strange time derivative $\mathbf{e}_{\bar{\alpha}} \wedge d_t$. The conventional (not super-) Hamiltonian 4-form H_g is

$$H_g = \frac{I}{8\pi} ((\mathbf{e} \wedge \mathbf{S})_{\bar{\alpha}} \wedge \mathbf{\Gamma}_{\bar{t}} + \mathbf{e}_{\bar{t}} \wedge (\mathbf{e} \wedge \mathbf{R})_{\bar{\alpha}}) . \quad (16.320)$$

The conventional Hamiltonian (16.320) is a sum of the Gaussian and Hamiltonian constraint variables $(\mathbf{e} \wedge \mathbf{S})_{\bar{\alpha}}$ and $(\mathbf{e} \wedge \mathbf{R})_{\bar{\alpha}}$, equations (16.312), wedged with the gauge variables $\mathbf{\Gamma}_{\bar{t}}$ and $\mathbf{e}_{\bar{t}}$.

The Hamiltonian (16.320) is fine as a conventional Hamiltonian in which the coordinates and momenta are the 18-component spatial connection $\mathbf{\Gamma}_{\bar{\alpha}}$ and the 12-component spatial line interval $\mathbf{e}_{\bar{\alpha}}$. But the Hamiltonian cannot be satisfactory because it yields only 12 equations of motion (16.313b) for the 18 components of $\mathbf{\Gamma}_{\bar{\alpha}}$, and because the time derivative in those equations is $\mathbf{e}_{\bar{\alpha}} \wedge d_t$ rather than d_t . Ultimately, these problems stem from the fact that there remain redundant degrees of freedom in $\mathbf{\Gamma}_{\bar{\alpha}}$ despite the 3+1 split.

16.15.3 3+1 split of the alternative gravitational Lagrangian in multivector forms notation

A 3+1 split of the alternative Lagrangian (16.259) yields a more promising result: a balanced set of equations of motion, and a time derivative that is just $d_t \equiv \partial/\partial t dt$ as opposed to $\mathbf{e}_{\bar{\alpha}} \wedge d_t$. In the alternative

Lagrangian, the gravitational coordinates are the line interval e , and their conjugate momenta are π defined by equation (16.255). After the 3+1 split, the coordinates are the spatial components $e_{\bar{\alpha}}$ of the line interval, which is a vector 1-form with $4 \times 3 = 12$ components, while the momenta are the spatial components $\pi_{\bar{\alpha}}$, which is a trivector 2-form also with $4 \times 3 = 12$ components.

Once again, the equations of motion (16.322) and constraints and identities (16.326) follow directly from splitting equations (16.268) into time and space parts, but they can be derived more fundamentally by splitting the variation δS of the action into time and space parts. Splitting the variation δS_g of the alternative gravitational action (16.261) into time and space parts gives

$$\delta S'_g = \frac{I}{8\pi} \oint_{t_i}^{t_f} (\pi \wedge \delta e)_{\bar{t}} + \frac{I}{8\pi} \left[\oint (\pi \wedge \delta e)_{\bar{\alpha}} \right]_{t_i}^{t_f} + \frac{I}{8\pi} \int \delta \pi_{\bar{\alpha}} \wedge S_{\bar{t}} - \Pi_{\bar{\alpha}} \wedge \delta e_{\bar{t}} + \delta \pi_{\bar{t}} \wedge S_{\bar{\alpha}} - \delta e_{\bar{\alpha}} \wedge \Pi_{\bar{t}}. \quad (16.321)$$

Variation of the combined gravitational and matter actions with respect to the variations $\delta e_{\bar{\alpha}}$ and $\delta \pi_{\bar{\alpha}}$ of the spatial coordinates and momenta yields $12 + 12 = 24$ equations of motion involving time derivatives,

$$12 \text{ equations of motion: } S_{\bar{t}} = 8\pi \tilde{\Sigma}_{\bar{t}}, \quad (16.322a)$$

$$12 \text{ equations of motion: } \Pi_{\bar{t}} = 8\pi \tilde{T}_{\bar{t}}. \quad (16.322b)$$

Variation of the action with respect to the variations $\delta e_{\bar{t}}$ and $\delta \pi_{\bar{t}}$ of the time components of the coordinates and momenta yields 6 identities and 10 constraint equations involving only spatial derivatives,

$$6 \text{ Gaussian constraints and 6 identities: } S_{\bar{\alpha}} = 8\pi \tilde{\Sigma}_{\bar{\alpha}}, \quad (16.323a)$$

$$4 \text{ Hamiltonian constraints: } \Pi_{\bar{\alpha}} = 8\pi \tilde{T}_{\bar{\alpha}}. \quad (16.323b)$$

The Gaussian constraints are the subset of equations (16.323a) comprising

$$6 \text{ Gaussian constraints: } (e \wedge S)_{\bar{\alpha}} = 8\pi (e \wedge \tilde{\Sigma})_{\bar{\alpha}}. \quad (16.324)$$

More explicitly, the equations of motion (16.322) are

$$12 \text{ equations of motion: } d_t e_{\bar{\alpha}} + \frac{1}{2} [\Gamma_{\bar{t}}, e_{\bar{\alpha}}] + d_{\alpha} e_{\bar{t}} + \frac{1}{2} [\Gamma_{\bar{\alpha}}, e_{\bar{t}}] = 8\pi \tilde{\Sigma}_{\bar{t}}, \quad (16.325a)$$

$$12 \text{ equations of motion: } d_t \pi_{\bar{\alpha}} + \frac{1}{2} [\Gamma_{\bar{t}}, \pi_{\bar{\alpha}}] + d_{\alpha} \pi_{\bar{t}} + \frac{1}{2} [\Gamma_{\bar{\alpha}}, \pi_{\bar{t}}] - \frac{1}{4} (e \wedge [\Gamma, \Gamma])_{\bar{t}} = 8\pi \tilde{T}_{\bar{t}}, \quad (16.325b)$$

and the constraints and identities (16.323) are

$$6 \text{ Gaussian constraints and 6 identities: } (de + \frac{1}{2} [\Gamma, e])_{\bar{\alpha}} = 8\pi \tilde{\Sigma}_{\bar{\alpha}}, \quad (16.326a)$$

$$4 \text{ Hamiltonian constraints: } (d\pi + \frac{1}{2} [\Gamma, \pi] - \frac{1}{4} e \wedge [\Gamma, \Gamma])_{\bar{\alpha}} = 8\pi \tilde{T}_{\bar{\alpha}}. \quad (16.326b)$$

The Gaussian constraints (16.324) are

$$6 \text{ Gaussian constraints: } (de^2 - e \cdot \pi)_{\bar{\alpha}} = 8\pi (e \wedge \tilde{\Sigma})_{\bar{\alpha}}. \quad (16.327)$$

Equations (16.325a) comprise 12 equations of motion for the 12 coordinates $e_{\bar{\alpha}}$, while equations (16.325b) comprise 12 equations of motion for the 12 momenta $\pi_{\bar{\alpha}}$. The equations of motion (16.325) do not suffer from the peculiarities of the earlier equations of motion (16.313): the time evolution operator is $d_t \equiv dt \partial / \partial t$; and the number of equations of motion matches the number of dynamical variables.

16.15.4 Gravomagnetic field

To solve the system of gravitational equations (16.325) and (16.326), it is necessary to isolate the 6 identities from the 6 Gaussian constraints in equation (16.326a). Whereas constraint equations can be discarded after being imposed in the initial conditions (because conservation laws ensure their ongoing satisfaction during subsequent evolution), identities must be calculated at each time step. The 6 identities are equations (16.344) below.

The spatial Lorentz connection $\Gamma_{\bar{\alpha}}$ has 18 components, whereas its contraction the spatial momentum $\pi_{\bar{\alpha}} = -(e \wedge \Gamma)_{\bar{\alpha}}$, equation (16.255), has only 12. The extra 6 components of the spatial Lorentz connection are redundant. The 6 identities (16.326a) can be interpreted as defining the 6 redundant components of the spatial Lorentz connection $\Gamma_{\bar{\alpha}}$ in terms of spatial exterior derivatives $d_{\alpha}e_{\bar{\alpha}}$ of the spatial line interval $e_{\bar{\alpha}}$. These 6 redundant components, denoted $\mathring{\Gamma}_{\bar{\alpha}}$ (with a slash, symbolizing that it is the trace-free part of $\Gamma_{\bar{\alpha}}$, equation (16.334)), can be called the **gravomagnetic field**, equation (16.336), since the situation is analogous to that in electromagnetism, where the 3-component magnetic field is redundant because it can be replaced by the spatial exterior derivative of the spatial components of the electromagnetic potential, equation (16.77b). Although the gravomagnetic field $\mathring{\Gamma}_{\bar{\alpha}}$ is redundant, it must still be calculated because the equations of motion (16.325) depend on it.

To isolate the gravomagnetic field from the other components of the spatial Lorentz connection $\Gamma_{\bar{\alpha}}$, let γ^0 denote the future-pointing tetrad vector normal to spatial hypersurfaces. The ADM formalism, Chapter 17, makes the gauge choice of imposing that the tetrad time vector γ^0 be normal to spatial hypersurfaces. Choosing the vector γ^0 to be normal to spatial hypersurfaces is equivalent to imposing the ADM conditions (17.7) on the vierbein and its inverse,

$$e^0_{\alpha} = e_a{}^t = 0 . \quad (16.328)$$

However, in the present context the vector γ^0 should be interpreted as the future-pointing normal to spatial hypersurfaces, regardless of whether it happens also to be the tetrad time vector. The normal to spatial hypersurfaces is related by a Lorentz boost to any arbitrary tetrad time vector. In what follows, the vector γ^0 will be referred to as the tetrad time vector, and 0 as the tetrad time index, on the grounds that γ^0 is timelike while the three vectors γ^a , $a = 1, 2, 3$, orthogonal to it are spacelike, regardless of whether γ^0 is or is not the chosen tetrad time axis. The point of requiring γ^0 to be normal to spatial hypersurfaces is that spatial tetrad and coordinate indices can then be transformed freely between each other using the spatial vierbein e^a_{α} and its inverse $e_a{}^{\alpha}$,

$$e^a_{\alpha}a_a = e^k_{\alpha}a_k = a_{\alpha} , \quad e_a{}^{\alpha}a_{\alpha} = e_a{}^{\kappa}a_{\kappa} = a_a . \quad (16.329)$$

The extension of the sum over 3 spatial indices a (or α) to 4 spacetime indices k (or κ) in equations (16.329) is thanks to the conditions (16.328), which hold as long as γ^0 is normal to the spatial hypersurface, as is being required.

It is convenient to extend the 3+1 form-splitting notation (16.306) to a double 3+1 split in which tetrad (Lorentz) indices as well as coordinate indices are split out. Thus a multivector form $\mathbf{a}$ splits into 4 components $\mathbf{a}_{\bar{0}\bar{t}}$, $\mathbf{a}_{\bar{0}\bar{\alpha}}$, $\mathbf{a}_{\bar{\alpha}\bar{t}}$, and $\mathbf{a}_{\bar{\alpha}\bar{\alpha}}$ that represent respectively the time-time, time-space, space-time, and space-

space components of the multivector form,

$$\mathbf{a} = \mathbf{a}_{\bar{0}\bar{t}} + \mathbf{a}_{\bar{0}\bar{\alpha}} + \mathbf{a}_{\bar{a}\bar{t}} + \mathbf{a}_{\bar{a}\bar{\alpha}} \equiv a_{0A\bar{t}\bar{\Lambda}} \gamma^0 \wedge \gamma^A dx^{t\bar{\Lambda}} + a_{0A\bar{\Lambda}} \gamma^0 \wedge \gamma^A dx^{p\bar{\Lambda}} + a_{A\bar{t}\bar{\Lambda}} \gamma^A dx^{t\bar{\Lambda}} + a_{A\bar{\Lambda}} \gamma^A dx^{p\bar{\Lambda}}, \quad (16.330)$$

implicitly summed over distinct antisymmetric sequences of tetrad and coordinate indices A and $\bar{\Lambda}$. In the notation (16.330), the ADM gauge condition (16.328) is

$$\mathbf{e}_{\bar{0}\bar{\alpha}} = 0. \quad (16.331)$$

The 18-component spatial Lorentz connection $\Gamma_{\bar{\alpha}}$ splits into 9+9 components,

$$\Gamma_{\bar{\alpha}} = \Gamma_{\bar{0}\bar{\alpha}} + \Gamma_{\bar{a}\bar{\alpha}} = (\Gamma_{0b\alpha} \gamma^0 + \Gamma_{ab\alpha} \gamma^a) \wedge \gamma^b dx^\alpha. \quad (16.332)$$

The time $\bar{0}$ tetrad components $\Gamma_{0b\alpha}$ are part of what is commonly called the extrinsic curvature, $K_{b\kappa} \equiv \Gamma_{b0\kappa} = -\Gamma_{0b\kappa}$, §17.1.4, while the spatial $\bar{a}$ tetrad components $\Gamma_{ab\alpha}$ are part of what is referred to elsewhere in this book as the restricted connection $\hat{\Gamma}_{ab\kappa} \equiv \Gamma_{ab\kappa}$, §17.1.5. The 9-component extrinsic curvature $\Gamma_{\bar{0}\bar{\alpha}}$ is invertibly related to the 9-component momentum, $\pi_{\bar{0}\bar{\alpha}}$

$$\pi_{\bar{0}\bar{\alpha}} = \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \Gamma_{\bar{0}\bar{\alpha}}, \quad (16.333)$$

which holds thanks to the ADM condition (16.331). The 9-component all-spatial Lorentz connection $\Gamma_{\bar{a}\bar{\alpha}}$ resolves into a 3-component spatial trace (16.335), and a 6-component trace-free part, the gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$, equation (16.336),

$$\Gamma_{\bar{a}\bar{\alpha}} = -\frac{1}{2} \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \pi_{\bar{a}\bar{\alpha}}^{**\top} + \mathbf{F}_{\bar{a}\bar{\alpha}}. \quad (16.334)$$

The slashed notation $\mathbf{F}_{\bar{a}\bar{\alpha}}$ for the 6-component gravomagnetic field symbolizes that it is trace-free, and also that it is the part of the 18-component spatial Lorentz connection $\Gamma_{\bar{\alpha}}$ not contained in the 12-component spatial momentum $\pi_{\bar{\alpha}}$. The 3-component spatial trace of the all-spatial Lorentz connection is

$$\text{Tr } \Gamma_{\bar{a}\bar{\alpha}} \equiv e^{a\alpha} \Gamma_{ab\alpha} \gamma^b = \Gamma_{b\alpha}^\alpha \gamma^b = -\pi_{\bar{a}\bar{\alpha}}^{**\top}, \quad (16.335)$$

where the vector 0-form $\pi_{\bar{a}\bar{\alpha}}^{**\top}$ is the transpose of the spatial double dual of the all-spatial momentum $\pi_{\bar{a}\bar{\alpha}}$. The spatial trace (16.335) is to be distinguished from the spacetime trace (16.258); the latter includes an additional contribution from $\Gamma_{\bar{0}}$. The 3-component all-spatial momentum $\pi_{\bar{a}\bar{\alpha}}$ may be called the BSSN variable, because the equation of motion for this variable is the key equation that distinguishes the BSSN formalism, §16.16.2, from the ADM formalism, §16.16.1.

The 6-component trace-free part of the all-spatial Lorentz connection defines the gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$,

$$\mathbf{F}_{\bar{a}\bar{\alpha}} \equiv \Gamma_{\bar{a}\bar{\alpha}} + \frac{1}{2} \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \pi_{\bar{a}\bar{\alpha}}^{**\top} = \mathbf{F}_{ab\alpha} \gamma^a \wedge \gamma^b dx^\alpha = (\Gamma_{ab\alpha} - e_{a\alpha} \Gamma_{b\gamma}^\gamma) \gamma^a \wedge \gamma^b dx^\alpha. \quad (16.336)$$

The 6 identities that define the gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$ are part of the 12-component expression (16.323a) for the spatial torsion $\mathbf{S}_{\bar{\alpha}}$ in terms of the spatial spin angular-momentum $\tilde{\mathbf{S}}_{\bar{\alpha}}$. The 12-component spatial torsion $\mathbf{S}_{\bar{\alpha}}$ splits into 3+9 components (the minus sign conforms to Cartan's convention, equation (16.211)),

$$\mathbf{S}_{\bar{\alpha}} = \mathbf{S}_{\bar{0}\bar{\alpha}} + \mathbf{S}_{\bar{a}\bar{\alpha}} = - (S_{0\alpha\beta} \gamma^0 + S_{a\alpha\beta} \gamma^a) d^2 x^{\alpha\beta}. \quad (16.337)$$

The 3-component time $\bar{0}$ tetrad part $\mathbf{S}_{\bar{0}\bar{\alpha}}$ is invertibly related to the 3-component $(\mathbf{e} \wedge \mathbf{S})_{\bar{0}\bar{\alpha}}$,

$$(\mathbf{e} \wedge \mathbf{S})_{\bar{0}\bar{\alpha}} = \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \mathbf{S}_{\bar{0}\bar{\alpha}} , \quad (16.338)$$

so the equation for $\mathbf{S}_{\bar{0}\bar{\alpha}}$ is part of the Gaussian constraints (16.324). The 9-component all-spatial torsion $\mathbf{S}_{\bar{a}\bar{\alpha}}$ resolves into a 3-component trace (16.340) and a 6-component trace-free part $\mathcal{S}_{\bar{a}\bar{\alpha}}$ (the slashed notation $\mathcal{S}_{\bar{a}\bar{\alpha}}$ symbolizing that it is trace-free),

$$\mathbf{S}_{\bar{a}\bar{\alpha}} = \frac{1}{2} \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \mathbf{s}_{\bar{a}\bar{\alpha}}^{**} + \mathcal{S}_{\bar{a}\bar{\alpha}} . \quad (16.339)$$

The 3-component spatial trace of the all-spatial torsion is, compare equation (16.131a) (the minus signs are Cartan, again),

$$\text{Tr } \mathbf{S}_{\bar{a}\bar{\alpha}} \equiv -e^{a\beta} S_{a\alpha\beta} dx^\alpha = -S_{\alpha\beta}^\beta dx^\alpha = -\mathbf{s}_{\bar{a}\bar{\alpha}}^{**\top} , \quad (16.340)$$

where the scalar 1-form $\mathbf{s}_{\bar{a}\bar{\alpha}}^{**\top}$ is the transpose of the spatial double dual of the spatial bivector 3-form $\mathbf{s}_{\bar{a}\bar{\alpha}}$ defined by

$$\mathbf{s}_{\bar{a}\bar{\alpha}} \equiv (\mathbf{e} \wedge \mathbf{S})_{\bar{a}\bar{\alpha}} . \quad (16.341)$$

The equation for $\mathbf{s}_{\bar{a}\bar{\alpha}}$ is part of the Gaussian constraints (16.324). The components of the double dual $\mathbf{s}^{**}$ and its transpose $\mathbf{s}^{**\top}$ are

$$\mathbf{s}_{\bar{a}\bar{\alpha}}^{**} = S_{ab}^b \gamma^a , \quad \mathbf{s}_{\bar{a}\bar{\alpha}}^{**\top} = S_{\alpha b}^b dx^\alpha . \quad (16.342)$$

The 6-component trace-free part $\mathcal{S}_{\bar{\alpha}}$ of the all-spatial torsion is, equation (16.339),

$$\mathcal{S}_{\bar{a}\bar{\alpha}} = \mathbf{S}_{\bar{a}\bar{\alpha}} - \frac{1}{2} \mathbf{e}_{\bar{a}\bar{\alpha}} \wedge \mathbf{s}_{\bar{a}\bar{\alpha}}^{**\top} = -\mathcal{S}_{a\alpha\beta} \gamma^a d^2 x^{\alpha\beta} = -(S_{a\alpha\beta} + e_{a\alpha} S_{\beta\gamma}^\gamma) \gamma^a d^2 x^{\alpha\beta} . \quad (16.343)$$

The 6 identities in equations (16.323a) are, finally,

$$6 \text{ identities: } \mathcal{S}_{\bar{a}\bar{\alpha}} = 8\pi \tilde{\mathbf{Z}}_{\bar{a}\bar{\alpha}} . \quad (16.344)$$

More explicitly, the 6 identities (16.344) are

$$6 \text{ identities: } (\mathfrak{d}\mathbf{e} + \frac{1}{2}[\mathbf{F}, \mathbf{e}])_{\bar{a}\bar{\alpha}} = 8\pi \tilde{\mathbf{Z}}_{\bar{a}\bar{\alpha}} , \quad (16.345)$$

where $\mathfrak{d}\mathbf{e}_{\bar{a}\bar{\alpha}}$ is the trace-free part of the all-spatial exterior derivative $d\mathbf{e}_{\bar{a}\bar{\alpha}}$ of the line interval. Equation (16.344) defines the gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$ in term of the spin angular-momentum $\tilde{\mathbf{Z}}_{\bar{a}\bar{\alpha}}$ and spatial derivatives of the line interval. Note that $[\mathbf{F}, \mathbf{e}]_{\bar{a}\bar{\alpha}}$ is invertibly related to $\mathbf{F}_{\bar{a}\bar{\alpha}}$,

$$\frac{1}{2}[\mathbf{F}, \mathbf{e}]_{\bar{a}\bar{\alpha}} = \mathbf{F}_{\bar{a}\bar{\alpha}} \cdot \mathbf{e}_{\bar{a}\bar{\alpha}} = \mathbf{F}_{ab\alpha} (e^b{}_\beta \gamma^a - e^a{}_\beta \gamma^b) d^2 x^{\alpha\beta} = -2\mathbf{F}_{a\alpha\beta} \gamma^a d^2 x^{\alpha\beta} . \quad (16.346)$$

16.15.5 Alternative conventional Hamiltonian

The conventional Hamiltonian is not the same as the super-Hamiltonian. The conventional Hamiltonian was discussed for the standard gravitational Lagrangian (16.236) in §16.15.2. The present section considers the conventional Hamiltonian for the alternative gravitational Lagrangian (16.259).

Splitting the alternative Lagrangian L'_g , equation (16.259), into time and space components, and rearranging along lines similar to those leading to the gravitational action (16.319), brings the alternative gravitational action S'_g to

$$S'_g = \frac{I}{8\pi} \oint_{t_i}^{t_f} \pi_{\bar{\alpha}} \wedge e_{\bar{t}} + \frac{I}{8\pi} \int \pi_{\bar{\alpha}} \wedge d_t e_{\bar{\alpha}} + \pi_{\bar{t}} \wedge S_{\bar{\alpha}} - \Pi_{\bar{\alpha}} \wedge e_{\bar{t}} . \quad (16.347)$$

With the surface term discarded, the gravitational action (16.347) is in conventional Hamiltonian form with coordinates $e_{\bar{\alpha}}$ and momenta $\pi_{\bar{\alpha}}/(8\pi)$. The alternative conventional (not super-) Hamiltonian H'_g is

$$H'_g = \frac{I}{8\pi} (-\pi_{\bar{t}} \wedge S_{\bar{\alpha}} + \Pi_{\bar{\alpha}} \wedge e_{\bar{t}}) . \quad (16.348)$$

Part of deriving equation (16.348) involves proving that

$$(e_{\bar{\alpha}} \wedge \Gamma_{\bar{t}}) \wedge (e \cdot \Gamma)_{\bar{\alpha}} = (e \wedge \Gamma)_{\bar{\alpha}} \wedge (e_{\bar{\alpha}} \cdot \Gamma_{\bar{t}}) . \quad (16.349)$$

The alternative conventional Hamiltonian (16.348) is a sum of constraint and identities variables $S_{\bar{\alpha}}$ and $\Pi_{\bar{\alpha}}$, equations (16.323), wedged with time components $e_{\bar{t}}$ and $\pi_{\bar{t}}$ of the coordinates and momenta. Whereas the standard conventional Hamiltonian (16.320) depended only on constraint and gauge variables, the alternative conventional Hamiltonian (16.348) depends in addition on the gravomagnetic field $\mathfrak{F}_{\bar{\alpha}\bar{\alpha}}$.

In contrast to the conventional Hamiltonian (16.320), the alternative conventional Hamiltonian (16.348) accomplishes the goal of a balanced number, 12 each, of coordinates and momenta.

16.15.6 WEBB formalism

The system of 24 Hamiltonian equations of motion (16.325) is a set of coupled first-order partial differential equations. The system is integrable, but integrability does not guarantee that their numerical integration is stable. A set of coupled partial differential equations is numerically stable if they are strongly hyperbolic, as described in §17.7.1.

The thing that complicates the analysis of the hyperbolicity of the equations of motion (16.325) is that they involve not only the coordinates and momenta $e_{\bar{\alpha}}$ and $\pi_{\bar{\alpha}}$ and their first derivatives, but also the gravomagnetic field $\mathfrak{F}_{\bar{\alpha}\bar{\alpha}}$, which itself depends on spatial derivatives $d_{\bar{\alpha}}$ of the coordinates $e_{\bar{\alpha}}$, equation (16.345). The term $d_{\bar{\alpha}} \pi_{\bar{t}}$ in the Einstein equations (16.325b) then includes some second-order spatial derivatives of $e_{\bar{\alpha}}$, while the terms $\frac{1}{2}[\Gamma_{\bar{\alpha}}, \pi_{\bar{t}}]$ and $\frac{1}{4}e_{\bar{t}} \wedge [\Gamma_{\bar{\alpha}}, \Gamma_{\bar{\alpha}}]$ include terms quadratic in spatial derivatives of $e_{\bar{\alpha}}$.

The difficulty can be overcome by promoting the gravomagnetic field $\mathfrak{F}_{\bar{\alpha}\bar{\alpha}}$ to a set of 6 independent variables governed by their own equation of motion. The operation of promoting derivatives of variables to independent variables and enlarging the system of differential equations is called **prolongation**. The system obtained by prolonging the gravomagnetic field is the WEBB formalism (Buchman and Bardeen, 2005), a system of tetrad-based equations proposed by Buchman and Bardeen (2003) based on the work of Estabrook, Robinson, and Wahlquist (1997). Buchman and Bardeen (2003) prove that the WEBB system is strongly hyperbolic for at least some prescriptions for the gauge variables $e_{\bar{t}}$ and $\Gamma_{\bar{t}}$.

The 6 equations of motion governing the prolonged gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$ are

$$6 \text{ equations of motion: } (d\mathcal{S})_{\bar{a}\bar{t}} = 8\pi(d\tilde{\mathbf{Z}})_{\bar{a}\bar{t}}. \quad (16.350)$$

Since the second exterior derivative vanishes identically, $d^2 = 0$, and the trace-free all-spatial torsion is given by the left hand side of equation (16.345), the equations of motion (16.350) reduce to

$$6 \text{ equations of motion: } (d\frac{1}{2}[e, \mathbf{F}])_{\bar{a}\bar{t}} = 8\pi(d\tilde{\mathbf{Z}})_{\bar{a}\bar{t}}. \quad (16.351)$$

The original 6 identities (16.344) become constraints, because although they must be arranged to be satisfied on the initial spatial hypersurface, they are guaranteed thereafter by the Bianchi identity (16.405a),

$$6 \text{ gravomagnetic constraints: } \mathcal{S}_{\bar{a}\bar{\alpha}} = 8\pi\tilde{\mathbf{Z}}_{\bar{a}\bar{\alpha}}. \quad (16.352)$$

In all, after the gravomagnetic field is prolonged, the original 40 Hamiltonian equations (16.322) and (16.323) become 46 equations, consisting of 30 equations of motion, 16 constraints, and zero identities.

16.16 ADM gauge condition

Section 16.15.1 rejected the possibility of treating the 18-component spatial connection $\mathbf{\Gamma}_{\bar{\alpha}}$ as the gravitational coordinates, and the 18-component spatial area element $e_{\bar{\alpha}}^2$ as their conjugate momenta, on the grounds that the area element contains excess degrees of freedom compared to the 12-component spatial line interval $e_{\bar{\alpha}}$. However, the idea of working with $\mathbf{\Gamma}_{\bar{\alpha}}$ and $e_{\bar{\alpha}}^2$, as opposed to $e_{\bar{\alpha}}$ and $\pi_{\bar{\alpha}}$, is attractive, firstly because in gauge theories such as electromagnetism the coordinates are the connections $\mathbf{A}$, §16.6, and the (Lorentz) connections of gravity are $\mathbf{\Gamma}$, and secondly because black hole thermodynamics points to area as the thing that is somehow quantized in general relativity.

One way to reduce the excess degrees of freedom in the area element is to impose gauge conditions on the spatial line interval $e_{\bar{\alpha}}$. A natural strategy is to impose the 3-component ADM gauge condition $e_{\bar{0}\bar{\alpha}} = 0$, which was invoked earlier, equation (16.331), to separate out the 6 identities of the Hamiltonian system of equations, §16.15.4. The gauge choice (16.331) is the starting point of the ADM formalism, Chapter 17, and is carried over into the BSSN formalism, §17.8. The gauge choice (16.331) is also a basic ingredient of Loop Quantum Gravity, §??.

The ADM gauge condition (16.331) reduces the number of degrees of freedom of the spatial line interval $e_{\bar{\alpha}}$ from 12 to $3 \times 3 = 9$, and of the spatial area element $e_{\bar{\alpha}}^2$ from 18 to the same number, $3 \times 3 = 9$. The 9 components of the spatial line interval and spatial area element subject to the ADM gauge condition are invertibly related to each other. The spatial area element $e_{\bar{a}\bar{\alpha}}^2$ is the 9-component bivector 2-form

$$e_{\bar{a}\bar{\alpha}}^2 = \frac{1}{2} (e \wedge e)_{\bar{a}\bar{\alpha}} = 2 e_{a\alpha} e_{b\beta} \gamma^a \wedge \gamma^b dx^{\alpha\beta}. \quad (16.353)$$

The momenta conjugate to the spatial area element $e_{\bar{a}\bar{\alpha}}^2$ are the $3 \times 3 = 9$ components of the spatial Lorentz connections $\mathbf{\Gamma}_{\bar{0}\bar{\alpha}}$ with one Lorentz index the tetrad time index 0, also called (minus) the extrinsic curvature, §17.1.4,

$$\mathbf{\Gamma}_{\bar{0}\bar{\alpha}} = \Gamma_{0a\alpha} \gamma^0 \wedge \gamma^a dx^{\alpha}. \quad (16.354)$$

It looks as though the goal of having the coordinates and conjugate momenta be the connection $\Gamma_{\bar{\alpha}}$ and area element $e_{\bar{\alpha}}^2$ has been achieved, but notice this success has been won by trickery. The ADM gauge choice (16.331) is a condition on the line element e , not on the area element e^2 . Imposing the ADM gauge condition still requires that the area element be a product $e^2 = \frac{1}{2}e \wedge e$ of the line element e .

Double-splitting the variation δS_g of the gravitational action, equation (16.249) into time and space parts gives

$$\begin{aligned} \delta S_g = & -\frac{I}{8\pi} \oint_{t_i}^{t_f} (e^2 \wedge \delta \Gamma)_{\bar{0}\bar{t}} - \left[\oint (e^2 \wedge \delta \Gamma)_{\bar{0}\bar{\alpha}} \right]_{t_i}^{t_f} \\ & - \frac{I}{8\pi} \int (e \wedge S)_{\bar{0}\bar{t}} \wedge \delta \Gamma_{\bar{a}\bar{\alpha}} + (e \wedge S)_{\bar{a}\bar{t}} \wedge \delta \Gamma_{\bar{0}\bar{\alpha}} + (e \wedge S)_{\bar{0}\bar{\alpha}} \wedge \delta \Gamma_{\bar{a}\bar{t}} + (e \wedge S)_{\bar{a}\bar{\alpha}} \wedge \delta \Gamma_{\bar{0}\bar{t}} \\ & + \delta e_{\bar{a}\bar{\alpha}} \wedge (e \wedge R)_{\bar{0}\bar{t}} + \delta e_{\bar{0}\bar{\alpha}} \wedge (e \wedge R)_{\bar{a}\bar{t}} + \delta e_{\bar{a}\bar{t}} \wedge (e \wedge R)_{\bar{0}\bar{\alpha}} + \delta e_{\bar{0}\bar{t}} \wedge (e \wedge R)_{\bar{a}\bar{\alpha}}. \end{aligned} \quad (16.355)$$

Variation of the combined gravitational and matter actions with respect to the variations $\delta \Gamma_{\bar{0}\bar{\alpha}}$ and $\delta e_{\bar{a}\bar{\alpha}}$ yields $9 + 9 = 18$ equations of motion for the area element $e_{\bar{a}\bar{\alpha}}^2$ and their conjugate momenta $\Gamma_{\bar{0}\bar{\alpha}}$,

$$9 \text{ equations of motion: } (e \wedge S)_{\bar{a}\bar{t}} = 8\pi \Sigma_{\bar{a}\bar{t}}^{**}, \quad (16.356a)$$

$$9 \text{ equations of motion: } (e \wedge R)_{\bar{0}\bar{t}} = 8\pi T_{\bar{0}\bar{t}}^{**}, \quad (16.356b)$$

while variation with respect to $\delta e_{\bar{0}\bar{\alpha}}$ yields the 3 equations of motion

$$3 \text{ equations of motion: } (e \wedge R)_{\bar{a}\bar{t}} = 8\pi T_{\bar{a}\bar{t}}^{**}. \quad (16.357)$$

Note that the ADM gauge condition $e_{\bar{0}\bar{\alpha}} = 0$ is a gauge condition, fixed after equations of motion are derived, so it is correct to vary $e_{\bar{0}\bar{\alpha}}$ in the action, leading to equation (16.357). Explicitly, the equations of motion (16.356) and (16.357) are, similarly to equations (16.313),

$$9 \text{ eqs. of motion: } -d_t e_{\bar{a}\bar{\alpha}}^2 + \frac{1}{2}[\Gamma_{\bar{a}\bar{t}}, e_{\bar{a}\bar{\alpha}}^2] - e_{\bar{a}\bar{\alpha}} \wedge (d_{\bar{\alpha}} e_{\bar{a}\bar{t}} + \frac{1}{2}[\Gamma_{\bar{a}\bar{\alpha}}, e_{\bar{a}\bar{t}}]) + e_{\bar{a}\bar{t}} \wedge S_{\bar{a}\bar{\alpha}} = 8\pi \Sigma_{\bar{a}\bar{t}}^{**}, \quad (16.358a)$$

$$9 \text{ eqs. of motion: } e_{\bar{a}\bar{\alpha}} \wedge (d_t \Gamma_{\bar{0}\bar{\alpha}} + \frac{1}{2}[\Gamma_{\bar{a}\bar{t}}, \Gamma_{\bar{0}\bar{\alpha}}] + d_{\bar{\alpha}} \Gamma_{\bar{0}\bar{t}} + \frac{1}{2}[\Gamma_{\bar{a}\bar{\alpha}}, \Gamma_{\bar{0}\bar{t}}]) + e_{\bar{a}\bar{t}} \wedge R_{\bar{0}\bar{\alpha}} = 8\pi T_{\bar{0}\bar{t}}^{**}, \quad (16.358b)$$

$$3 \text{ eqs. of motion: } e_{\bar{a}\bar{\alpha}} \wedge (d_t \Gamma_{\bar{a}\bar{\alpha}} + d_{\bar{\alpha}} \Gamma_{\bar{t}} + \frac{1}{2}[\Gamma_{\bar{a}\bar{\alpha}}, \Gamma_{\bar{t}}]) + e_{\bar{a}\bar{t}} \wedge R_{\bar{a}\bar{\alpha}} = 8\pi T_{\bar{a}\bar{t}}^{**}. \quad (16.358c)$$

Equation (16.357) is an equation of motion in the sense that it involves a time derivative $d_t \Gamma_{\bar{a}\bar{\alpha}}$; but $\Gamma_{\bar{a}\bar{\alpha}}$ is not one of the momenta $\Gamma_{\bar{0}\bar{\alpha}}$ conjugate to the area element $e_{\bar{a}\bar{\alpha}}^2$, so equation (16.357) has a different status from the $9 + 9$ equations of motion (16.356). In the ADM formalism, §16.16.1, equation (16.357) is discarded as redundant with the 3 momentum constraints (16.359d), on the grounds that the energy-momentum tensor is symmetric (for vanishing torsion). The BSSN formalism on the other hand, §16.16.2, retains equation (16.357) as a distinct equation of motion.

The earlier equations (16.313) had the problem that the time derivative in the equation of motion for $\Gamma_{\bar{\alpha}}$ was $e_{\bar{\alpha}} \wedge d_t$ as opposed to just d_t . Equation (16.358b) seems to have the same difficulty, but here it is no longer a problem, because the 9-component trivector 3-form $e_{\bar{a}\bar{\alpha}} \wedge d_t \Gamma_{\bar{0}\bar{\alpha}}$ is invertibly related to the 9-component bivector 2-form $d_t \Gamma_{\bar{0}\bar{\alpha}}$, so equation (16.358b) can be rearranged as an equation for $d_t \Gamma_{\bar{0}\bar{\alpha}}$.

Variation of the action with respect to $\delta\Gamma_{\bar{a}\bar{\alpha}}$, $\delta\Gamma_{\bar{a}\bar{t}}$, $\delta\Gamma_{\bar{0}\bar{t}}$, $\delta e_{\bar{a}\bar{t}}$ and $\delta e_{\bar{0}\bar{t}}$ yields 9 identities and 10 constraints involving only spatial derivatives

$$9 \text{ identities: } (e \wedge S)_{\bar{0}\bar{t}} = 8\pi \Sigma_{\bar{0}\bar{t}}^{**}, \quad (16.359a)$$

$$3 \text{ Gaussian constraints: } (e \wedge S)_{\bar{0}\bar{\alpha}} = 8\pi \Sigma_{\bar{0}\bar{\alpha}}^{**}, \quad (16.359b)$$

$$3 \text{ Gaussian constraints: } (e \wedge S)_{\bar{a}\bar{\alpha}} = 8\pi \Sigma_{\bar{a}\bar{\alpha}}^{**}, \quad (16.359c)$$

$$3 \text{ momentum constraints: } (e \wedge R)_{\bar{0}\bar{\alpha}} = 8\pi T_{\bar{0}\bar{\alpha}}^{**}, \quad (16.359d)$$

$$1 \text{ Hamiltonian constraint: } (e \wedge R)_{\bar{a}\bar{\alpha}} = 8\pi T_{\bar{a}\bar{\alpha}}^{**}. \quad (16.359e)$$

The 9 identities (16.359a) are not equations of motion (they involve no time derivatives), despite having a form index $\bar{t}$. Explicitly,

$$9 \text{ identities: } \frac{1}{2}[\Gamma_{\bar{0}\bar{t}}, e_{\bar{a}\bar{\alpha}}^2] + \frac{1}{2}[\Gamma_{\bar{0}\bar{\alpha}}, e_{\bar{a}\bar{t}}^2] = 8\pi \Sigma_{\bar{0}\bar{t}}^{**}. \quad (16.360)$$

16.16.1 ADM formalism

The previous section 16.16 explored the form of the Hamiltonian system of equations when the ADM gauge condition (16.331) is imposed, and the coordinates and momenta are taken to be the extrinsic curvature $\Gamma_{\bar{0}\bar{\alpha}}$ and the spatial area element $e_{\bar{a}\bar{\alpha}}^2$. However, the traditional ADM formalism goes further than just imposing a gauge condition. The ADM formalism is pursued at length in Chapter 17 in traditional (coordinate and tetrad) index notation. Here it is useful to offer a few comments on the ADM formalism in the present context of multivector forms notation.

The ADM formalism imposes the ADM gauge condition (16.331), $e_{\bar{0}\bar{\alpha}} = 0$, from the outset, reducing the degrees of freedom of the spatial line interval e_{α} from 12 to 9. The ADM formalism further assumes from the outset that torsion vanishes. One of the consequences of vanishing torsion is that the energy-momentum tensor is symmetric, equation (16.284). This motivates the ADM strategy of simply discarding the 6 antisymmetric components of the Einstein equations, the 6 antisymmetric components of the 12 equations of motion (16.325b). Discarding the antisymmetric Einstein's equations seems innocent enough, until one realises that antisymmetric part of the energy-momentum tensor is a source in the law of conservation of spin angular-momentum, equation (16.283), which law is responsible for the 6 Gaussian constraints (16.324). Thus discarding the 6 antisymmetric Einstein equations is equivalent to using up the 6 Gaussian constraints. As a corollary, the 6 Gaussian constraints can no longer be treated as constraints; rather, they must be treated as identities.

Finally, the usual ADM strategy (though not a necessary one — see Chapter 17), is to work entirely with coordinate-frame quantities. An advantage of this approach is that all quantities are spatially Lorentz gauge-invariant (the ADM gauge choice (16.331) removes the gauge freedom of Lorentz boosts). In particular, the 9 components $e_{a\alpha}$ of the spatial line interval reduce to the 6 components of the Lorentz gauge-invariant spatial metric $g_{\alpha\beta}$, and the 24 Lorentz connections are replaced by the 6 components of the symmetric (for vanishing torsion) extrinsic curvatures $\Gamma_{\alpha 0 \beta}$ together with the $3 \times 6 = 18$ torsion-free coordinate-frame spatial connections (Christoffel symbols) $\Gamma_{\alpha\beta\gamma}$.

In all, in the ADM formalism there are $6 + 6 = 12$ equations of motion for $g_{\alpha\beta}$ and $\Gamma_{\alpha 0\beta}$, 18 identities for $\Gamma_{\alpha\beta\gamma}$, and 4 Hamiltonian constraints, a total of 34 equations altogether. The 6 equations lost compared to the 40 of the Hamiltonian system (16.322) and (16.323) are the 6 antisymmetric Einstein equations.

16.16.2 BSSN formalism

The BSSN formalism, discussed further in Chapter 17, §17.8, has gained popularity because it is strongly hyperbolic, and therefore has better numerical stability when applied to problems such as the merger of two black holes.

The BSSN formalism follows ADM for the most part, in particular imposing the ADM gauge choice (16.331). However, instead of discarding all 6 antisymmetric components of the 12 Einstein equations (16.325b), BSSN retains the 3 antisymmetric components $\Pi_{\bar{a}\bar{i}}$, which govern the evolution of the 3-component all-spatial momentum $\pi_{\bar{a}\bar{\alpha}}$, equation (16.335). BSSN thereby keeps 3 Gaussian constraints, the ones governing the evolution of the 3-component all-spatial contracted torsion ($e \wedge S$) $_{\bar{a}\bar{\alpha}}$ equation (16.341).

In all, the BSSN equations constitute $6 + 6 + 3 = 15$ equations of motion for the spatial metric $g_{\alpha\beta}$, the (symmetric, for vanishing torsion) extrinsic curvature $\Gamma_{\alpha 0\beta}$, and the BSSN variable $\Gamma_{b\alpha}^{\alpha}$ (eq. 16.335), 15 identities for the torsion-free coordinate-frame connections $\Gamma_{\alpha\beta\gamma}$, 4 Hamiltonian constraints, and 3 Gaussian constraints, a total of 37 equations altogether. The 3 equations lost compared to the 40 of the Hamiltonian system (16.322) and (16.323) are the 3 antisymmetric spatial Einstein equations.

Exercise 16.14 Gravitational equations in arbitrary spacetime dimensions. In multivector forms language in N spacetime dimensions:

1. What is the Hilbert gravitational Lagrangian? What is the gravitational super-Hamiltonian?
2. What is the variation of the gravitational Lagrangian?
3. What are the gravitational equations of motion?
4. What is the space+time $(N-1)+1$ split of the gravitational equations of motion?
5. What is the alternative Hilbert gravitational Lagrangian?
6. What is the variation of the alternative gravitational Lagrangian?
7. What is the space+time $(N-1)+1$ split of the alternative gravitational equations of motion?
8. What is the space+time $(N-1)+1$ split of the gravitational equations of motion when the ADM gauge condition (16.331) is imposed?

Solution.

1. The Hilbert gravitational Lagrangian in N spacetime dimensions is the scalar N -form, generalizing equation (16.236),

$$L_g = -\frac{I_N}{\kappa_N} e^{N-2} \wedge R, \quad (16.361)$$

where I_N is the N -dimensional spacetime pseudoscalar, and κ_N is Newton's gravitational constant,

suitably normalized, in N spacetime dimensions ($\kappa_4 = 8\pi G$). The Lagrangian (16.361) is in super-Hamiltonian form

$$L_g = -\frac{I_N}{\kappa_N} e^{N-2} \wedge d\mathbf{\Gamma} - H_g, \quad (16.362)$$

with super-Hamiltonian, generalizing equation (16.238),

$$H_g = \frac{I_N}{4\kappa_N} e^{N-2} \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}]. \quad (16.363)$$

2. The variation of the gravitational action in N spacetime dimensions is, generalizing equation (16.249),

$$\delta S_g = (-)^{N-1} \frac{I_N}{\kappa_N} \oint e^{N-2} \wedge \delta \mathbf{\Gamma} - \frac{I_N}{\kappa_N} \int (e^{N-3} \wedge \mathbf{S}) \wedge \delta \mathbf{\Gamma} + \delta e \wedge (e^{N-3} \wedge \mathbf{R}). \quad (16.364)$$

The variation of the matter action is defined by equations (16.250) in any spacetime dimension. With matter, the equations of motion generalizing equations (16.252) are

$$\frac{1}{2}N^2(N-1) \text{ equations of motion: } e^{N-3} \wedge \mathbf{S} = \kappa_N \mathbf{\Sigma}^{**}, \quad (16.365a)$$

$$N^2 \text{ equations of motion: } e^{N-3} \wedge \mathbf{R} = \kappa_N \mathbf{T}^{**}. \quad (16.365b)$$

The pseudobivector pseudo 1-form set of equations (16.365a) governing the torsion have the same number of components as the vector 2-form torsion $\mathbf{S}$ defined by equation (16.211), so completely determine the torsion in terms of the spin angular-momentum $\mathbf{\Sigma}$. Thus in N spacetime dimensions, as in 4 spacetime dimensions, torsion vanishes in empty space, and does not propagate. By contrast, the pseudovector pseudo 1-form set of equations (16.365b) constitute N^2 equations governing the $(\frac{1}{2}N(N-1))^2$ components of the bivector 2-form Riemann curvature $\mathbf{R}$ defined by equation (16.207); the equations of motion (16.365b) determine only the contracted components of the Riemann tensor. The remaining $\frac{1}{4}(N+1)N^2(N-3)$ components of the Riemann tensor are governed by Bianchi identities, §16.17. In $N > 3$ spacetime dimensions, as in 4 spacetime dimensions, Riemann curvature does not vanish in empty space, but rather propagates as a wave.

3. Split into time and space parts, the spacetime equations of motion (16.365) split into equations of motion that involve time derivatives d_t of the $N(N-1)$ spatial momenta e and $\frac{1}{2}N(N-1)^2$ spatial coordinates $\mathbf{\Gamma}$, generalizing equations (16.311),

$$\frac{1}{2}N(N-1)^2 \text{ equations of motion: } (e^{N-3} \wedge \mathbf{S})_{\bar{t}} = \kappa_N \mathbf{\Sigma}_{\bar{t}}^{**}, \quad (16.366a)$$

$$N(N-1) \text{ equations of motion: } (e^{N-3} \wedge \mathbf{R})_{\bar{t}} = \kappa_N \mathbf{T}_{\bar{t}}^{**}, \quad (16.366b)$$

and purely spatial constraint equations involving no time derivatives d_t , generalizing equations (16.312),

$$\frac{1}{2}N(N-1) \text{ Gaussian constraints: } (e^{N-3} \wedge \mathbf{S})_{\bar{\alpha}} = \kappa_N \mathbf{\Sigma}_{\bar{\alpha}}^{**}, \quad (16.367a)$$

$$N \text{ Hamiltonian constraints: } (e^{N-3} \wedge \mathbf{R})_{\bar{\alpha}} = \kappa_N \mathbf{T}_{\bar{\alpha}}^{**}. \quad (16.367b)$$

4. The alternative Hilbert gravitational Lagrangian in N spacetime dimensions is, generalizing equation (16.259),

$$L'_g = (-)^{N-1} \frac{I_N}{\kappa_N} \boldsymbol{\pi} \wedge d\mathbf{e} - H_g , \quad (16.368)$$

where $\boldsymbol{\pi}$ is momentum pseudovector $(N-2)$ -form, generalizing equation (16.255),

$$\boldsymbol{\pi} \equiv (-)^{N-3} \mathbf{e}^{N-3} \wedge \boldsymbol{\Gamma} , \quad (16.369)$$

and H_g is the same super-Hamiltonian (16.363) as before.

5. The variation of the alternative action (16.368) is, generalizing equation (16.261)

$$\delta S'_g = \frac{I_N}{\kappa_N} \oint \boldsymbol{\pi} \wedge \delta \mathbf{e} + (-)^{N-1} \frac{I_N}{\kappa_N} \int \delta \boldsymbol{\pi} \wedge \mathbf{S} + \boldsymbol{\Pi} \wedge \delta \mathbf{e} , \quad (16.370)$$

where the curvature pseudovector $(N-1)$ -form $\boldsymbol{\Pi}$ is, generalizing equations (16.262),

$$\boldsymbol{\Pi} \equiv \mathbf{e}^{N-3} \wedge \mathbf{R} - \mathbf{e}^{N-4} \wedge \mathbf{S} \wedge \boldsymbol{\Gamma} = d\boldsymbol{\pi} + \frac{1}{2} [\boldsymbol{\Gamma}, \boldsymbol{\pi}] + \frac{1}{4} \mathbf{e}^{N-3} \wedge [\boldsymbol{\Gamma}, \boldsymbol{\Gamma}] . \quad (16.371)$$

Note that the $\mathbf{e}^{N-4}$ term in the middle expression vanishes for $N = 3$. The variation of the matter action is defined by equations (16.263) in any spacetime dimension. With matter, the equations of motion generalizing equations (16.268) are

$$\frac{1}{2} N^2 (N-1) \text{ equations of motion: } \mathbf{S} = \kappa_N \tilde{\boldsymbol{\Sigma}} , \quad (16.372a)$$

$$N^2 \text{ equations of motion: } \boldsymbol{\Pi} = \kappa_N \tilde{\mathbf{T}} . \quad (16.372b)$$

6. Split into time and space parts, the alternative spacetime equations of motion (16.372) split into equations of motion that involve time derivatives d_t of the $N(N-1)$ spatial coordinates $\mathbf{e}$ and $N(N-1)$ spatial momenta $\boldsymbol{\pi}$, generalizing equations (16.322),

$$N(N-1) \text{ equations of motion: } \mathbf{S}_{\bar{t}} = \kappa_N \tilde{\boldsymbol{\Sigma}}_{\bar{t}} , \quad (16.373a)$$

$$N(N-1) \text{ equations of motion: } \boldsymbol{\Pi}_{\bar{t}} = \kappa_N \tilde{\mathbf{T}}_{\bar{t}} , \quad (16.373b)$$

and purely spatial equations involving no time derivatives d_t , generalizing equations (16.323),

$$\frac{1}{2} N(N-1)(N-3) \text{ gravomagnetic identities: } \mathbf{S}_{\bar{a}\bar{\alpha}} = \kappa_N \tilde{\boldsymbol{\Sigma}}_{\bar{a}\bar{\alpha}} \quad (16.374a)$$

$$\frac{1}{2} N(N-1) \text{ Gaussian constraints: } (\mathbf{e}^{N-3} \wedge \mathbf{S})_{\bar{\alpha}} = \kappa_N (\mathbf{e}^{N-3} \wedge \tilde{\boldsymbol{\Sigma}})_{\bar{\alpha}} , \quad (16.374b)$$

$$N \text{ Hamiltonian constraints: } \boldsymbol{\Pi}_{\bar{\alpha}} = \kappa_N \tilde{\mathbf{T}}_{\bar{\alpha}} . \quad (16.374c)$$

Prolonging the gravomagnetic field $\mathbf{F}_{\bar{a}\bar{\alpha}}$ replaces the identities (16.374a) by the same number each of equations of motion and constraints, generalizing equations (16.350) and (16.352),

$$\frac{1}{2} N(N-1)(N-3) \text{ equations of motion: } (d[\frac{1}{2} \mathbf{e}, \mathbf{F}])_{\bar{a}\bar{t}} = \kappa_N (d\tilde{\boldsymbol{\Sigma}})_{\bar{a}\bar{t}} , \quad (16.375a)$$

$$\frac{1}{2} N(N-1)(N-3) \text{ gravomagnetic constraints: } (\mathbf{e}^{N-3} \wedge \mathbf{S})_{\bar{\alpha}} = \kappa_N (\mathbf{e}^{N-3} \wedge \tilde{\boldsymbol{\Sigma}})_{\bar{\alpha}} . \quad (16.375b)$$

7. ADM imposes the $N - 1$ ADM gauge conditions $e_{\bar{0}\bar{\alpha}} = 0$, equation (16.331), reducing the number of degrees of freedom of the spatial line-element $\mathbf{e}$ to $(N-1)^2$, and likewise the number of degrees of freedom of the spatial area element $\mathbf{e}^{N-2}$ to $(N-1)^2$. The momenta conjugate to the spatial area element are $-\mathbf{\Gamma}_{\bar{0}\bar{\alpha}}$, again with $(N-1)^2$ degrees of freedom. There are $2(N-1)^2$ equations of motion for the spatial area element and their conjugate momenta, generalizing equations (16.356),

$$(N-1)^2 \text{ equations of motion: } (e^{N-3} \wedge \mathbf{S})_{\bar{a}\bar{t}} = \kappa_N \mathbf{\Sigma}_{\bar{a}\bar{t}}^{**}, \quad (16.376a)$$

$$(N-1)^2 \text{ equations of motion: } (e^{N-3} \wedge \mathbf{R})_{\bar{0}\bar{t}} = \kappa_N \mathbf{T}_{\bar{0}\bar{t}}^{**}. \quad (16.376b)$$

There are a further $N-1$ equations of motion that are discarded in the ADM formalism (incidentally demoting the Gaussian constraints (16.378c) from constraints to identities) but retained in the BSSN formalism, generalizing equation (16.357),

$$N-1 \text{ equations of motion: } (e^{N-3} \wedge \mathbf{R})_{\bar{a}\bar{t}} = \kappa_N \mathbf{T}_{\bar{a}\bar{t}}^{**}. \quad (16.377)$$

The remaining equations, containing no time derivatives d_t , comprise $\frac{1}{2}(N-1)^2(N-2)$ identities and $\frac{1}{2}N(N+1)$ constraints, generalizing equations (16.359),

$$\frac{1}{2}(N-1)^2(N-2) \text{ identities: } (e^{N-3} \wedge \mathbf{S})_{\bar{0}\bar{t}} = \kappa_N \mathbf{\Sigma}_{\bar{0}\bar{t}}^{**}, \quad (16.378a)$$

$$\frac{1}{2}(N-1)(N-2) \text{ Gaussian constraints: } (e^{N-3} \wedge \mathbf{S})_{\bar{0}\bar{\alpha}} = \kappa_N \mathbf{\Sigma}_{\bar{0}\bar{\alpha}}^{**}, \quad (16.378b)$$

$$N-1 \text{ Gaussian constraints: } (e^{N-3} \wedge \mathbf{S})_{\bar{a}\bar{\alpha}} = \kappa_N \mathbf{\Sigma}_{\bar{a}\bar{\alpha}}^{**}, \quad (16.378c)$$

$$N-1 \text{ momentum constraints: } (e^{N-3} \wedge \mathbf{R})_{\bar{0}\bar{\alpha}} = \kappa_N \mathbf{T}_{\bar{0}\bar{\alpha}}^{**}, \quad (16.378d)$$

$$1 \text{ Hamiltonian constraint: } (e^{N-3} \wedge \mathbf{R})_{\bar{a}\bar{\alpha}} = \kappa_N \mathbf{T}_{\bar{a}\bar{\alpha}}^{**}. \quad (16.378e)$$

Exercise 16.15 Volume of a ball and area of a sphere. What is the volume V_N of a unit N -ball, and the area A_N of a unit N -dimensional sphere? A unit N -ball is the interior of a unit $(N-1)$ -sphere, and an N -sphere is the boundary of a unit $(N+1)$ -ball.

Solution. The volume V_N of an N -ball is the area $A_{N-1}R^{N-1}$ of an $(N-1)$ -sphere of radius R integrated over R from 0 to 1,

$$V_N = A_{N-1} \int_0^1 R^{N-1} dR, \quad (16.379)$$

yielding

$$V_N = \frac{A_{N-1}}{N}. \quad (16.380)$$

The volume of an N -ball is also the volume $V_{N-1}r^{N-1}$ of an $(N-1)$ -ball of radius $r = \sin \theta$ integrated over height $z = \cos \theta$ from -1 to 1 ,

$$V_N = V_{N-1} \int_{-1}^1 r^{N-1} dz = V_{N-1} \int_0^\pi \sin^N \theta d\theta. \quad (16.381)$$

The integral $\int_0^\pi \sin^N \theta d\theta$ can be expressed in terms of Γ functions. Iterated twice, equation (16.381) gives the recurrence relation

$$V_N = \frac{2\pi V_{N-2}}{N} . \quad (16.382)$$

Equations (16.380) and (16.382) imply

$$A_N = 2\pi V_{N-1} . \quad (16.383)$$

Initial values of the recurrence are $V_1 = 2$ and $V_2 = \pi$. General expressions for the volume and area are

$$V_N = \frac{\pi^{N/2}}{\Gamma(\frac{N}{2} + 1)} , \quad A_N = \frac{2\pi^{(N+1)/2}}{\Gamma(\frac{N+1}{2})} . \quad (16.384)$$

16.17 Bianchi identities in multivector forms notation

The Bianchi identities, equations (16.405) below, are differential identities satisfied by the torsion $\mathbf{S}$ and Riemann $\mathbf{R}$ tensors. The Bianchi identities are identities in the sense that if the torsion and Riemann tensors are expressed in terms of the line interval $\mathbf{e}$ and Lorentz connection $\mathbf{\Gamma}$ in accordance with Cartan's equations (16.214) and (16.210), then the Bianchi identities are satisfied automatically. The contracted Bianchi identities (16.408) enforce conservation laws on the total spin angular-momentum $\mathbf{\Sigma}$ and matter energy-momentum $\mathbf{T}$.

In this section the number N of spacetime dimensions is arbitrary. The caret on various symbols in this section, such as $\hat{\mathbf{\Gamma}}$, equation (16.388), signifies that they are operators; $\hat{\mathbf{\Gamma}}$ should not be confused with the restricted Lorentz connection $\hat{\Gamma}_{mnl}$ considered in the next Chapter, equation (17.27).

16.17.1 Covariant exterior derivative of a multivector form

The exterior derivatives d of the multivector forms $\mathbf{\Gamma}$ and $\mathbf{e}$ in equations (16.208) and (16.212) were applied to the coordinate indices, but not to the tetrad indices. A covariant exterior derivative D , distinguished like the coordinate exterior derivative d by latin font, can be defined that is covariant not only with respect to coordinate transformations but also with respect to Lorentz transformations. In this context, covariance means that D commutes with both coordinate and Lorentz transformations. There is a torsion-free covariant exterior derivative $\mathring{D}$, and a torsion-full covariant exterior derivative D .

If $\mathbf{a}$ is a multivector p -form, then its torsion-free covariant exterior derivative $\mathring{D}\mathbf{a}$ is a sum of the coordinate exterior derivative plus a torsion-free Lorentz connection term, equation (15.4),

$$\mathring{D}\mathbf{a} \equiv d\mathbf{a} + \hat{\mathbf{\Gamma}}\mathbf{a} , \quad (16.385)$$

where the torsion-free Lorentz connection operator $\hat{\mathbf{\Gamma}}$ acting on the multivector form $\mathbf{a}$ is, equation (15.19),

$$\hat{\mathbf{\Gamma}}\mathbf{a} \equiv \frac{1}{2}[\mathring{\mathbf{\Gamma}}, \mathbf{a}] , \quad (16.386)$$

with $\mathring{\Gamma} \equiv \mathring{\Gamma}_{kl\kappa} \gamma^k \wedge \gamma^l dx^\kappa$ the torsion-free bivector 1-form, the torsion-free version of equation (16.203a).

The torsion-full covariant exterior derivative $D\mathbf{a}$ is a sum of the coordinate exterior derivative plus a torsion-full Lorentz connection term plus a torsion term,

$$D\mathbf{a} \equiv d\mathbf{a} + \hat{\Gamma}\mathbf{a} + \hat{S}\mathbf{a} , \quad (16.387)$$

where the Lorentz connection operator $\hat{\Gamma}$ acting on the multivector form $\mathbf{a}$ is, equation (15.19),

$$\hat{\Gamma}\mathbf{a} \equiv \frac{1}{2}[\Gamma, \mathbf{a}] . \quad (16.388)$$

The torsion-full Lorentz connection bivector 1-form Γ , equation (16.203a), is as usual the sum of the torsion-free Lorentz connection $\mathring{\Gamma}$ and the contortion K , equations (11.55),

$$\Gamma = \mathring{\Gamma} + K . \quad (16.389)$$

The coordinate exterior derivative d is a torsion-free covariant curl, so the torsion operator $\hat{S}$ in equation (16.387) must be included when torsion does not vanish, as for example in equation (2.71). The torsion operator $\hat{S}$ is essentially the antisymmetric part of the coordinate connection, which is the only part of the coordinate connection that survives in a (covariant) exterior derivative. The torsion operator $\hat{S}$ acts only on the coordinate indices of the form $\mathbf{a}$, while the Lorentz connection operator $\hat{\Gamma}$ acts only on the tetrad indices of the multivector $\mathbf{a}$. If $\mathbf{a} = \mathbf{a}_{\lambda\Pi} d^p x^{\lambda\Pi}$ is a multivector p -form, with implicit summation over distinct antisymmetric sequences $\lambda\Pi$ of p indices, then the torsion term $\hat{S}\mathbf{a}$ is the multivector $(p+1)$ -form defined by

$$\hat{S}\mathbf{a} \equiv \frac{p(p+1)}{2} S_{\kappa\lambda}^\mu \mathbf{a}_{\mu\Pi} d^{p+1} x^{\kappa\lambda\Pi} , \quad (16.390)$$

implicitly summed over distinct antisymmetric sequences $\kappa\lambda\Pi$ of $p+1$ indices. If $\mathbf{a}$ is a 0-form (a coordinate scalar), then the torsion term vanishes, $\hat{S}\mathbf{a} = 0$. In components, the covariant exterior derivative $D\mathbf{a}$, equation (16.387), of the multivector p -form $\mathbf{a}$ is the $(p+1)$ -form

$$D\mathbf{a} = (p+1) D_\kappa \mathbf{a}_{\lambda\Pi} d^{p+1} x^{\kappa\lambda\Pi} = (p+1) \left(\partial_\kappa \mathbf{a}_{\lambda\Pi} + \frac{1}{2}[\Gamma_\kappa, \mathbf{a}_{\lambda\Pi}] + \frac{1}{2} p S_{\kappa\lambda}^\mu \mathbf{a}_{\mu\Pi} \right) d^{p+1} x^{\kappa\lambda\Pi} , \quad (16.391)$$

with the implicit summation over distinct antisymmetric sequences $\kappa\lambda\Pi$ of $p+1$ indices.

The covariant exterior derivative D (in both torsion-free and torsion-full versions) acting on the product of a multivector p -form $\mathbf{a}$ and a multivector q -form $\mathbf{b}$ satisfies the same Leibniz-like rule as the exterior derivative d , equation (16.200),

$$D(\mathbf{a}\mathbf{b}) = (D\mathbf{a})\mathbf{b} + (-)^p \mathbf{a}(D\mathbf{b}) . \quad (16.392)$$

16.17.2 A third, Lorentz-covariant, exterior derivative

A third exterior derivative that is Lorentz-covariant but not coordinate-covariant crops up often enough to warrant a special notation. The Lorentz-covariant derivative D_Γ , subscripted Γ as a reminder that it is covariant only with respect to Lorentz indices, is

$$D_\Gamma \equiv d + \hat{\Gamma} , \quad (16.393)$$

which is torsion-free acting on coordinate indices, and torsion-full acting on multivector indices. The Lorentz-covariant derivative D_Γ satisfies the same Leibniz-like rule (16.392) as the other exterior derivatives.

The derivative D_Γ is not coordinate-covariant in the sense that it does not commute with the vierbein, that is, acting on the line interval e , it yields the torsion S , equation (16.395),

$$D_\Gamma e = S . \quad (16.394)$$

However, the derivative D_Γ satisfies other conditions for being a covariant derivative: it yields a (coordinate and tetrad) tensor when acting on a (coordinate and tetrad) tensor.

16.17.3 Torsion from the covariant exterior derivative

By construction, the covariant exterior derivative D (in either torsion-free or torsion-full versions) commutes with both coordinate and Lorentz transformations. Thus the covariant exterior derivative of the line-element e defined by equation (16.203b) vanishes, $De = 0$. Applied to the line interval e , equation (16.387) recovers the definition (16.214) of the torsion vector 2-form S , Cartan's first equation of structure,

$$0 = De = de + \frac{1}{2}[\Gamma, e] - S , \quad (16.395)$$

since $\hat{S}e$ is just minus the vector 2-form torsion S ,

$$\hat{S}e = S_{\kappa\lambda}^\mu e_{m\mu} \gamma^m d^2x^{\kappa\lambda} = S_{m\kappa\lambda} \gamma^m d^2x^{\kappa\lambda} = -S . \quad (16.396)$$

The torsion-free version of Cartan's equation (16.395) is $de + \frac{1}{2}[\overset{\circ}{\Gamma}, e] = 0$. Subtracting this torsion-free Cartan's equation from the torsion-full Cartan's equation (16.395) yields the relation between the torsion S and the contortion K ,

$$S = \frac{1}{2}[K, e] . \quad (16.397)$$

Equation (16.397) can be inverted to yield K in terms of S . The relation between torsion and contortion was given previously in index notation as equations (11.56).

16.17.4 Riemann curvature from the covariant exterior derivative

Whereas the square of the coordinate exterior derivative vanishes because of the commutation of coordinate partial derivatives, $dd = 0$, the square of the covariant exterior derivative does not vanish. In components, the square DD is

$$DD \equiv [D_\kappa, D_\lambda] d^2x^{\kappa\lambda} , \quad (16.398)$$

implicitly summed over distinct antisymmetric pairs of indices $\kappa\lambda$. Acting on any multivector form a , the square of the covariant exterior derivative gives (compare equation (15.21))

$$DDa = \hat{R}a + \hat{S}Da . \quad (16.399)$$

If $\mathbf{a} = \mathbf{a}_{\mu\Pi} d^p x^{\mu\Pi}$ is a multivector p -form, then the Riemann operator $\hat{\mathbf{R}}$ acting on $\mathbf{a}$ is the $(p+2)$ -form (the $D_{\Gamma}\hat{\mathbf{S}}$ term in the following equation was given previously in components by equation (11.70))

$$\hat{\mathbf{R}}\mathbf{a} = (D_{\Gamma}D_{\Gamma} + D_{\Gamma}\hat{\mathbf{S}})\mathbf{a} = \frac{(p+1)(p+2)}{2} \left(\frac{1}{2}[\mathbf{R}_{\kappa\lambda}, \mathbf{a}_{\mu\Pi}] + \frac{1}{3}p R_{\kappa\lambda\mu}{}^{\nu} \mathbf{a}_{\nu\Pi} \right) d^{p+2} x^{\kappa\lambda\mu\Pi}, \quad (16.400)$$

implicitly summed over distinct antisymmetric sequences $\kappa\lambda\mu\Pi$ of $p+2$ indices. The components of the Riemann tensor are those of the Riemann bivector 2-form $\mathbf{R} \equiv \mathbf{R}_{\kappa\lambda} d^2 x^{\kappa\lambda}$, equation (16.207). Equation (16.400) recovers the definition (16.210) of the Riemann curvature $\mathbf{R}$ in terms of the Lorentz connection $\mathbf{\Gamma}$, Cartan's second equation of structure. In equation (16.399), the scalar 2-form covariant derivative operator $\hat{\mathbf{S}}D$ acting on the multivector p -form $\mathbf{a} = \mathbf{a}_{\Pi} d^p x^{\Pi}$ is, from equations (16.390) and (16.391), the $(p+2)$ -form

$$\hat{\mathbf{S}}D\mathbf{a} = \frac{(p+1)^2(p+2)}{2} S_{\kappa\lambda}^{\mu} D_{[\mu} \mathbf{a}_{\Pi]} d^{p+2} x^{\kappa\lambda\Pi}, \quad (16.401)$$

implicitly summed over distinct antisymmetric sequences $\kappa\lambda\Pi$ of $p+2$ indices.

16.17.5 Bianchi identities

The Jacobi identity for the covariant exterior derivative is

$$D(DD) - (DD)D = 0. \quad (16.402)$$

Applied to an arbitrary multivector form $\mathbf{a}$, the Jacobi identity (16.402) implies

$$0 = D(DD)\mathbf{a} - (DD)D\mathbf{a} = D(\hat{\mathbf{R}} + \hat{\mathbf{S}}D)\mathbf{a} - (\hat{\mathbf{R}} + \hat{\mathbf{S}}D)D\mathbf{a} = (D\hat{\mathbf{R}} - \hat{\mathbf{S}}\hat{\mathbf{R}})\mathbf{a} + (D\hat{\mathbf{S}} - \hat{\mathbf{S}}\hat{\mathbf{S}} - \hat{\mathbf{R}})D\mathbf{a}. \quad (16.403)$$

Equation (16.403) holds for arbitrary $\mathbf{a}$, so the coefficients of $\mathbf{a}$ and $D\mathbf{a}$ must vanish, implying the Bianchi identities

$$D\mathbf{S} - \hat{\mathbf{S}}\mathbf{S} + \frac{1}{2}[\mathbf{e}, \mathbf{R}] = 0, \quad (16.404a)$$

$$D\mathbf{R} - \hat{\mathbf{S}}\mathbf{R} = 0, \quad (16.404b)$$

where $\mathbf{S}$ is the vector 2-form torsion (16.211) with $\frac{1}{2}N^2(N-1)$ components, and $\mathbf{R}$ is the bivector 2-form Riemann curvature (16.207) with $(\frac{1}{2}N(N-1))^2$ components. These equations (16.404) were given previously in component form by equations (11.69) and (11.91). Equation (16.404a) is a vector 3-form, with $\frac{1}{6}N^2(N-1)(N-2)$ components, while equation (16.404b) is a bivector 3-form, with $\frac{1}{12}N^2(N-1)^2(N-2)$ components. Equivalently, in terms of the exterior derivative d instead of the covariant exterior derivative D , the Bianchi identities (16.404) are (note that $\frac{1}{2}[\mathbf{e}, \mathbf{R}] = \mathbf{e} \cdot \mathbf{R}$, the dot product being defined by equation (13.35))

$$d\mathbf{S} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{S}] + \mathbf{e} \cdot \mathbf{R} = 0, \quad (16.405a)$$

$$d\mathbf{R} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{R}] = 0. \quad (16.405b)$$

The Bianchi identities (16.405) follow alternatively from taking the exterior derivatives of the torsion and Riemann curvature forms defined by Cartan's equations of structure (16.214) and (16.210), Exercise 16.16.

Thus the Bianchi identities (16.405) are identically true as long as Cartan's equations of structure hold, and do not impose additional constraints on torsion $\mathbf{S}$ or curvature $\mathbf{R}$.

Exercise 16.16 Bianchi identities. Confirm that the Bianchi identities (16.405) follow from taking exterior derivatives of torsion and Riemann curvature forms $\mathbf{S}$ and $\mathbf{R}$ defined by Cartan's equations of structure (16.214) and (16.210).

Solution. The commutation rules for multivector-valued forms are (16.197), and the exterior derivative of a product of multivector-valued forms satisfies the rule (16.200).

$$\begin{aligned} d\mathbf{S} &= d\left(\mathbf{d}\mathbf{e} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{e}]\right) = -\frac{1}{2}([\mathbf{\Gamma}, d\mathbf{e}] + [d\mathbf{e}, \mathbf{\Gamma}]) = -\frac{1}{2}[\mathbf{\Gamma}, \mathbf{S}] + \frac{1}{4}[\mathbf{\Gamma}, [\mathbf{\Gamma}, \mathbf{e}]] - \frac{1}{2}[d\mathbf{e}, \mathbf{R}] + \frac{1}{8}[\mathbf{e}, [\mathbf{\Gamma}, \mathbf{\Gamma}]] \\ &= -\frac{1}{2}[\mathbf{\Gamma}, \mathbf{S}] - \frac{1}{2}[d\mathbf{e}, \mathbf{R}] + \frac{1}{4}(\mathbf{\Gamma}(\mathbf{\Gamma}\mathbf{e} + \mathbf{e}\mathbf{\Gamma}) - (\mathbf{\Gamma}\mathbf{e} + \mathbf{e}\mathbf{\Gamma})\mathbf{\Gamma} + \mathbf{e}(\mathbf{\Gamma}\mathbf{\Gamma}) - (\mathbf{\Gamma}\mathbf{\Gamma})\mathbf{e}) \\ &= -\frac{1}{2}[\mathbf{\Gamma}, \mathbf{S}] - \frac{1}{2}[d\mathbf{e}, \mathbf{R}] , \end{aligned} \tag{16.406a}$$

$$\begin{aligned} d\mathbf{R} &= d\left(\mathbf{d}\mathbf{\Gamma} + \frac{1}{4}[\mathbf{\Gamma}, \mathbf{\Gamma}]\right) = -\frac{1}{2}[\mathbf{\Gamma}, d\mathbf{\Gamma}] = -\frac{1}{2}[\mathbf{\Gamma}, \mathbf{R}] + \frac{1}{8}[\mathbf{\Gamma}, [\mathbf{\Gamma}, \mathbf{\Gamma}]] \\ &= -\frac{1}{2}[\mathbf{\Gamma}, \mathbf{R}] . \end{aligned} \tag{16.406b}$$

16.17.6 Interpretation of the Bianchi identities

The torsion Bianchi identity (16.405a) looks like a covariant conservation equation for torsion $\mathbf{S}$, except that there is a source term $\mathbf{e} \cdot \mathbf{R}$, a vector 3-form whose $\frac{1}{6}N^2(N-1)(N-2)$ components are

$$\mathbf{e} \cdot \mathbf{R} = \frac{1}{2}[d\mathbf{e}, \mathbf{R}] = R_{[\kappa\lambda\mu]n} \gamma^n d^3x^{\kappa\lambda\mu} . \tag{16.407}$$

Since torsion $\mathbf{S}$ is, according to its equation of motion (16.372a), everywhere equal to the spin angular-momentum $\tilde{\mathbf{\Sigma}}$, the torsion Bianchi identity (16.405a) can be interpreted as determining $\mathbf{e} \cdot \mathbf{R}$ in terms of the torsion and its derivatives. I thank Fred Hehl for pointing out (2017, private communication) that the $\mathbf{e} \cdot \mathbf{R}$ term can be interpreted as the covariant exterior derivative of orbital angular momentum, §19(c) of Corson (1953), so that the Bianchi identity (16.405a) can be interpreted as enforcing conservation of total angular momentum, spin plus orbital. If torsion $\mathbf{S}$ vanishes, or more generally if it satisfies the covariant conservation equation $d\mathbf{S} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{S}] = 0$, then $\mathbf{e} \cdot \mathbf{R} = 0$. The remaining $\left(\frac{1}{2}N(N-1)\right)^2 - \frac{1}{6}N^2(N-1)(N-2) = \frac{1}{12}(N+1)N^2(N-1)$ components of the Riemann tensor constitute its torsion-free part $\tilde{\mathbf{R}}$, equation (15.49).

The Riemann Bianchi identity (16.405b) looks like a covariant conservation equation for the Riemann tensor $\mathbf{R}$. In contrast to the torsion $\mathbf{S}$, the equation of motion (16.375b) for the Riemann tensor determines only the contracted (Ricci) part $\mathbf{e}^{N-3} \wedge \mathbf{R}$ of the Riemann tensor in terms of the matter energy-momentum $\mathbf{T}$. That is, even though the equations of motion (16.375) along with initial conditions completely determine the Riemann tensor and its evolution, there is a part of the Riemann tensor, the non-contracted (Weyl) part, that is not fixed by the local matter energy-momentum at each point. The fact that the Riemann curvature tensor has a non-local part is what allows there to be waves of curvature — gravitational waves. The Riemann Bianchi identity (16.405b) has the form of a linear differential constraint on the Riemann tensor. As already emphasized, the Riemann Bianchi identity (16.405b) is just the exterior derivative of the

definition (16.210) of the Riemann tensor $\mathbf{R}$, and does not impose any additional constraint beyond those set by the equations of motion and initial conditions. Nevertheless, the fact that the Riemann tensor satisfies the differential identity (16.405b) means that the Weyl tensor satisfies Maxwell-like equations which describe gravitational waves, as discussed for the case of 4 spacetime dimensions in §3.2.

16.17.7 Contracted Bianchi identities

The equations of motion (16.365) for torsion and curvature are equations for $e^{N-3} \wedge \mathbf{S}$ and $e^{N-3} \wedge \mathbf{R}$. Differential constraints on these combinations are obtained by contracting the Bianchi identities (16.405) by pre-multiplying by $e^{N-3} \wedge$. The contracted Bianchi identities for torsion and Riemann curvature constitute respectively a pseudobivector N -form with $\frac{1}{2}N(N-1)$ components, and a pseudovector N -form with N components,

$$d(e^{N-3} \wedge \mathbf{S}) + \frac{1}{2}[\Gamma, e^{N-3} \wedge \mathbf{S}] + (-)^{N-3} e \cdot (e^{N-3} \wedge \mathbf{R}) = 0, \quad (16.408a)$$

$$d(e^{N-3} \wedge \mathbf{R}) + \frac{1}{2}[\Gamma, e^{N-3} \wedge \mathbf{R}] + (-)^{N-3} e^{N-4} \wedge \mathbf{S} \wedge \mathbf{R} = 0. \quad (16.408b)$$

The final term in the contracted torsion Bianchi identity (16.408a) is a pseudobivector N -form whose components constitute the antisymmetric part $R_{[\mu\nu]}$ of the Ricci tensor,

$$e \cdot (e^{N-3} \wedge \mathbf{R}) = e^{N-3} \wedge (e \cdot \mathbf{R}) = \frac{N-2}{2} [e^{N-2}, \mathbf{R}] = -e_{\kappa\kappa\dots} e_{l\lambda} R_{\mu\nu} \gamma^k \wedge \dots \wedge \gamma^l d^N x^{\kappa\dots\lambda\mu\nu}, \quad (16.409)$$

implicitly summed as usual over distinct antisymmetric sequences $k\dots l$ and $\kappa\dots\lambda\mu\nu$ of indices.

Combining the contracted Bianchi identity (16.408b) with the torsion Bianchi identity (16.405a) yields the pseudovector N -form identity for the curvature $\mathbf{\Pi}$ defined by equation (16.371),

$$d\mathbf{\Pi} + \frac{1}{2}[\Gamma, \mathbf{\Pi}] + (-)^{N-3} e^{N-4} ((e \cdot \mathbf{R}) \wedge \Gamma - \frac{1}{4} \mathbf{S} \wedge [\Gamma, \Gamma]) = 0. \quad (16.410)$$

16.17.8 Interpretation of the contracted Bianchi identities

The contracted torsion Bianchi identity (16.408a) is the $\frac{1}{2}N(N-1)$ -component conservation law associated with invariance of the gravitational Lagrangian under Lorentz transformations. The contracted Riemann identity (16.408b), or equivalently (16.410), is the N -component conservation law associated with invariance of the gravitational Lagrangian under coordinate transformations.

The contracted torsion Bianchi identity (16.408a) enforces continued satisfaction of the Gaussian constraint (16.312a). The contracted Riemann Bianchi identity (16.408b) enforces continued satisfaction of the Hamiltonian constraint (16.312b).

Conventional Hamiltonian (3+1) approach

In the previous Chapter, gravitational equations of motion were derived from the Hilbert Lagrangian in a fully covariant fashion, and the (super-)Hamiltonian form of the Hilbert Lagrangian was emphasized. The present Chapter explores a more traditional non-covariant 3+1 approach, in which the spacetime is foliated into hypersurfaces of constant time t , and the system of Einstein (and other) equations is evolved by integrating from one spacelike hypersurface of constant time to the next.

The traditional 3+1 formalism is called the **Arnowitt-Deser-Misner (ADM) formalism**, introduced by Arnowitt, Deser & Misner (1959; 1963). The original purpose of ADM was to cast the gravitational equations of motion into conventional Hamiltonian form, to facilitate quantization. The goal of quantizing general relativity failed, but the ADM formalism revealed fundamental insights into the structure of the Einstein equations (see §16.16.1 of the previous Chapter). The ADM formalism provides the backbone for modern codes that implement numerical general relativity.

The ADM formalism reveals that, for vanishing torsion, the 6 physical degrees of freedom of the gravitational field can be regarded as being carried by the 6 spatial components $g_{\alpha\beta}$ of the coordinate metric. The 6 spatial Einstein equations constitute partial differential equations of motion of second order in time t for the 6 physical degrees of freedom. The remaining 4 degrees of freedom of the coordinate metric can be treated as gauge degrees of freedom, which can be chosen arbitrarily. The 4 non-spatial Einstein equations are partial differential equations of first order in time t , and they are not equations of motion, but rather constraint equations, which must be arranged to be satisfied in the initial conditions (on the initial hypersurface of constant time t), but which are guaranteed thereafter by the contracted Bianchi identities, which enforce conservation of energy-momentum.

The mere fact that the 6 spatial components $g_{\alpha\beta}$ of the coordinate metric *can* (if torsion vanishes) be taken to be the 6 gravitational physical degrees of freedom, and that the remaining 4 degrees of freedom of the coordinate metric *can* be treated as gauge degrees of freedom, does not mean that these choices *must* be made. In cosmology for example, the preferred gauge choice is conformal Newtonian (Copernican) gauge, §29.8, in which only 3 of the 6 physical perturbations are part of the spatial coordinate metric $g_{\alpha\beta}$ (the scalar Φ and the 2 components of the tensor h_{ab}), while the remaining 3 physical perturbations are part of the time components g_{tt} and $g_{t\alpha}$ of the metric (the scalar Ψ and the 2 components of the vector W_a).

Numerical experiments during the 1990s established that the ADM equations are numerically unstable. The

community of numerical relativists engaged in an intensive effort to understand the cause of the instability, and to find a numerically stable formalism. The challenge problem was to compute reliably the evolution of the merger of a pair of black holes, and to calculate the general relativistic radiation produced as a result. The effort was rewarded in 2005–6 when a number of groups (Pretorius, 2005a; Pretorius, 2006; Baker et al., 2006b; Baker et al., 2006a; Campanelli et al., 2006; Campanelli, Lousto, and Zlochower, 2006; Diener et al., 2006; Sopuerta, Sperhake, and Laguna, 2006) reported successful evolution of a binary black hole (or black hole plus neutron star) merger. The most popular formalism for long-term evolution of spacetimes is the **Baumgarte-Shapiro-Shibata-Nakamura (BSSN) formalism** (Shibata and Nakamura, 1995; Baumgarte and Shapiro, 1998).

This Chapter starts with an exposition of the ADM formalism, §17.1. It goes on to apply the ADM formalism to Bianchi spacetimes, §17.4, which provide a fine example of the application of the formalism in a non-trivial case. The gravitational collapse of Bianchi spacetimes reveals that collapse to a singularity can show a complicated oscillatory behaviour called **Belinskii-Khalatnikov-Lifshitz (BKL) oscillations** (Belinskii, Khalatnikov, and Lifshitz, 1970; Belinskii and Khalatnikov, 1971; Belinskii, Khalatnikov, and Lifshitz, 1972; Belinskii, Khalatnikov, and Lifshitz, 1982; Belinski, 2014), §17.6. The Chapter concludes with an exposition of the BSSN formalism, §17.8, and the elegant 4-dimensional version of it proposed by Pretorius (2005), §17.9.

In this Chapter, torsion is assumed to vanish.

17.1 ADM formalism

The ADM formalism splits the spacetime coordinates x^μ into a time coordinate t and spatial coordinates x^α , $\alpha = 1, 2, 3$,

$$x^\mu \equiv \{t, x^\alpha\} . \quad (17.1)$$

At each point of spacetime, the spacelike hypersurface of constant time t has a unique future-pointing unit normal γ_0 , defined to have unit length and to be orthogonal to the spatial tangent axes e_α ,

$$\gamma_0 \cdot \gamma_0 = -1 , \quad \gamma_0 \cdot e_\alpha = 0 \quad \alpha = 1, 2, 3 . \quad (17.2)$$

The central idea of the ADM approach is to work in a tetrad frame γ_m consisting of this time axis γ_0 , together with three spatial tetrad axes γ_a , also called the **triad**, that are orthogonal to the tetrad time axis γ_0 , and therefore lie in the 3D spatial hypersurface of constant time,

$$\gamma_0 \cdot \gamma_a = 0 \quad a = 1, 2, 3 . \quad (17.3)$$

The tetrad metric γ_{mn} in the ADM formalism is thus

$$\gamma_{mn} = \begin{pmatrix} -1 & 0 \\ 0 & \gamma_{ab} \end{pmatrix} , \quad (17.4)$$

and the inverse tetrad metric γ^{mn} is correspondingly

$$\gamma^{mn} = \begin{pmatrix} -1 & 0 \\ 0 & \gamma^{ab} \end{pmatrix}, \quad (17.5)$$

whose spatial part γ^{ab} is the inverse of γ_{ab} . Given the conditions (17.2) and (17.3), the vierbein $e^m{}_\mu$ and inverse vierbein $e_m{}^\mu$ take the form

$$e^m{}_\mu = \begin{pmatrix} \alpha & 0 \\ -e^a{}_\alpha \beta^\alpha & e^a{}_\alpha \end{pmatrix}, \quad e_m{}^\mu = \begin{pmatrix} 1/\alpha & \beta^\alpha/\alpha \\ 0 & e_a{}^\alpha \end{pmatrix}, \quad (17.6)$$

where α and β^α are the lapse and shift (see next paragraph), and $e^a{}_\alpha$ and $e_a{}^\alpha$ represent the spatial vierbein and inverse vierbein, which are inverse to each other, $e^a{}_\alpha e_b{}^\alpha = \delta_b^a$. As can be read off from equations (17.6), the following off-diagonal time-space components of the vierbein and its inverse vanish, as a direct consequence of the ADM gauge choices (17.2),

$$e^0{}_\alpha = e_a{}^t = 0. \quad (17.7)$$

The ADM line-element is

$$ds^2 = -\alpha^2 dt^2 + g_{\alpha\beta} (dx^\alpha - \beta^\alpha dt) (dx^\beta - \beta^\beta dt), \quad (17.8)$$

where $g_{\alpha\beta}$ is the spatial coordinate metric

$$g_{\alpha\beta} = \gamma_{ab} e^a{}_\alpha e^b{}_\beta. \quad (17.9)$$

Essentially all the tetrad formalism developed in Chapter 11 carries through, subject only to the conditions (17.2) and (17.3). As usual in the tetrad formalism, coordinate indices are lowered and raised with the coordinate metric, tetrad indices are lowered and raised with the tetrad metric, and coordinate and tetrad indices can be transformed to each other with the vierbein and its inverse.

The vierbein coefficient α is called the **lapse**, while β^α is called the **shift**. Physically, the lapse α is the rate at which the proper time τ of the tetrad rest frame elapses per unit coordinate time t , while the shift β^α is the velocity at which the tetrad rest frame moves through the spatial coordinates x^α per unit coordinate time t ,

$$\alpha = \frac{d\tau}{dt}, \quad \beta^\alpha = \frac{dx^\alpha}{dt}. \quad (17.10)$$

These relations (17.10) follow from the fact that the 4-velocity in the tetrad rest frame is by definition $u^m \equiv \{1, 0, 0, 0\}$, so the coordinate 4-velocity $u^\mu \equiv e_m{}^\mu u^m$ of the tetrad rest frame is

$$\frac{dx^\mu}{d\tau} \equiv u^\mu = e_0{}^\mu = \frac{1}{\alpha} \{1, \beta^\alpha\}. \quad (17.11)$$

The proper time derivative $d/d\tau$ in the tetrad rest frame is just equal to the directed derivative ∂_0 in the time direction γ_0 ,

$$\frac{d}{d\tau} = u^\mu \frac{\partial}{\partial x^\mu} = \partial_0. \quad (17.12)$$

Coordinate and tetrad derivatives $\partial/\partial x^\mu$ and ∂_m are related to each other as usual by the vierbein and its inverse,

$$\frac{\partial}{\partial t} = \alpha \partial_0 - \beta^a \partial_a, \quad \partial_0 = \frac{1}{\alpha} \left(\frac{\partial}{\partial t} + \beta^\alpha \frac{\partial}{\partial x^\alpha} \right), \quad (17.13a)$$

$$\frac{\partial}{\partial x^\alpha} = e^\alpha_a \partial_a, \quad \partial_a = e_a^\alpha \frac{\partial}{\partial x^\alpha}, \quad (17.13b)$$

where $\beta^a \equiv e^a_\alpha \beta^\alpha$. By construction, the only coordinate derivative involving the directed time derivative ∂_0 is the coordinate time derivative $\partial/\partial t$, and conversely the only directed derivative involving a coordinate time derivative $\partial/\partial t$ is the directed time derivative ∂_0 .

Concept question 17.1 Does Nature pick out a preferred foliation of time? In the ADM formalism, spacetime must be foliated into spacelike hypersurfaces of constant time, but the choice of foliation can be made arbitrarily. Does Nature pick out any particular foliation? **Answer.** Yes, apparently. The Cosmic Microwave Background defines a preferred frame of reference in cosmology. What caused the preferred frame to be established is mysterious, but it must have happened during or before early inflation, when the different parts of what became our Universe were in causal contact.

17.1.1 Traditional ADM approach

The traditional ADM approach sets the spatial tetrad axes γ_a equal to the spatial coordinate tangent axes e_α ,

$$\gamma_a = \delta_a^\alpha e_\alpha \quad (\text{traditional ADM}), \quad (17.14)$$

equivalent to choosing the spatial vierbein to be the unit matrix, $e_a^\alpha = \delta_a^\alpha$. It is natural however to extend the ADM approach into a full tetrad approach, allowing the spatial tetrad axes γ_a to be chosen more generally, subject only to the condition (17.3) that they be orthogonal to the tetrad time axis, and therefore lie in the hypersurface of constant time t . For example, the spatial tetrad γ_a can be chosen to form 3D orthonormal axes, $\gamma_{ab} \equiv \gamma_a \cdot \gamma_b = \delta_{ab}$, so that the full 4D tetrad metric γ_{mn} is Minkowski.

This Chapter follows the full tetrad approach to the ADM formalism, but all the results hold for the traditional case where the spatial tetrad axes are set equal to the coordinate spatial axes, equation (17.14).

Bianchi spacetimes, discussed in §17.4, provide an illustrative example of the application of the ADM formalism to a case where it is advantageous to choose the tetrad to be neither orthonormal nor equal to the coordinate tangent axes.

17.1.2 Spatial vectors and tensors

Since the tetrad time axis γ_0 in the ADM formalism is defined uniquely by the choice of hypersurfaces of constant time t , there is no freedom of tetrad transformations of the time axis distinct from temporal coordinate transformations (no distinct freedom of Lorentz boosts). However, there is still freedom of coordinate

transformations of the spatial coordinate axes e_α , and tetrad transformations of the spatial tetrad axes γ_a (spatial rotations).

A covariant **spatial coordinate vector** A_α is defined to be a vector that transforms like the spatial coordinate axes e_α . Likewise a covariant **spatial tetrad vector** A_a is defined to be a vector that transforms like the spatial tetrad axes γ_a . The usual apparatus of vectors and tensors carries through. For spatial tensors, coordinate and tetrad spatial indices are lowered and raised with respectively the spatial coordinate and tetrad 3-metrics $g_{\alpha\beta}$ and γ_{ab} and their inverses, and spatial indices are transformed between coordinate and tetrad frames with the spatial vierbein e^a_α and its inverse.

17.1.3 ADM gravitational coordinates and momenta

The ADM formalism follows the conventional Hamiltonian approach of regarding the velocities of the fields as being their time derivatives $\partial/\partial t$ (as opposed to their 4-gradients $\partial/\partial x^\kappa$), and the momenta as derivatives of the Lagrangian with respect to these velocities.

If the Lorentz connections $\Gamma_{mn\lambda}$ are taken to be the coordinates of the gravitational field, then the corresponding conjugate momenta are, equation (16.89) with the factor 8π replaced by $16\pi\alpha$ for convenience,

$$16\pi\alpha \frac{\delta L_g}{\delta(\partial\Gamma_{mn\lambda}/\partial t)} = \alpha(e^{mt}e^{n\lambda} - e^{m\lambda}e^{nt}) . \quad (17.15)$$

But ADM imposes $e^{at} = 0$, equations (17.6), so for the momentum to be non-vanishing, one of m or n , say n , must be the tetrad time index 0. Since the momentum is antisymmetric in mn , the other tetrad index m must be a spatial tetrad index a . Moreover since the momentum is antisymmetric in $t\lambda$, the coordinate index λ must be a spatial coordinate index α . Finally, with $e^{0t} = -1/\alpha$, the non-vanishing momenta conjugate to the Lorentz connections are

$$16\pi\alpha \frac{\delta L_g}{\delta(\partial\Gamma_{a0\alpha}/\partial t)} = e^{a\alpha} . \quad (17.16)$$

This shows that the coordinates $\Gamma_{mn\lambda}$ with non-vanishing conjugate momenta are $\Gamma_{a0\alpha}$ with middle (or first) index the tetrad time index 0 and the other two indices spatial, and that the momenta conjugate to these coordinates are the spatial vierbein $e^{a\alpha}$.

If on the other hand the vierbein $e^{n\lambda}$ are taken to be the coordinates of the gravitational field, then the corresponding canonically conjugate momenta are, with a factor of $8\pi\alpha$ thrown in for convenience,

$$8\pi\alpha \frac{\delta L'_g}{\delta(\partial e^{n\lambda}/\partial t)} = \alpha e^{mt} \pi_{nm\lambda} , \quad (17.17)$$

where $\pi_{nm\lambda}$ are related to the Lorentz connections $\Gamma_{nm\lambda}$ by equation (16.114). But again ADM imposes $e^{at} = 0$, so the tetrad index m must be the time tetrad index 0. Since $\pi_{nm\lambda}$ is antisymmetric in its first two indices, the tetrad index n must be a spatial tetrad index a . And then the non-vanishing of the coordinate $e^{n\lambda} = e^{a\lambda}$ requires that λ also be a spatial coordinate index β . Thus the non-vanishing momenta conjugate

to the vierbein coordinates are

$$8\pi\alpha \frac{\delta L'_g}{\partial(\partial e^{a\beta}/\partial t)} = -\pi_{a0\beta} , \quad (17.18)$$

in which the momenta $\pi_{a0\beta}$ are related to the Lorentz connections $\Gamma_{a0\beta}$ by, from equations (16.114) with $e_{0\beta} = 0$,

$$\pi_{a0\beta} \equiv \Gamma_{a0\beta} - e_{a\beta}\Gamma_{0c}^c , \quad \Gamma_{a0\beta} = \pi_{a0\beta} - \frac{1}{2}e_{a\beta}\pi_{0c}^c . \quad (17.19)$$

This shows that the coordinates $e^{n\lambda}$ with non-vanishing conjugate momenta are the spatial vierbein $e^{a\alpha}$, and that the momenta conjugate to these coordinates are $\pi_{a0\beta}$ with middle (or first) index the tetrad time index 0 and the other two indices spatial.

As remarked before equation (16.116), the same equations of motion are obtained whether the action is varied with respect to either $\pi_{a0\beta}$ or $\Gamma_{a0\beta}$, so one can choose either $\pi_{a0\beta}$ or $\Gamma_{a0\beta}$ as the momentum variables conjugate to the coordinates $e^{a\alpha}$. The original choice of Arnowitt, Deser, and Misner (1963) was $\pi_{a0\beta}$, but equations using $\Gamma_{a0\beta}$ were proposed by Smarr and York (1978) and York (1979).

A reminder: do not confuse the Lorentz connections $\Gamma_{mn\lambda}$ (of which there are 24) with the coordinate connections $\Gamma_{\mu\nu\lambda}$ (of which there are 40, for vanishing torsion). The Lorentz connections $\Gamma_{mn\lambda}$ with final index a coordinate index λ are related to the Lorentz connections Γ_{mnl} with all tetrad indices by, equation (15.20),

$$\Gamma_{mn\lambda} \equiv e^l{}_{\lambda}\Gamma_{mnl} . \quad (17.20)$$

17.1.4 ADM acceleration and extrinsic curvature

In the previous subsection 17.1.3 it was found that, given the choice (17.6) of ADM vierbein, the momentum variables that emerge naturally are the Lorentz connections Γ_{a0b} whose middle (or first) index is the tetrad time index 0, and whose other two indices ab are both spatial indices. This set of Lorentz connections is called the **extrinsic curvature**, commonly denoted K_{ab} . As will be shown momentarily, the extrinsic curvature K_{ab} is a spatial tetrad tensor. The other set of Lorentz connections that transforms like a spatial tensor are the connections Γ_{a00} , which are called the **acceleration** K_a . The combined set of connections with middle index 0 is called the **generalized extrinsic curvature** $K_{m0l} \equiv \Gamma_{m0l}$. The non-vanishing components of the generalized extrinsic curvature constitute the acceleration and the extrinsic curvature (the remaining components vanish, $\Gamma_{000} = \Gamma_{00a} = 0$):

$$K_a \equiv K_{a00} \equiv \Gamma_{a00} \equiv \boldsymbol{\gamma}_a \cdot \partial_0 \boldsymbol{\gamma}_0 \quad \text{a spatial vector} , \quad (17.21a)$$

$$K_{ab} \equiv K_{a0b} \equiv \Gamma_{a0b} \equiv \boldsymbol{\gamma}_a \cdot \partial_b \boldsymbol{\gamma}_0 \quad \text{a spatial tensor} . \quad (17.21b)$$

The acceleration K_a and the extrinsic curvature K_{ab} form a spatial vector and tensor because the time axis $\boldsymbol{\gamma}_0$ is a spatial scalar, so its derivatives $\partial_0 \boldsymbol{\gamma}_0$ and $\partial_b \boldsymbol{\gamma}_0$ constitute respectively a spatial scalar and a spatial vector. The vanishing of the ADM tetrad metric γ_{0a} with one time index 0 and one spatial index a , equation (17.4), implies that the generalized extrinsic curvature is antisymmetric in its first two indices,

$$K_{0al} = -K_{a0l} , \quad (17.22)$$

which remains true even in the traditional ADM case, equation (17.14), where the spatial tetrad metric γ_{ab} is not constant. The unique non-vanishing contraction of the generalized extrinsic curvature is

$$K_n \equiv K_{nm}^m = \{K_{0m}^m, K_{am}^m\} = \{K_0, K_a\} , \quad (17.23)$$

whose space part is the acceleration K_a , and whose time part is the trace K of the extrinsic curvature K_{ab} ,

$$K_0 = K \equiv K_a^a . \quad (17.24)$$

The acceleration K_a is justly named because the geodesic equation shows that its contravariant components K^a constitute the acceleration experienced in the tetrad rest frame, where $u^m = \{1, 0, 0, 0\}$,

$$\frac{Du^a}{D\tau} = u^n \partial_n u^a + \Gamma_{mn}^a u^m u^n = K_{00}^a = K^a . \quad (17.25)$$

The extrinsic curvature K_{ab} describes how the unit normal γ_0 to the 3-dimensional spatial hypersurface of constant time changes over the hypersurface, and can therefore be regarded as embodying the curvature of the 3-dimensional spatial hypersurface embedded in the 4-dimensional spacetime.

Momenta π_{ab} analogous to those defined by equations (17.19) are related to the extrinsic curvatures K_{ab} by

$$\pi_{ab} = K_{ab} - \gamma_{ab} K , \quad K_{ab} = \pi_{ab} - \frac{1}{2} \gamma_{ab} \pi , \quad (17.26)$$

where $\pi \equiv \pi_a^a = -2K$ is the trace of π_{ab} .

17.1.5 Decomposition of connections and curvatures

As seen in the previous subsection 17.1.4, the Lorentz connections decompose into a part, the generalized extrinsic curvature $K_{m0l} \equiv \Gamma_{m0l}$ with middle (or first) index the tetrad time index 0, that transforms like a tensor under spatial tetrad transformations, and a remainder, the restricted connections $\hat{\Gamma}_{abl} \equiv \Gamma_{abl}$ with first two indices ab spatial, that does not transform like a spatial tensor,

$$\Gamma_{mnl} = \hat{\Gamma}_{mnl} + K_{mnl} . \quad (17.27)$$

Although the acceleration and extrinsic curvature arise in the first instance as Lorentz connections, for which the tetrad metric γ_{mn} is constant, it is useful to allow a more general situation in which the spatial tetrad metric γ_{ab} is arbitrary. Whereas K_{mnl} is necessarily antisymmetric in its first two indices mn , equation (17.22), the restricted connection $\hat{\Gamma}_{mnl}$ need not be (it is antisymmetric in its first two indices if the spatial tetrad metric γ_{ab} is constant, but not for example in the traditional ADM case (17.14) where the spatial tetrad metric equals the spatial coordinate metric). The vanishing components of $\hat{\Gamma}_{mnl}$ and K_{mnl} are

$$\hat{\Gamma}_{a0l} = \hat{\Gamma}_{0al} = 0 , \quad K_{abl} = 0 . \quad (17.28)$$

As a result of the decomposition (17.27) of the connections, the Riemann curvature tensor R_{klmn} decomposes into a restricted part $\hat{R}_{klmn}$, and a part that depends on the generalized extrinsic curvature K_{mnl} .

Rather than specializing immediately to the ADM case, consider the more general situation in which, under some restricted subgroup of tetrad transformations, the tetrad-frame connections Γ_{mnl} decompose

as equation (17.27) into a non-tensorial part $\hat{\Gamma}_{mnl}$ and a tensorial part K_{mnl} . The resemblance of the decomposition (17.27) to the split (11.55) between the torsion-free and contortion parts of the tetrad-frame connection is deliberate: in both cases, the tetrad-frame connection Γ_{mnl} is decomposed into non-tensorial and tensorial parts. The resulting decomposition of the Riemann curvature tensor is consequently quite similar in the two cases. However, here K_{mnl} is not the contortion, but rather some part of the tetrad-frame connections that is tensorial under the restricted group of tetrad transformations.

The unique non-vanishing contraction of the tensor K_{mnl} is the vector

$$K_n \equiv K_{nl}^l . \quad (17.29)$$

The placement of indices in equation (17.29) follows the usual convention for general relativistic connections, that $K_{nl}^k = \gamma^{km} K_{mnl}$.

The restricted tetrad-frame derivative $\hat{D}_k$ with restricted tetrad-frame connection $\hat{\Gamma}_{mnl}$ is a covariant derivative with respect to the restricted group of tetrad transformations. Since the generalized extrinsic curvature K_{mnl} is a tensor with respect to the restricted group, its restricted covariant derivative is also a restricted tensor. Among other things, this implies that the restricted covariant derivatives $\hat{D}_k$ of the vanishing components (17.28) of K_{mnl} vanish identically.

The tetrad metric γ_{lm} commutes by construction with the total covariant derivative D_k , and it also commutes (even when the tetrad metric is not constant) with the restricted covariant derivative $\hat{D}_k$, as follows from

$$0 = D_k \gamma_{lm} = \hat{D}_k \gamma_{lm} - K_{lk}^n \gamma_{nm} - K_{mk}^n \gamma_{ln} = \hat{D}_k \gamma_{lm} - K_{mlk} - K_{lmk} = \hat{D}_k \gamma_{lm} , \quad (17.30)$$

the last step of which is a consequence of the antisymmetry of the extrinsic curvature in its first two indices. Therefore tensors involving the restricted covariant derivative can be contracted in the usual way.

In ADM, the extrinsic curvature is tensorial not only with respect to spatial tetrad transformations, but also with respect to spatial coordinate transformations. In this case, the restricted covariant derivative $\hat{D}_k$ commutes not only with the tetrad metric, equation (17.30), but also with the vierbein $e^m{}_\mu$ and its inverse $e_m{}^\mu$,

$$0 = D_k e^m{}_\mu = \hat{D}_k e^m{}_\mu + K_{nk}^m e^n{}_\mu - K_{\mu k}^\nu e^m{}_\nu = \hat{D}_k e^m{}_\mu + K_{\mu k}^m - K_{\mu k}^m = \hat{D}_k e^m{}_\mu . \quad (17.31)$$

Therefore, provided that the extrinsic curvature is tensorial with respect to both coordinate and tetrad spatial transformations, tensors involving the restricted covariant derivative can be flipped between coordinate and tetrad indices in the usual way.

The tetrad-frame Riemann tensor R_{klmn} decomposes into a restricted part $\hat{R}_{klmn}$ and a remainder that depends on the generalized extrinsic curvature K_{mnl} and its restricted covariant derivatives. The derivation of the decomposition of the Riemann tensor is most elegant in terms of the multivector Riemann tensor $\mathbf{R}_{\kappa\lambda}$ given by equation (15.25). The decomposition of the Riemann tensor into restricted and extrinsic curvature parts is, analogously to the decomposition (15.49) of the Riemann tensor into torsion-free and contortion

parts,

$$\begin{aligned} \mathbf{R}_{\kappa\lambda} &= \frac{\partial(\hat{\Gamma}_\lambda + \mathbf{K}_\lambda)}{\partial x^\kappa} - \frac{\partial(\hat{\Gamma}_\kappa + \mathbf{K}_\kappa)}{\partial x^\lambda} + \frac{1}{2}[\hat{\Gamma}_\kappa + \mathbf{K}_\kappa, \hat{\Gamma}_\lambda + \mathbf{K}_\lambda] \\ &= \hat{\mathbf{R}}_{\kappa\lambda} + \hat{D}_\kappa \mathbf{K}_\lambda - \hat{D}_\lambda \mathbf{K}_\kappa + \frac{1}{2}[\mathbf{K}_\kappa, \mathbf{K}_\lambda] , \end{aligned} \quad (17.32)$$

where $\mathbf{K}_\kappa \equiv \frac{1}{2}K_{mn\kappa}\gamma^m \wedge \gamma^n$ is the generalized extrinsic curvature vector of bivectors. The restricted Riemann tensor $\hat{\mathbf{R}}_{\kappa\lambda}$ is

$$\hat{\mathbf{R}}_{\kappa\lambda} = \frac{\partial \hat{\Gamma}_\lambda}{\partial x^\kappa} - \frac{\partial \hat{\Gamma}_\kappa}{\partial x^\lambda} + \frac{1}{2}[\hat{\Gamma}_\kappa, \hat{\Gamma}_\lambda] . \quad (17.33)$$

In components, the tetrad-frame Riemann tensor decomposes as

$$R_{klmn} = \hat{R}_{klmn} + \hat{D}_k K_{mnl} - \hat{D}_l K_{mnk} + K_{ml}^p K_{pnk} - K_{mk}^p K_{pnl} + (K_{kl}^p - K_{lk}^p) K_{mnp} . \quad (17.34)$$

The restricted Riemann tensor $\hat{R}_{klmn}$ is

$$\hat{R}_{klmn} = \partial_k \hat{\Gamma}_{mnl} - \partial_l \hat{\Gamma}_{mnk} + \hat{\Gamma}_{ml}^p \hat{\Gamma}_{pnk} - \hat{\Gamma}_{mk}^p \hat{\Gamma}_{pnl} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \hat{\Gamma}_{mnp} . \quad (17.35)$$

Equation (17.35) looks like the usual tetrad-frame formula (11.62), with connections replaced by restricted connections, except that the final term on the right hand side involves the difference $\Gamma_{kl}^p - \Gamma_{lk}^p$ of the full tetrad-frame connection, not just the restricted connection. The part of the Riemann tensor (17.34) that depends on the generalized extrinsic curvature is manifestly antisymmetric in kl and in mn , but it is not necessarily symmetric under $kl \leftrightarrow mn$. Thus the restricted Riemann tensor $\hat{R}_{klmn}$ is antisymmetric in kl and in mn , but not necessarily symmetric under $kl \leftrightarrow mn$.

Contracting the Riemann tensor (17.34) gives the Ricci tensor R_{km} ,

$$R_{km} = \hat{R}_{km} - \hat{D}_k K_m + \hat{D}_n K_{mk}^n - K_{kn}^p K_{mp}^n + K_{mk}^p K_p , \quad (17.36)$$

with $\hat{R}_{km} \equiv \gamma^{ln} \hat{R}_{klmn}$ the restricted Ricci tensor. Contracting the Ricci tensor (17.36) yields the Ricci scalar R ,

$$R = \hat{R} - 2\hat{D}_m K^m - K^{pmn} K_{nmp} - K^p K_p , \quad (17.37)$$

with $\hat{R} \equiv \gamma^{km} \hat{R}_{km}$ the restricted Ricci scalar.

A restricted covariant divergence $\hat{D}_m A^m$ can be converted to a total covariant divergence $D_m A^m$ through

$$D_m A^m = \hat{D}_m A^m + K_{pm}^m A^p = \hat{D}_m A^m + K_p A^p . \quad (17.38)$$

With the restricted covariant divergence converted to a total covariant divergence, the Ricci scalar (17.37) is

$$R = \hat{R} - 2D_m K^m - K^{pmn} K_{nmp} + K^p K_p . \quad (17.39)$$

17.1.6 ADM Riemann and Ricci tensors

For ADM, the components of the Riemann curvature tensor R_{klmn} are, from equations (17.34) with the generalized extrinsic curvature K_{mnl} replaced by the acceleration $K_a \equiv K_{a00}$ and the extrinsic curvature $K_{ab} \equiv K_{a0b}$,

$$R_{a0c0} = \hat{D}_a K_c - \hat{D}_0 K_{ca} + K_a K_c - K_a^b K_{cb} , \quad (17.40a)$$

$$R_{abc0} = \hat{D}_a K_{cb} - \hat{D}_b K_{ca} - (K_{ba} - K_{ab}) K_c \quad (17.40b)$$

$$= R_{c0ab} = \hat{R}_{c0ab} + K_b K_{ac} - K_a K_{bc} , \quad (17.40c)$$

$$R_{abcd} = \hat{R}_{abcd} + K_{ca} K_{db} - K_{cb} K_{da} . \quad (17.40d)$$

Equations (17.40a), (17.40b), (17.40c), and (17.40d) are called respectively the **Ricci**, **Codazzi-Mainardi**, **BSSN**, and **Gauss** equations. After equations of motion have been obtained, the extrinsic curvature K_{ab} will prove to be symmetric (given the ADM gauge condition $e^0_\alpha = 0$, and assuming vanishing torsion), and consequently the final term on the right hand side of equation (17.40b) vanishes. At this point however no equations of motion have yet been obtained: equations are obtained later, §17.2, from variation of the action. If torsion vanishes, then the Riemann tensor R_{klmn} is symmetric in $kl \leftrightarrow mn$, Exercise 11.7. If the tetrad connections are replaced by their usual torsion-free expressions in terms of derivatives of the vierbein, then the symmetries of the Riemann tensor are satisfied identically, so that the right hand sides of the expressions (17.40b) and (17.40c) for R_{abc0} and R_{c0ab} become identical, and one of them can be discarded. In the ADM formalism, equation (17.40c) for R_{c0ab} is discarded as redundant. However, in the BSSN formalism, §17.8, equation (17.40c) is retained as a distinct equation, and some of the equations relating the tetrad connections to derivatives of the vierbein are discarded instead.

The restricted Riemann tensor $\hat{R}_{kla0}$ with one of the final two indices the time index 0 vanishes since $\hat{\Gamma}_{a0l}$ vanishes, equations (17.28),

$$\hat{R}_{kla0} = \partial_k \hat{\Gamma}_{a0l} - \partial_l \hat{\Gamma}_{a0k} + \hat{\Gamma}_{al}^p \hat{\Gamma}_{p0k} - \hat{\Gamma}_{ak}^p \hat{\Gamma}_{p0l} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \hat{\Gamma}_{a0p} = 0 . \quad (17.41)$$

The restricted Riemann tensor $\hat{R}_{klmn}$ with one time 0 index does not satisfy the $kl \leftrightarrow mn$ symmetry of the full Riemann tensor R_{klmn} . The restricted Riemann tensor $\hat{R}_{c0ab}$ with one of the first two indices the time index 0 and the last two indices spatial is

$$\hat{R}_{c0ab} = \partial_c \hat{\Gamma}_{ab0} - \partial_0 \hat{\Gamma}_{abc} + \hat{\Gamma}_{a0}^d \hat{\Gamma}_{dbc} - \hat{\Gamma}_{ac}^d \hat{\Gamma}_{db0} + (\Gamma_{c0}^p - \Gamma_{0c}^p) \hat{\Gamma}_{abp} . \quad (17.42)$$

The restricted Riemann tensor $\hat{R}_{abcd}$ with all spatial indices is

$$\hat{R}_{abcd} = \partial_a \hat{\Gamma}_{cdb} - \partial_b \hat{\Gamma}_{cda} + \hat{\Gamma}_{cb}^e \hat{\Gamma}_{eda} - \hat{\Gamma}_{ca}^e \hat{\Gamma}_{edb} + (\hat{\Gamma}_{ab}^e - \hat{\Gamma}_{ba}^e) \hat{\Gamma}_{cde} + (K_{ab} - K_{ba}) \hat{\Gamma}_{cd0} . \quad (17.43)$$

Again, after equations of motion have been obtained, the extrinsic curvature K_{ab} will proved to be symmetric, equation (17.50), so the final term on the right hand side of equation (17.43) vanishes. Consequently the spatial restricted Riemann tensor $\hat{R}_{abcd}$ depends only on spatial components $\hat{\Gamma}_{abc}$ of the restricted connections and their spatial derivatives (not on the restricted connections $\hat{\Gamma}_{ab0}$ with one time index). For ADM, the restricted spatial connections coincide with the full spatial connections, $\hat{\Gamma}_{abc} = \Gamma_{abc}$, so for ADM the spatial

restricted Riemann curvature equals the Riemann curvature tensor restricted to the 3-dimensional spatial hypersurface of constant time. The spatial restricted Riemann tensor $\hat{R}_{abcd}$ satisfies the usual $ab \leftrightarrow cd$ symmetry.

Contracting the Riemann tensor yields the Ricci tensor R_{km} ,

$$R_{00} = \hat{D}_m K^m - K^{ba} K_{ab} + K^a K_a, \quad (17.44a)$$

$$R_{a0} = -\hat{D}_a K + \hat{D}^b K_{ba} - K^b (K_{ab} - K_{ba}) \quad (17.44b)$$

$$= R_{0a} = \hat{R}_{0a} - K^b K_{ab} + K_a K, \quad (17.44c)$$

$$R_{ab} = \hat{R}_{ab} - \hat{D}_a K_b + \hat{D}_0 K_{ba} + K_{ba} K - K_a K_b. \quad (17.44d)$$

If torsion vanishes, then the Ricci tensor R_{km} is symmetric. Again, if the tetrad connections are replaced by their torsion-free expressions in terms of vierbein derivatives, then the symmetry of the Ricci tensor is satisfied identically, so that two expressions (17.44b) and (17.44c) are identical, and one of them can be discarded as redundant. In the ADM formalism, equation (17.44c) is discarded. In the BSSN formalism however, §17.8, equation (17.44c) is retained, and some of the equations relating the tetrad connections to derivatives of the vierbein are discarded instead. Like the restricted Riemann tensor, the restricted Ricci tensor $\hat{R}_{km}$ with one time 0 index is not symmetric. While $\hat{R}_{a0}$ vanishes, $\hat{R}_{0a}$ does not. The purely spatial Ricci tensor $\hat{R}_{ab}$ is on the other hand symmetric in ab . For ADM, the purely spatial Ricci tensor $\hat{R}_{ab}$ is the Ricci tensor restricted to the 3-dimensional spatial hypersurface of constant time.

Contracting the Ricci tensor yields the Ricci scalar R ,

$$R = \hat{R} - 2\hat{D}_m K^m + K^{ba} K_{ab} + K^2 - 2K^a K_a. \quad (17.45)$$

For ADM, the restricted Ricci scalar $\hat{R}$ is the Ricci scalar restricted to the 3-dimensional spatial hypersurface of constant time.

Converting the restricted covariant divergence $\hat{D}_m K^m$ to a total covariant derivative $D_m K^m$ using equation (17.38) brings the Ricci scalar to

$$R = \hat{R} - 2D_m K^m + K^{ba} K_{ab} - K^2. \quad (17.46)$$

At this point it is common to argue that the covariant divergence $D_m K^m$ has no effect on equations of motion, so can be dropped from the Ricci scalar, yielding the so-called ADM Lagrangian

$$L_{\text{ADM}} = \frac{1}{16\pi} \left(\hat{R} + K^{ba} K_{ab} - K^2 \right). \quad (17.47)$$

The ADM Lagrangian (17.47) is fine as a Lagrangian, but it is not in Hamiltonian form. Rather, the ADM Lagrangian (17.47) is in a form analogous to the quadratic Lagrangian (16.160). As discussed in §16.12.1, the quadratic Lagrangian is valid provided that the tetrad connections satisfy their equations of motion (in particle physics jargon, the tetrad connections are “on shell”).

The original purpose of the ADM formalism was to bring the gravitational Lagrangian into (conventional) Hamiltonian form. As has been seen, §16.7, the Hilbert Lagrangian is already in (super-)Hamiltonian form. In dropping the covariant divergence $D_m K^m$ to arrive at the Lagrangian (17.47), one both implicitly assumes

that the equation of motion for K^m is satisfied, and loses the ability to derive that equation of motion from the Lagrangian. One could attempt to recover the Hamiltonian form of the Lagrangian from the ADM Lagrangian (17.47), which would involve re-assuming the equation of motion for K^m , but such a procedure (widely repeated in the physics literature) seems like shooting oneself in the foot. A sensible approach is to stick with the Hilbert Lagrangian, which is already in (super-)Hamiltonian form. The (super-)Hamiltonian approach has already identified the gravitational coordinates and momenta for ADM, §17.1.3, and it also supplies the equations of motion for ADM, §17.2.

17.2 ADM gravitational equations of motion

As shown in §17.1.3, the gravitational coordinates and momenta in the ADM formalism are the spatial components $e^{a\beta}$ of the vierbein, and the extrinsic curvatures $K_{a\beta} \equiv \Gamma_{a0\beta}$, equation (17.21b) (or alternatively, in place of $K_{a\beta}$, the trace-corrected extrinsic curvatures $\pi_{a\beta}$ defined by equations (17.26)).

Gravitational equations of motion in the ADM formalism follow from varying the Hilbert action. All the equations obtained from varying the Hilbert action in super-Hamiltonian form continue to hold in the ADM formalism, namely the 24 equations for the (torsion-free) Lorentz connections, and the 10 Einstein equations (the Einstein tensor is symmetric if torsion vanishes). The difference is that only some of the equations, namely those that come from varying the action with respect to the gravitational coordinates and momenta $e^{a\beta}$ and $K_{a\beta}$, are interpreted as equations of motion that determine the time evolution of those coordinates and momenta. The remaining equations are interpreted either as identities (in the case of the Lorentz connections), or as constraints (in the case of the Einstein equations). A constraint equation is one that must be satisfied in the initial conditions, but is thereafter guaranteed to be satisfied by conservation laws, here conservation of energy-momentum, guaranteed by the contracted Bianchi identities.

Because the tetrad in this Chapter is being allowed a general form, with not necessarily constant tetrad metric, the connections are not necessarily Lorentz connections, and the relation between the connections and derivatives of the vierbein and metric, equation (11.53), is more general than that derived from an action principle in Chapter 16. Suffice to say that the relation can be derived from an action principle, but that will not be done here.

17.2.1 ADM connections

Start by considering the equations of motion for the tetrad-frame connections, determined by varying the Hilbert action with respect to the connections. The connections are given by the usual expressions (11.53) in terms of the vierbein derivatives d_{lmn} defined by equation (11.33) (equations (11.53) allow for a non-constant spatial tetrad metric γ_{ab} , thus admitting the traditional ADM approach in which the spatial tetrad γ_a are set equal to the spatial coordinate tangent axes e_α , equation (17.14)). The non-vanishing tetrad connections

are, from the general formula (11.53) with vanishing torsion (note that $d_{0an} = 0$ since $e^0_{\alpha} = 0$),

$$\Gamma_{a00} = -\Gamma_{0a0} \equiv K_a = -d_{00a} , \quad (17.48a)$$

$$\Gamma_{a0b} = -\Gamma_{0ab} \equiv K_{ab} = \frac{1}{2} (\partial_0 \gamma_{ab} + d_{ab0} + d_{ba0} - d_{a0b} - d_{b0a}) , \quad (17.48b)$$

$$\Gamma_{ab0} \equiv \hat{\Gamma}_{ab0} = K_{ab} - d_{ab0} + d_{a0b} , \quad (17.48c)$$

$$\Gamma_{abc} \equiv \hat{\Gamma}_{abc} = \text{same as eq. (11.53)} , \quad (17.48d)$$

where the relevant vierbein derivatives d_{lmn} are

$$d_{00a} = -\frac{1}{\alpha} \partial_a \alpha , \quad d_{a0b} = -\frac{1}{\alpha} e_{a\alpha} \partial_b \beta^{\alpha} , \quad d_{ab0} = \gamma_{ac} e_b^{\beta} \partial_0 e^c_{\beta} . \quad (17.49)$$

Equation (17.48b) shows that the extrinsic curvature is symmetric,

$$K_{ab} = K_{ba} \quad (17.50)$$

(and consequently so also is the momentum π_{ab} , equations (17.26)). The symmetry of the extrinsic curvature is a consequence of the ADM gauge choice $e^0_{\alpha} = 0$ along with the assumption of vanishing torsion. The connections (17.48a) and (17.48b) form, as remarked after equations (17.21), a spatial tetrad vector the acceleration K_a , and a spatial tetrad tensor the extrinsic curvature K_{ab} , but the remaining connections (17.48c) and (17.48d) are not spatial tetrad tensors. Note that the purely spatial tetrad connections Γ_{abc} , like the spatial tetrad axes γ_a , transform under temporal coordinate transformations despite the absence of temporal indices. If the spatial tetrad metric γ_{ab} is taken to be constant, which is true if for example the spatial tetrad axes γ_a are taken to be orthonormal, then the tetrad connections Γ_{ab0} and Γ_{abc} , equations (17.48c) and (17.48d), are antisymmetric in their first two indices. However, equations (17.48) are valid in general, including in the traditional case where the spatial axes are taken equal to the spatial coordinate tangent axes, equation (17.14), in which case Γ_{ab0} and Γ_{abc} are not antisymmetric in their first two indices.

In the ADM formalism, an equation of motion for the ADM spatial coordinate metric $g_{\alpha\beta}$ follows from the vanishing of the restricted covariant time derivative of the spatial tetrad metric γ_{ab} , equation (17.30),

$$\boxed{\hat{D}_0 \gamma_{ab} = 0} . \quad (17.51)$$

With the expressions (17.48c) for the connections $\hat{\Gamma}_{ab0}$, the covariant time derivative is

$$\begin{aligned} \hat{D}_0 \gamma_{ab} &= \partial_0 \gamma_{ab} - \hat{\Gamma}_{a0}^c \gamma_{cb} - \hat{\Gamma}_{b0}^c \gamma_{ca} \\ &= \partial_0 \gamma_{ab} + d_{ab0} + d_{ba0} - d_{a0b} - d_{b0a} - 2K_{ab} . \end{aligned} \quad (17.52)$$

The time derivatives in expression (17.52) are the directed time derivatives $\partial_0 \gamma_{ab}$ of the spatial tetrad metric (the tetrad metric γ_{ab} is not being assumed constant, so as to allow the traditional ADM approach, equation (17.14)), and the directed time derivatives $d_{\beta 0}^c \equiv \partial_0 e^c_{\beta}$ of the spatial vierbein. These time derivatives appear in the expression (17.52) only in the combination

$$\partial_0 \gamma_{ab} + d_{ab0} + d_{ba0} = e_a^{\alpha} e_b^{\beta} \partial_0 (\gamma_{cd} e^c_{\alpha} e^d_{\beta}) = e_a^{\alpha} e_b^{\beta} \partial_0 g_{\alpha\beta} . \quad (17.53)$$

Thus the equation of motion (17.51) effectively governs the time evolution of not all 9 components of the

spatial vierbein $e_{a\beta}$, but rather only the 6 components $g_{\alpha\beta}$ of the spatial coordinate metric, equation (17.9). Recast in the coordinate frame, the equation of motion (17.51) is

$$\frac{\partial g_{\alpha\beta}}{\partial t} + \beta^\gamma \frac{\partial g_{\alpha\beta}}{\partial x^\gamma} + g_{\alpha\gamma} \frac{\partial \beta^\gamma}{\partial x^\beta} + g_{\beta\gamma} \frac{\partial \beta^\gamma}{\partial x^\alpha} = 2\alpha K_{\alpha\beta} . \quad (17.54)$$

Equation (17.54) may also be written

$$\boxed{\mathcal{L}_u g_{\alpha\beta} = \frac{1}{\alpha} \left(\frac{\partial g_{\alpha\beta}}{\partial t} + \hat{\mathcal{L}}_\beta g_{\alpha\beta} \right) = 2K_{\alpha\beta}} , \quad (17.55)$$

where $\mathcal{L}_u$ denotes the Lie derivative (7.157) with respect to the 4-velocity $u^\mu = \{1/\alpha, \beta^\gamma/\alpha\}$, equation (17.11), and $\hat{\mathcal{L}}_\beta$ denotes the Lie derivative (7.157) with respect to the shift β^γ , restricted to the hypersurface of constant time (hence the restricted $\hat{\mathcal{L}}$ overscript),

$$\hat{\mathcal{L}}_\beta g_{\alpha\beta} = \beta^\gamma \frac{\partial g_{\alpha\beta}}{\partial x^\gamma} + g_{\alpha\gamma} \frac{\partial \beta^\gamma}{\partial x^\beta} + g_{\beta\gamma} \frac{\partial \beta^\gamma}{\partial x^\alpha} . \quad (17.56)$$

As is usual with a Lie derivative, equation (7.158), the coordinate derivatives $\partial/\partial x^\alpha$ in equation (17.56) can be replaced, if desired, by the restricted covariant derivatives $\hat{D}_\alpha$. Since the restricted covariant derivative of the spatial coordinate metric $g_{\alpha\beta}$ vanishes, the Lie derivative $\hat{\mathcal{L}}_\beta g_{\alpha\beta}$ can be written (compare equation (7.160)),

$$\hat{\mathcal{L}}_\beta g_{\alpha\beta} = \hat{D}_\beta \beta_\alpha + \hat{D}_\alpha \beta_\beta . \quad (17.57)$$

The spatial trace of equation (17.52) provides an equation of motion for the determinant $\gamma \equiv |\gamma_{ab}|$ of the spatial tetrad metric, since $\gamma^{ab} \partial_0 \gamma_{ab} = \partial_0 \ln \gamma$. With, from equations (17.49),

$$d_{0a}^a = -\frac{1}{\alpha} e^a{}_\alpha \partial_a \beta^\alpha = -\frac{1}{\alpha} \frac{\partial \beta^\alpha}{\partial x^\alpha} , \quad d_{a0}^a = e_a{}^\beta \partial_0 e^\alpha{}_\beta = \partial_0 \ln e , \quad (17.58)$$

where $e \equiv |e^a{}_\alpha|$ is the determinant of the spatial vierbein, the spatial trace of equation (17.52) provides the equation of motion

$$\partial_0 \ln(\gamma e^2) + \frac{2}{\alpha} \frac{\partial \beta^\alpha}{\partial x^\alpha} = 2K . \quad (17.59)$$

In the coordinate frame, the trace equation is (see equation (7.24) for the Lie derivative of a metric determinant)

$$\mathcal{L}_u \ln g = \frac{1}{\alpha} \left(\frac{\partial \ln g}{\partial t} + \beta^\alpha \frac{\partial \ln g}{\partial x^\alpha} + 2 \frac{\partial \beta^\alpha}{\partial x^\alpha} \right) = 2K , \quad (17.60)$$

where $g = |g_{\alpha\beta}| = \gamma e^2$ is the determinant of the coordinate-frame spatial metric.

The expression (17.48b) for the extrinsic curvature K_{ab} has thus provided an equation of motion (17.55) for the spatial ADM metric $g_{\alpha\beta}$. Of the remaining connections (17.48), the acceleration K_a , equation (17.48a), and the purely spatial connections Γ_{abc} , equation (17.48d), involve only spatial derivatives of the vierbein, not time derivatives. These connections are needed in the ADM equations, but are treated as identities rather than equations of motion. That is, the equation of motion (17.55) determines the time evolution of the spatial vierbein $e_{a\beta}$, or rather of the spatial coordinate metric $g_{\alpha\beta}$, which is the quadratic combination (17.9) of

the spatial vierbein. With the spatial vierbein on a hypersurface of constant time determined, their spatial derivatives on the hypersurface follow. These spatial derivatives of the vierbein determine the acceleration K_a and purely spatial connections Γ_{abc} through equations (17.48a) and (17.48d).

The final set of connections is Γ_{ab0} , equation (17.48c). These connections do depend on time derivatives, and they do appear in the equations of motion (17.52) and (17.63), but they cease to appear explicitly when the equations of motion are expressed as equations for the coordinate metric $g_{\alpha\beta}$ and the coordinate-frame extrinsic curvature $K_{\alpha\beta}$, equations (17.55) and (17.68).

17.2.2 ADM Einstein equations

The Einstein equations follow from varying the Hilbert action with respect to the vierbein $e^{k\kappa}$, equation (16.104). In the ADM formalism, only the spatial Einstein equations, which come from varying with respect to the spatial vierbein $e^{a\alpha}$, are interpreted as equations of motion governing the time evolution of the system. The remaining equations are interpreted as constraints, §17.2.3.

The spatial Einstein equations are

$$G_{ab} = 8\pi T_{ab} , \quad (17.61)$$

which are symmetric for vanishing torsion. Equivalently, with the trace $R = -8\pi T$ transferred to the right hand side,

$$R_{ab} = 8\pi \left(T_{ab} - \frac{1}{2} \gamma_{ab} T \right) . \quad (17.62)$$

Substituting the spatial Ricci tensor from equation (17.44d) transforms the spatial Einstein equations (17.62) into equations of motion for the extrinsic curvature K_{ab} ,

$$\boxed{\hat{D}_0 K_{ab} = \hat{D}_a K_b - K_{ab} K + K_a K_b - \hat{R}_{ab} + 8\pi \left(T_{ab} - \frac{1}{2} \gamma_{ab} T \right)} . \quad (17.63)$$

The restricted covariant time derivative $\hat{D}_0 K_{ba}$ on the left hand side of equation (17.63) is, with formula (17.48c) for the connections $\hat{\Gamma}_{ab0}$,

$$\begin{aligned} \hat{D}_0 K_{ab} &= \partial_0 K_{ab} - \hat{\Gamma}_{b0}^c K_{ca} - \hat{\Gamma}_{a0}^c K_{cb} \\ &= \partial_0 K_{ab} + d_{cb0} K_a^c + d_{ca0} K_b^c - d_{c0b} K_a^c - d_{c0a} K_b^c - 2K_a^c K_{cb} . \end{aligned} \quad (17.64)$$

The time derivatives in equation (17.64) are the directed time derivatives $\partial_0 K_{ab}$ of the extrinsic curvature, and the directed time derivatives $d_{\beta 0}^c \equiv \partial_0 e_{\beta}^c$ of the spatial vierbein. These time derivatives appear in the expression (17.64) only in a combination analogous to that in equation (17.53),

$$\partial_0 K_{ab} + d_{cb0} K_a^c + d_{ca0} K_b^c = e_a^{\alpha} e_b^{\beta} \partial_0 K_{\beta\alpha} . \quad (17.65)$$

Just as equation (17.53) picked out the spatial coordinate metric $g_{\alpha\beta}$, so also equation (17.65) picks out the coordinate-frame extrinsic curvature $K_{\alpha\beta}$ as the fundamental object whose time evolution is being governed. Recast in the coordinate frame using equation (17.65), equation (17.64) is

$$\hat{D}_0 K_{\alpha\beta} = \mathcal{L}_u K_{\alpha\beta} - 2K_{\alpha}^{\gamma} K_{\gamma\beta} = \frac{1}{\alpha} \left(\frac{\partial K_{\alpha\beta}}{\partial t} + \hat{\mathcal{L}}_{\beta} K_{\alpha\beta} \right) - 2K_{\alpha}^{\gamma} K_{\gamma\beta} . \quad (17.66)$$

where again $\mathcal{L}_u$ denotes the Lie derivative (7.157) with respect to the 4-velocity u^μ , equation (17.11), and $\hat{\mathcal{L}}_\beta$ denotes the Lie derivative (7.157) with respect to the shift β^γ , restricted to the hypersurface of constant time,

$$\hat{\mathcal{L}}_\beta K_{\alpha\beta} = \beta^\gamma \frac{\partial K_{\alpha\beta}}{\partial x^\gamma} + K_{\gamma\alpha} \frac{\partial \beta^\gamma}{\partial x^\beta} + K_{\gamma\beta} \frac{\partial \beta^\gamma}{\partial x^\alpha} . \quad (17.67)$$

As usual with a Lie derivative, equation (7.157), the coordinate derivatives $\partial/\partial x^\alpha$ in equation (17.67) can be replaced, if desired, by the restricted covariant derivatives $\hat{D}_\alpha$. Substituting equation (17.66) into equation (17.63) brings the equation of motion for the coordinate-frame extrinsic curvature to

$$\boxed{\mathcal{L}_u K_{\alpha\beta} = \frac{1}{\alpha} \left(\frac{\partial K_{\alpha\beta}}{\partial t} + \hat{\mathcal{L}}_\beta K_{\alpha\beta} \right) = \hat{D}_\alpha K_\beta + 2K_\alpha^\gamma K_{\gamma\beta} - K_{\alpha\beta} K + K_\alpha K_\beta - \hat{R}_{\alpha\beta} + 8\pi (T_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} T)} . \quad (17.68)$$

All the terms in equation (17.68) are manifestly symmetric in $\alpha\beta$ except for $\hat{D}_\alpha K_\beta$, but this too is symmetric, for vanishing torsion, as follows from

$$\hat{D}_\alpha K_\beta = \frac{\partial^2 \ln \alpha}{\partial x^\alpha \partial x^\beta} - \hat{\Gamma}_{\beta\alpha}^\gamma K_\gamma = \hat{D}_\beta K_\alpha , \quad (17.69)$$

the coordinate connection $\hat{\Gamma}_{\beta\alpha}^\gamma$ being symmetric in its last two indices, for vanishing torsion. Equations (17.55) and (17.68) constitute the two fundamental sets of equations of motion for the coordinates $g_{\alpha\beta}$ and momenta $K_{\alpha\beta}$ in the ADM formalism.

The spatial trace of equation (17.63) (which is straightforward to take because the tetrad metric γ_{ab} commutes with the restricted covariant derivative $\hat{D}_k$) is

$$\partial_0 K = \hat{D}_a K^a - K^2 + K^a K_a - \hat{R} + 12\pi(\rho - p) , \quad (17.70)$$

where the spatial trace $T_\alpha^\alpha = 3p$ defines the proper monopole pressure p , and the full spacetime trace is $T = -\rho + 3p$, with ρ the proper energy density. In the coordinate frame, equation (17.70) becomes

$$\mathcal{L}_u K = \frac{1}{\alpha} \left(\frac{\partial K}{\partial t} + \beta^\alpha \frac{\partial K}{\partial x^\alpha} \right) = \hat{D}_\alpha K^\alpha - K^2 + K^\alpha K_\alpha - \hat{R} + 12\pi(\rho - p) . \quad (17.71)$$

17.2.3 ADM constraint equations

Unlike the spatial vierbein $e^{a\beta}$, the vierbein $e^{0\mu}$ with a tetrad time index 0, whose components define the lapse and shift, equation (17.11), have vanishing canonically conjugate momenta, as shown in §17.1.3. Consequently, in the ADM formalism, the lapse and shift are not considered to be part of the system of coordinates and momenta that encode the physical gravitational degrees of freedom. Rather, the lapse α and shift β^α are interpreted as gauge variables that can be chosen arbitrarily. The 4 gauge degrees of freedom in the lapse and shift embody the 4 gauge degrees of freedom of coordinate transformations.

Nevertheless, varying the Hilbert action with respect to $e^{0\mu}$ does yield equations of motion, which are the 4 Einstein equations with one tetrad time index 0,

$$G_{m0} = 8\pi T_{m0} . \quad (17.72)$$

Combining equations (17.44a) and (17.45) yields an expression for the time-time Einstein component $G_{00} \equiv R_{00} - \frac{1}{2}\gamma_{00}R = R_{00} + \frac{1}{2}R$, while equation (17.44b) gives the space-time Einstein component $G_{a0} \equiv R_{a0}$,

$$G_{00} = \frac{1}{2}(\hat{R} - K^{ab}K_{ba} + K^2) = \frac{1}{2}(\hat{R} - \pi^{ab}K_{ba}) , \quad (17.73a)$$

$$G_{a0} = \hat{D}^b K_{ba} - \hat{D}_a K = \hat{D}^b \pi_{ba} . \quad (17.73b)$$

Whereas the spatial Einstein equations yielded time evolution equations (17.63) or (17.68) for the momenta, the expressions (17.73) for the time-time and space-time Einstein components involve only spatial derivatives of the coordinates and momenta, no time derivatives. Since the coordinates and momenta are determined fully by their equations of motion, equations (17.52) and (17.63), or (17.55) and (17.68), the Einstein equations (17.72) with at least one time index cannot be independent equations. However, the equations (17.72) cannot be discarded completely. Rather, the Einstein equations (17.72) must be arranged to be satisfied in the initial conditions (on the initial hypersurface of constant time t), whereafter the Bianchi identities ensure that the constraints are satisfied automatically, as you will confirm in Exercise 17.2. This kind of equation, which must be satisfied on the initial hypersurface but is thereafter guaranteed by conservation laws, is called a constraint equation. In the ADM formalism, the time-time Einstein equation is called the **energy constraint** or **Hamiltonian constraint**, while the space-time Einstein equations are called the **momentum constraints**:

$$\frac{1}{2}(\hat{R} - K^{ab}K_{ba} + K^2) = 8\pi T_{00} \quad \text{Hamiltonian constraint} , \quad (17.74a)$$

$$\hat{D}^b K_{ba} - \hat{D}_a K = 8\pi T_{a0} \quad \text{momentum constraints} . \quad (17.74b)$$

Exercise 17.2 Energy and momentum constraints. Confirm the argument of this section. Suppose that the spatial Einstein equations are true, $G^{ab} = 8\pi T^{ab}$. Show that if the time-time and space-time Einstein equations $G^{m0} = 8\pi T^{m0}$ are initially true, then conservation of energy-momentum implies that these equations must necessarily remain true at all times. [Hint: Conservation of energy-momentum requires that $D_n T^{mn} = 0$, and the Bianchi identities require that the Einstein tensor satisfies $D_n G^{mn} = 0$, so

$$D_n (G^{mn} - 8\pi T^{mn}) = 0 . \quad (17.75)$$

By expanding out these equations in full, or otherwise, show that the solution satisfying $G^{ab} - 8\pi T^{ab} = 0$ at all times, and $G^{m0} - 8\pi T^{m0} = 0$ initially, is $G^{m0} - 8\pi T^{m0} = 0$ at all times.]

17.2.4 ADM Raychaudhuri equation

If the Hamiltonian constraint (17.74a) is used to eliminate the restricted Ricci scalar $\hat{R}$, then the trace equation (17.71) becomes

$$\boxed{\mathcal{L}_u K = \frac{1}{\alpha} \left(\frac{\partial K}{\partial t} + \beta^\alpha \frac{\partial K}{\partial x^\alpha} \right) = \hat{D}_\alpha K^\alpha - K^{\alpha\beta} K_{\beta\alpha} + K^\alpha K_\alpha - 4\pi(\rho + 3p)} . \quad (17.76)$$

Equation (17.76) is the Raychaudhuri equation (18.23a) with vanishing vorticity and non-vanishing acceleration K_α .

17.3 Conformally scaled ADM

A common modification of the ADM formalism is to separate out a spatial conformal factor a , which may be an arbitrary function of coordinates.

It is neater to separate the conformal factor from the vierbein than from the tetrad metric, so that the tetrad metric γ_{ab} can still be allowed to be constant, as in the case of an orthonormal tetrad. If the spatial vierbein e^a_α is factored as a product of the conformal factor a and a conformal vierbein $\tilde{e}^a_\alpha$, then the vierbein and inverse vierbein become

$$e^a_\alpha = a \tilde{e}^a_\alpha, \quad e_a^\alpha = \tilde{e}_a^\alpha / a. \quad (17.77)$$

The conformal vierbein and inverse conformal vierbein are inverse to each other, $\tilde{e}^a_\alpha \tilde{e}_b^\alpha = \delta_b^a$. The lapse α and shift β^α are unchanged by the conformal scaling. The spatial conformal coordinate metric defined by $\tilde{g}_{\alpha\beta} \equiv \gamma_{ab} \tilde{e}^a_\alpha \tilde{e}^b_\beta$ is related to the spatial coordinate metric $g_{\alpha\beta}$ by

$$g_{\alpha\beta} = a^2 \tilde{g}_{\alpha\beta}. \quad (17.78)$$

Section 17.1.5 discussed the splitting of tetrad-frame connections into a generalized extrinsic curvature K_{lmn} that behaves like a tensor under some restricted group of transformations, and a restricted connection $\hat{\Gamma}_{lmn}$ that does not transform like a tensor. In the case of ADM, the restricted group of transformations was spatial transformations of the tetrad γ_m (that is, transformations that leave the time axis γ_0 unchanged). The conformal factor a is a scalar with respect to the subgroup of spatial tetrad transformations that leave the conformal factor a unchanged. Thus all of the discussion in §17.1.5 carries through with the restricted group of transformations taken to be spatial transformations that preserve the conformal factor.

The conformal decomposition of the spatial vierbein implies a corresponding conformal decomposition of the vierbein derivatives d_{lmn} defined by equation (11.33). The vierbein derivatives d_{lmn} with either of the first two indices lm the time index 0 are unaffected, but the vierbein derivatives d_{abn} with first two indices ab spatial decompose as

$$d_{abn} \equiv \gamma_{ac} e_b^\alpha \partial_n e^c_\alpha = \gamma_{ac} e_b^\alpha e^c_\alpha \partial_n \ln a + \gamma_{ac} \tilde{e}_b^\alpha \partial_n \tilde{e}^c_\alpha = \gamma_{ab} \partial_n \ln a + \tilde{d}_{abn}, \quad (17.79)$$

which is a sum of a part $\gamma_{ab} \partial_n \ln a$ that depends on derivatives of the conformal factor a , and a conformal part $\tilde{d}_{abn}$ that depends on derivatives of the conformal vierbein $\tilde{e}^c_\alpha$. The part $\gamma_{ab} \partial_n \ln a$ is a spatial tensor under the restricted group of spatial transformations that leave the conformal factor a unchanged. It then follows that the spatial tetrad-frame connections Γ_{abc} split into a restricted part $\hat{\Gamma}_{abc}$ and a tensorial part K_{abc} ,

$$\Gamma_{abc} = \hat{\Gamma}_{abc} + K_{abc}, \quad (17.80)$$

where K_{abc} is the spatial tensor

$$K_{abc} = \gamma_{ac} \partial_b \ln a - \gamma_{bc} \partial_a \ln a . \quad (17.81)$$

The acceleration $K_a \equiv K_{a00}$, extrinsic curvature $K_{ab} \equiv K_{a0b}$, and restricted connections $\hat{\Gamma}_{ab0}$ with final index the time index 0 are unchanged by the conformal decomposition. Thus the generalized extrinsic curvature K_{lmn} now consists of the acceleration K_{a00} , the extrinsic curvature K_{a0b} , and the derivatives K_{abc} of the conformal factor defined by equation (17.81). The generalized extrinsic curvature K_{lmn} remains antisymmetric in its first two indices,

$$K_{lmn} = -K_{mln} . \quad (17.82)$$

The unique non-vanishing contraction K_m of the generalized extrinsic curvature is (this repeats equation (17.23))

$$K_m \equiv K_{mn}^n = \{K_{0n}^n, K_{an}^n\} = \{K_0, K_a\} , \quad (17.83)$$

whose time part remains equal to the trace K of the extrinsic curvature K_{ab} , but whose spatial part K_a is modified to equal the sum of the acceleration K_{a00} and a derivative of the conformal factor,

$$K_a = K_{a00} + 2 \partial_a \ln a = \partial_a \ln(\alpha a^2) . \quad (17.84)$$

Unlike in ADM, K_a is not the same as the acceleration K_{a00} .

The restricted tetrad-frame derivative $\hat{D}_k$ with restricted tetrad-frame connections $\hat{\Gamma}_{lmn}$ is a covariant derivative with respect to the restricted group of spatial transformations that preserve the conformal factor a . The restricted covariant derivative $\hat{D}_k$ differs from ADM only in that the restricted connections now exclude the part depending on derivatives of the conformal factor, which have been absorbed into the spatial components K_{abc} of the generalized extrinsic curvature. The vierbein e^m_μ and the tetrad metric γ_{lm} continue to commute with the restricted covariant derivative D_k , equations (17.30) and (17.31). All of the discussion and equations in §17.1.5 carry through unchanged.

The various expressions for the Riemann and Ricci tensors given in §17.1.6 are modified to include additional terms involving the spatial components K_{abc} of the generalized extrinsic curvature. In particular, the expressions for the Ricci tensor R_{km} are modified to, from the general equation (17.36),

$$R_{00} = -\hat{D}_0 K + \hat{D}_a K_{00}^a - K^{ba} K_{ab} + K_{00}^a K_a , \quad (17.85a)$$

$$R_{a0} = -\hat{D}_a K + \hat{D}^b K_{ba} - K_{ab} K_{00}^b + K_{ba} K^b - K_{cab} K^{bc} , \quad (17.85b)$$

$$= R_{0a} = \hat{R}_{0a} - \hat{D}_0 K_{ab}^b - K_{00}^b K_{ab} + K_{a00} K - K^{bc} K_{cab} , \quad (17.85c)$$

$$R_{ab} = \hat{R}_{ab} - \hat{D}_a K_b + \hat{D}_0 K_{ba} + \hat{D}^c K_{cba} + K_{ba} K - K_{a00} K_{b00} + K_{cba} K^c - K_{ad}^c K_{bc}^d . \quad (17.85d)$$

Like the time-space restricted Ricci tensor $\hat{R}_{0a}$, the spatial restricted Ricci tensor $\hat{R}_{ab}$ is not symmetric in ab .

The equations of motion (17.51) or (17.55) for the spatial metric $g_{\alpha\beta}$ remain unchanged by the conformal decomposition. The equation of motion (17.63) for the extrinsic curvature K_{ab} is modified in accordance

with the modified expression (17.85d) for the spatial Ricci tensor R_{ab} to

$$\hat{D}_0 K_{ab} = \hat{D}_a K_b - \hat{D}^c K_{cba} - K_{ab} K + K_{a00} K_{b00} - K_{cba} K^c + K_{ad}^c K_{bc}^d - \hat{R}_{ab} + 8\pi (T_{ab} - \tfrac{1}{2} \gamma_{ab} T) . \quad (17.86)$$

Equation (17.86) is essentially the same as the earlier equation of motion (17.63), but it redistributes terms involving derivatives of the conformal factor a out of the spatial restricted Ricci tensor $\hat{R}_{ab}$ into terms involving K_a and K_{abc} . The coordinate-frame version of the equation of motion (17.68) for the extrinsic curvature $K_{\alpha\beta}$ is modified similarly to

$$\begin{aligned} \mathcal{L}_u K_{\alpha\beta} &= \frac{1}{\alpha} \left(\frac{\partial K_{\alpha\beta}}{\partial t} + \hat{\mathcal{L}}_\beta K_{\alpha\beta} \right) \\ &= \hat{D}_\alpha K_\beta - \hat{D}^\gamma K_{\gamma\beta\alpha} + 2K_\alpha^\gamma K_{\gamma\beta} - K_{\alpha\beta} K + K_{\alpha 00} K_{\beta 00} - K_{\gamma\beta\alpha} K^\gamma + K_{\alpha\delta}^\gamma K_{\beta\gamma}^\delta - \hat{R}_{\alpha\beta} + 8\pi (T_{\alpha\beta} - \tfrac{1}{2} g_{\alpha\beta} T) . \end{aligned} \quad (17.87)$$

Again, this equation of motion is essentially the same as the earlier equation of motion (17.68), with a redistribution of terms out of $\hat{R}_{\alpha\beta}$ into generalized extrinsic curvatures.

17.4 Bianchi spacetimes

A Bianchi space (Bianchi, 1898) is a 3-dimensional manifold whose directed derivatives ∂_a form a Lie algebra, meaning that the structure coefficients c_{ab}^c defined by the commutators of the directed derivatives, equation (17.92), are spatially constant. A Bianchi space is consequently homogeneous. The assumption that the structure coefficients are constant is stronger than the assumption that the space is homogeneous, which requires merely that the tetrad-frame Riemann tensor be spatially constant.

Because a Bianchi space is homogeneous, it possesses a complete set of Killing fields ξ_a that generate those homogeneities. The Killing fields ξ_a are not the same as the directed derivatives ∂_a . Rather, as shown in §17.4.1, the Killing fields commute with the directed derivatives, equation (17.93). The commutators of the Killing fields themselves prove to have structure coefficients that are minus those of the directed derivatives, equation (17.94). The Killing fields generate a Lie group of symmetries of the 3-dimensional manifold. There is a discrete symmetry (17.117) that exchanges the Killing fields and directed derivatives.

Bianchi spaces are homogeneous but not necessarily isotropic. However, FLRW spacetimes, which are by definition both spatially homogeneous and isotropic, can be recast as Bianchi spaces, §17.5. For example, the closed spatial FLRW geometry, which has the geometry of a 3-sphere, coincides with the Bianchi Type IX geometry in the case of all equal axes, Exercise 17.4. Thus, most homogeneous 3-dimensional spaces are Bianchi spaces, the notable exception being the closed cylindrical geometry, equation (17.160).

Bianchi spacetimes, also known as Bianchi universes, are Bianchi spaces that evolve in time while preserving the homogeneous spatial Bianchi structure. Bianchi spacetimes offer a framework for addressing possible large scale departures from isotropy in cosmology, and provide the prototype for the Belinskii-Khalatnikov-Lifshitz (BKL) (Belinskii, Khalatnikov, and Lifshitz, 1982; Belinski, 2014) model of anisotropic gravitational collapse, §17.6. Bianchi spacetimes present a fine application of both the ADM formalism and the tetrad formalism, in a situation where the tetrad is neither orthonormal, nor aligned with the coordinates, nor is the tetrad metric constant (in time).

17.4.1 Bianchi structure coefficients

A symmetry of a manifold is described by a Killing field, §7.32. A Bianchi space has 3 independent Killing fields ξ_a , whose coefficients define Killing vectors ξ_a^α ,

$$\xi_a \equiv \xi_a^\alpha \frac{\partial}{\partial x^\alpha} . \quad (17.88)$$

The Killing fields define a triad of axes ξ_a at each point on the manifold, related to the spatial tangent vectors e_α by

$$\xi_a \equiv \xi_a^\alpha e_\alpha . \quad (17.89)$$

Because the space is homogeneous, it is possible to introduce a triad of axes γ_a such that the triad metric $\gamma_{ab} \equiv \gamma_a \cdot \gamma_b$ is spatially constant. It is possible to choose an origin of the manifold and to co-align the triad axes γ_a with the Killing fields ξ_a at the origin. However, the scalar product $\xi_a \cdot \xi_b$ of Killing vectors is generically not constant if the manifold is rotated slightly away from the origin, so the Killing axes ξ_a cease to remain aligned with the triad axes γ_a away from the origin. Therefore the symmetry generators ξ_a must be distinguished from the spatial directed derivatives ∂_a , which are defined as usual by, equation (11.30),

$$\partial_a \equiv e_a^\alpha \frac{\partial}{\partial x^\alpha} , \quad (17.90)$$

with e_a^α the spatial inverse dreibein. The spatial triad γ_a axes are related to the coordinate tangent vectors e_α as usual by

$$\gamma_a \equiv e_a^\alpha e_\alpha . \quad (17.91)$$

The coordinates of the manifold can be chosen so that the Killing fields ξ_a and triad axes γ_a co-align with the coordinate axes e_α at the origin, so that the Killing vectors ξ_a^α and dreibein e_a^α are the unit matrix at the origin. The Killing fields ξ_a and triad γ_a are independent of the choice of spatial coordinates x^α . However, convenient coordinates can be constructed by exponentiating the generators ∂_a of displacements from the origin. Exercise 17.4 gives an example for the case of Bianchi Type IX.

The commutators of the directed derivatives ∂_a define the structure coefficients c_{ab}^c ,

$$[\partial_a, \partial_b] = c_{ab}^c \partial_c . \quad (17.92)$$

As shown below, equation (17.97), the Killing fields ξ_a commute with the directed derivatives ∂_a ,

$$[\xi_a, \partial_b] = 0 . \quad (17.93)$$

It then follows, equation (17.107), that the commutators of the Killing fields ξ_a have the same structure coefficients as the directed derivatives ∂_a , but flipped in sign,

$$[\xi_a, \xi_b] = -c_{ab}^c \xi_c . \quad (17.94)$$

The structure coefficients c_{ab}^c are necessarily antisymmetric in their last two indices ab . The structure coefficients must be such that Jacobi identity $[\partial_a, [\partial_b, \partial_c]] = 0$ is satisfied. For spatially constant structure

coefficients, the Jacobi identity requires

$$c_{d[a}^e c_{bc]}^d = 0 . \quad (17.95)$$

As discussed in §7.34, a Killing vector ξ satisfies Killing's equation (7.123), which says that the Lie derivative $\mathcal{L}_\xi$ of the metric along the direction ξ vanishes. Equivalently, the Lie derivative of the inverse metric vanishes, equation (7.162). With the tetrad metric γ_{ab} spatially constant, Killing's equation becomes a condition on the (inverse) dreibein e_a^α ,

$$\mathcal{L}_\xi g^{\alpha\beta} = \gamma^{ab} (e_b^\beta \mathcal{L}_\xi e_a^\alpha + e_a^\alpha \mathcal{L}_\xi e_b^\beta) = 0 . \quad (17.96)$$

A sufficient condition for Killing's equation (17.96) to hold is that the Lie derivative of the inverse dreibein vanishes,

$$\mathcal{L}_\xi e_b^\beta = \xi_a^\alpha \frac{\partial e_b^\beta}{\partial x^\alpha} - e_b^\alpha \frac{\partial \xi_a^\beta}{\partial x^\alpha} = 0 . \quad (17.97)$$

Not all symmetries satisfy the condition (17.97), but the condition is necessary and sufficient for the structure coefficients to be spatially constant, equation (17.99), the defining property of a Bianchi space. The condition (17.97) is equivalent to the vanishing of the commutators of the symmetry generators ξ_a with the directed derivatives ∂_b , as asserted earlier, equation (17.93). The Jacobi identity

$$\begin{aligned} 0 &= [\partial_a, [\partial_b, \xi_c]] + [\partial_b, [\xi_c, \partial_a]] + [\xi_c, [\partial_a, \partial_b]] \\ &= [\xi_c, [\partial_a, \partial_b]] \\ &= \xi_c^\gamma (\partial_\gamma c_{ab}^d) \partial_d + c_{ab}^d [\xi_c, \partial_d] \\ &= \xi_c^\gamma (\partial_\gamma c_{ab}^d) \partial_d \end{aligned} \quad (17.98)$$

then implies that the structure constants c_{ab}^d are all constant, independent of the coordinates

$$c_{ab}^d = \text{constant} . \quad (17.99)$$

The Killing vectors ξ_a^α can be expressed as linear combinations of the dreibein e_a^α ,

$$\xi_a^\alpha = \Xi_a^b e_b^\alpha , \quad (17.100)$$

which defines the 3×3 matrix Ξ_a^b . Equivalently, the Killing axes ξ_a are related to the triad γ_a by

$$\xi_a = \Xi_a^b \gamma_b . \quad (17.101)$$

The vanishing commutator (17.93) implies an expression for the commutator $[\partial_c, \partial_b]$,

$$0 = [\xi_a, \partial_b] = [\Xi_a^c \partial_c, \partial_b] = \Xi_a^c [\partial_c, \partial_b] - (\partial_b \Xi_a^c) \partial_c = \Xi_a^c (c_{cb}^d - (\Xi^{-1})_c^e (\partial_b \Xi_e^d)) \partial_d , \quad (17.102)$$

which implies, in abbreviated notation (with c_b denoting the matrix c_{cb}^d), that the structure coefficients of the commutator $[\partial_c, \partial_b]$ are

$$c_b = \Xi^{-1} \partial_b \Xi . \quad (17.103)$$

Since the structure coefficients are spatially constant, the expression (17.103) can be evaluated anywhere, in particular at the origin where Ξ equals (by construction) the unit matrix, and equation (17.103) reduces to

$$c_b = \partial_b \ln \Xi|_{\text{origin}} . \quad (17.104)$$

Alternatively, the vanishing commutator (17.93) implies an expression for the commutator $[\xi_b, \xi_c]$,

$$0 = [\xi_b, \partial_a] = [\xi_b, (\Xi^{-1})^c_a \xi_c] = (\Xi^{-1})^c_a [\xi_b, \xi_c] + (\xi_b (\Xi^{-1})^c_a) \xi_c = (\Xi^{-1})^c_a \left(-c_{bc}^d + \Xi_c^e (\xi_b (\Xi^{-1})^e_d) \right) \xi_d , \quad (17.105)$$

which implies, in abbreviated notation, (again with c_b denoting $c_{cb}^d = -c_{bc}^d$), that the structure coefficients of the commutator $[\xi_b, \xi_c]$ are

$$-c_b = \Xi \xi_b \Xi^{-1} . \quad (17.106)$$

Again, the constancy of the structure coefficients allows the expression (17.106) to be evaluated anywhere, in particular at the origin where Ξ is the unit matrix and $\xi_b = \partial_b$, where equation (17.106) reduces to

$$-c_b = -\partial_b \ln \Xi|_{\text{origin}} . \quad (17.107)$$

Equation (17.104) is an expression for the structure coefficients of the directed derivatives ∂_a , while equation (17.107) is an expression for the structure coefficients of the Killing vectors ξ_a . Equations (17.104) and (17.107) establish that the structure coefficients of the Killing vectors are indeed minus those of the directed derivatives, as earlier asserted in equation (17.94).

In general relativity, the structure coefficients c_{ab}^c defined by the commutators of the directed derivatives do not usually constitute a tensor. But in Bianchi spaces, the spatial structure coefficients $d\sigma$ constitute a spatial tensor under global transformations of the spatial tetrad γ_a , global meaning that the transformation is the same at every spatial point. Such transformations must of course respect the symmetries of the system. Generically, the only transformations of γ_a that respect homogeneity are scale transformations of each γ_a ; but other transformations are available if the space happens to possess additional rotational symmetries, as for example FLRW spaces. In Bianchi spaces, quantities that depend on the spatial structure coefficients, notably the spatial tetrad connections Γ_{abc} , equation (17.126), are tensors under global spatial transformations.

17.4.2 Bianchi line-element

A Bianchi spacetime is a Bianchi space that evolves in time while preserving the homogeneous Bianchi spatial structure. The time coordinate t is chosen so that spatial surfaces of constant time are homogeneous. To preserve spatial homogeneity, the tetrad metric $\gamma_{ab} \equiv \gamma_a \cdot \gamma_b$ must be independent of the spatial coordinates, but it may depend on time t . As usual in the ADM formalism, the tetrad time axis γ_0 is chosen to be orthogonal to the spatial tetrad axes γ_a , which lie in the surfaces of constant time. The line-element can thus be taken to be

$$ds^2 = -dt^2 + g_{\alpha\beta} dx^\alpha dx^\beta = -dt^2 + \gamma_{ab}(t) e^a_\alpha e^b_\beta dx^\alpha dx^\beta , \quad (17.108)$$

which is in ADM form with unit lapse and zero shift. The spatial metric $g_{\alpha\beta} = \gamma_{ab} e^a{}_\alpha e^b{}_\beta$ depends on spatial coordinates x^α only through the dreibein $e^a{}_\alpha$. The vierbein and its inverse are

$$e^m{}_\mu = \begin{pmatrix} 1 & 0 \\ 0 & e^a{}_\alpha \end{pmatrix}, \quad e_m{}^\mu = \begin{pmatrix} 1 & 0 \\ 0 & e_a{}^\alpha \end{pmatrix}. \quad (17.109)$$

The tetrad time derivative coincides with the coordinate time derivative, $\partial_0 = \partial/\partial t$. The condition that the homogeneous spatial structure be preserved in time means that the spatial directed derivatives commute with the time derivative, $[\partial_0, \partial_a] = 0$, so the vierbein, and the inverse vierbein, are independent of time.

17.4.3 Classification of Bianchi spaces

Bianchi spaces are classified according to the invariant properties of their constant structure coefficients c_{ab}^c . Choose a point of the spacetime. The structure coefficients at that point can be written in terms of a 3×3 matrix n^{dc} , which can be decomposed into symmetric $n^{(dc)}$ and antisymmetric $n^{[dc]}$ parts,

$$c_{ab}^c = \varepsilon_{abd} n^{dc} = \varepsilon_{abd} (n^{(dc)} + n^{[dc]}). \quad (17.110)$$

By a (global) orthogonal rotation of axes the symmetric matrix $n^{(dc)}$ can be brought to diagonal form with eigenvalues n_c , while the antisymmetric part can be written in terms of a vector k_e ,

$$c_{ab}^c = \varepsilon_{abd} (\delta^{dc} n_c + \varepsilon^{dce} k_e) \quad (\text{no sum over } c). \quad (17.111)$$

The Jacobi identity (17.95) implies that $0 = \varepsilon^{abc} c_{da}^e c_{bc}^d = 2\varepsilon_{daf} n^{fe} n^{ad} = -4n^{fe} k_f$, which equals $-4n_e k_e$ (no sum over e) in each direction e , thus

$$n^{fe} k_f = n_e k_e = 0 \quad (\text{each direction } e, \text{ no sum over } e). \quad (17.112)$$

Thus in each direction, either the eigenvalue n_e or the component k_e in the same direction equals zero. If the vector k_e is non-vanishing, then without loss of generality it can be chosen to lie along the 1-direction, $k_e = \{k, 0, 0\}$. The real number k can be non-zero only if $n_1 = 0$. The commutators (17.94) of the directed derivatives ∂_a then reduce to (with at least one of n_1 and k zero)

$$[\partial_3, \partial_2] = n_1 \partial_1, \quad [\partial_1, \partial_3] = n_2 \partial_2 + k \partial_3, \quad [\partial_2, \partial_1] = n_3 \partial_3 - k \partial_2. \quad (17.113)$$

Under a rescaling of axes $\partial_c \propto 1/a_c$, the eigenvalues scale as $n_1 \propto a_1/(a_2 a_3)$ and cyclically for n_2 and n_3 . Thus by a rescaling of axes, each of the non-zero eigenvalues n_c can be scaled to any other value of the same sign. Flipping any axis changes the signs of all the n_c , so the number of positive eigenvalues can always be chosen to be greater than or equal to the number of negative eigenvalues. Finally, the axes can be reordered arbitrarily. Thus the invariant properties of the eigenvalues n_c are the numbers of negative, zero, and positive eigenvalues. If the parameter k is non-zero, and if n_2 and n_3 are non-zero (Bianchi Types VI and VII), then k cannot be rescaled independently, since $k \propto 1/a_1 \propto |n_2 n_3|^{1/2}$ is fixed by the scaling of n_2 and n_3 . If on the other hand either of n_2 and n_3 are non-zero (Bianchi Types V and IV), then k can be rescaled independently. The sign of k changes under a flip of the 1-axis, so k can be taken to be positive.

Table 17.1 lists the distinct possibilities for the 3 eigenvalues n_e , and gives the corresponding traditional

Table 17.1: Classification of Bianchi spaces

Eigenvalues			Type	
n_1	n_2	n_3	$k = 0$	$k \neq 0$
0	0	0	I	V
0	0	+	II	IV
0	−	+	VI ₀	VI
0	+	+	VII ₀	VII
−	+	+	VIII	
+	+	+	IX	

Bianchi type. Missing from the Table is Type III, which is a special case of Type VI with $k = 1$, if n_2 and n_3 are scaled to ∓ 1 . Type III is distinguished by the fact that all three eigenvalues of the matrix n^{dc} (the full matrix, including both symmetric and antisymmetric parts) degenerate to zero.

In summary, the matrix n^{dc} is, in the frame where its symmetric part is diagonal,

$$n^{dc} = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & k \\ 0 & -k & n_3 \end{pmatrix}, \quad (17.114)$$

in which k can be non-zero only if $n_1 = 0$.

Table 17.2 gives spatial dreibein $e^a{}_\alpha$ and inverse dreibein $e_a{}^\alpha$ in a frame where the symmetric part of the matrix n^{dc} is diagonal, equation (17.114). The Table first lists Bianchi spaces with $k = 0$, then in the last line the Table gives the spatial dreibein and inverse dreibein for Bianchi spaces with $k \neq 0$, which are related to those with $k = 0$ by

$$\text{dreibein}(k \neq 0) = \text{dreibein}(k = 0) \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}. \quad (17.115)$$

For example, the dreibein for Type VII is

$$\text{Type VII: } e^a{}_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} \cos x & e^{kx} \sin x \\ 0 & -e^{kx} \sin x & e^{kx} \cos x \end{pmatrix}. \quad (17.116)$$

Table 17.3 gives the corresponding Killing vectors $\xi_a{}^\alpha$ and their inverses $\xi^a{}_\alpha$ (note that $\xi^a{}_\alpha$ is defined to be the (matrix) inverse of $\xi_a{}^\alpha$, which is *not* the same as the Killing vector with tetrad and coordinate indices raised and lowered with the tetrad metric and spatial metric.) In the row for Types I–VII, the values of n_c are 0 or ± 1 , as listed in Table 17.1. In view of the complementarity of the structure coefficients of ξ_a and ∂_a , equations (17.94) and (17.92), there is a discrete symmetry in which the inverse dreibein can be exchanged

Table 17.2: Bianchi spatial dreibein $e^a{}_\alpha$ and their inverses $e_a{}^\alpha$

Type	dreibein $e^a{}_\alpha$	inverse dreibein $e_a{}^\alpha$
I	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
II	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -x & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & x \\ 0 & 0 & 1 \end{pmatrix}$
VI ₀	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cosh x & -\sinh x \\ 0 & -\sinh x & \cosh x \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cosh x & \sinh x \\ 0 & \sinh x & \cosh x \end{pmatrix}$
VII ₀	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & \sin x \\ 0 & -\sin x & \cos x \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & \sin x \\ 0 & -\sin x & \cos x \end{pmatrix}$
VIII	$\begin{pmatrix} 1 & 0 & \sinh y \\ 0 & \cos x & \sin x \cosh y \\ 0 & -\sin x & \cos x \cosh y \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ -\sin x \tanh y & \cos x & \sin x \operatorname{sech} y \\ -\cos x \tanh y & -\sin x & \cos x \operatorname{sech} y \end{pmatrix}$
IX	$\begin{pmatrix} 1 & 0 & \sin y \\ 0 & \cos x & \sin x \cos y \\ 0 & -\sin x & \cos x \cos y \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ \sin x \tan y & \cos x & \sin x \sec y \\ \cos x \tan y & -\sin x & \cos x \sec y \end{pmatrix}$
$k \neq 0$	$e^a{}_\alpha(k=0) \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$	$e_a{}^\alpha(k=0) \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{-kx} & 0 \\ 0 & 0 & e^{-kx} \end{pmatrix}$

with the Killing vectors, with a flip of the sign of the structure coefficients c_{ab}^c , or equivalently of the sign of the structure matrix n^{dc} ,

$$e_a{}^\alpha \rightarrow \xi_a{}^\alpha (n^{dc} \rightarrow -n^{dc}) , \quad \xi_a{}^\alpha \rightarrow e_a{}^\alpha (n^{dc} \rightarrow -n^{dc}) . \quad (17.117)$$

The flip of the sign of the structure coefficients can equivalently be accomplished by flipping the signs of all coordinates, $x^\alpha \rightarrow -x^\alpha$.

Table 17.4 gives the transformation matrix $\Xi_a{}^b$ that transforms the inverse dreibein into the Killing vectors, equation (17.100). Cases I, II, VI₀, and VII₀ are subcases of V, IV, VI, and VII with $k = 0$. The transformation matrix can be written in terms of the Killing vectors and dreibein

$$\Xi_a{}^b = \xi_a{}^\alpha e_\alpha{}^b . \quad (17.118)$$

Table 17.3: Bianchi Killing vectors ξ_a^α and their inverses ξ^a_α

Type	Killing vectors ξ_a^α	inverse Killing vectors ξ^a_α
I–VII	$\begin{pmatrix} 1 & -n_2 z - ky & n_3 y - kz \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ n_2 z + ky & 1 & 0 \\ -n_3 y + kz & 0 & 1 \end{pmatrix}$
VIII	$\begin{pmatrix} \operatorname{sech} y \cosh z & -\sinh z & \tanh y \cosh z \\ -\operatorname{sech} y \sinh z & \cosh z & -\tanh y \sinh z \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} \cosh y \cosh z & \sinh z & 0 \\ \cosh y \sinh z & \cosh z & 0 \\ -\sinh y & 0 & 1 \end{pmatrix}$
IX	$\begin{pmatrix} \sec y \cos z & -\sin z & \tan y \cos z \\ \sec y \sin z & \cos z & \tan y \sin z \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} \cos y \cos z & -\sin z & 0 \\ \cos y \sin z & \cos z & 0 \\ -\sin y & 0 & 1 \end{pmatrix}$

The matrices Ξ in Table 17.4 are written as products of matrices in order to make their structure more manifest. For Type IX, the structure matrix Ξ is orthogonal.

Exercise 17.3 Geodesics in Bianchi spacetimes. Solve for the geodesics of particles in a Bianchi spacetime. Is the comoving frame geodesic? Are there any other cases where geodesic motion is particularly simple?

Solution. With each symmetry ξ_a is associated a constant of motion C_a for a particle following a geodesic,

$$C_a = \xi_a^\alpha p_\alpha = \Xi_a^b p_b, \quad (17.119)$$

where $p_\alpha \equiv g_{\alpha\beta} p^\beta$ are the covariant coordinate-frame spatial components of the 4-momentum $p^\mu \equiv dx^\mu/d\lambda$ of the particle, and $p_b \equiv e_b^\alpha p_\alpha$ are its covariant tetrad-frame components. Equations (17.119) invert to yield the covariant coordinate-frame and tetrad-frame spatial momentum p_α and p_a of the particle in terms of the constants C_a of motion,

$$p_\alpha = \xi^a_\alpha C_a, \quad p_a = (\Xi^{-1})_a^b C_b, \quad (17.120)$$

where ξ^a_α is the (matrix) inverse of ξ_a^α . The contravariant spatial components of the momentum are $p^a = \gamma^{ab} p_b = e^a_\alpha p^\alpha$ and $p^\alpha = g^{\alpha\beta} p_\beta = e_a^\alpha p_a$. The time component $p^0 = p^t$ of the momentum can be deduced from conservation of rest mass, $p^m p_m = p^\mu p_\mu = -m^2$.

Yes, the comoving frame is geodesic. Zero covariant momentum in the comoving frame, $p_\alpha = 0$, is a solution of the geodesic equation, so zero contravariant momentum, $p^\alpha = 0$, is a solution.

Are there special geodesic directions? The tetrad-frame covariant momentum p_a is constant along all directions for which the eigenvalue n_a , Table 17.1, vanishes.

Exercise 17.4 Bianchi Type IX space with all equal axes is a 3-sphere. The Bianchi Type IX spatial geometry coincides with that of a 3-sphere in the special case where all three axes are equal. Find

Table 17.4: Bianchi transformation matrix Ξ

Type	Transformation matrix Ξ_a^b
V	$\begin{pmatrix} 1 & -ky & -kz \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$
IV	$\begin{pmatrix} 1 & -ky & y - kz \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -x \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$
VI	$\begin{pmatrix} 1 & z - ky & y - kz \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cosh x & -\sinh x \\ 0 & -\sinh x & \cosh x \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$
VII	$\begin{pmatrix} 1 & -z - ky & y - kz \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & -\sin x \\ 0 & \sin x & \cos x \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$
VIII	$\begin{pmatrix} \cosh z & -\sinh z & 0 \\ -\sinh z & \cosh z & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cosh y & 0 & \sinh y \\ 0 & 1 & 0 \\ \sinh y & 0 & \cosh y \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & -\sin x \\ 0 & \sin x & \cos x \end{pmatrix}$
IX	$\begin{pmatrix} \cos z & -\sin z & 0 \\ \sin z & \cos z & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos y & 0 & \sin y \\ 0 & 1 & 0 \\ -\sin y & 0 & \cos y \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & -\sin x \\ 0 & \sin x & \cos x \end{pmatrix}$

the relation between Bianchi IX coordinates x, y, z and the Cartesian coordinates X^m , $m = 0, 1, 2, 3$, of the 4-dimensional Euclidean space in which the 3-sphere can be embedded.

Solution. A 3-sphere transforms to itself under the orthogonal group $\text{SO}(4)$ of rotations of 4-dimensional Euclidean space. The covering group (double cover) of $\text{SO}(4)$ is $\text{Spin}(4)$. Exercise 39.2 shows that $\text{Spin}(4)$ is isomorphic to $\text{SU}(2)_L \times \text{SU}(2)_L$. The generators of $\text{Spin}(4)$ are its 6 bivectors, which split into two commuting sets of 3 bivectors of respectively right and left chirality. The 2 commuting sets of 3 bivector generators of $\text{Spin}(4)$ are

$$\frac{1}{2}(1 \pm I)\gamma_{a0}, \quad a = 1, 2, 3, \quad (17.121)$$

where $I \equiv \gamma_{1230}$ is the 4-dimensional Euclidean pseudoscalar, and $\frac{1}{2}(1 \pm I)$ are the right-handed (+) and left-handed (−) chiral projection operators. The 3 Killing vectors ξ_a and the 3 vectors ∂_a have the correct commutation relations for two sets of commuting bivectors of $\text{SU}(2)_L$, equations (17.94), (17.92), and (17.93), and can be identified with the two sets of bivector generators of $\text{Spin}(4)$. The Pauli algebra of $\text{SU}(2)_L$ is conventionally normalized to $[i\sigma_a, i\sigma_b] = -2\varepsilon_{abc}i\sigma_c$, so the eigenvalue n in the Bianchi IX geometry,

Table 17.1, must be set to $n = 2$ to normalize the commutation rules of ξ_a and ∂_a to the Pauli convention, in place of the conventional Bianchi choice $n = 1$. Without loss of generality, the 3 Killing vectors ξ_a can be taken to be right-handed (consistent with the $-$ sign in equation (17.94)), while the 3 vectors ∂_a can be taken to be left-handed. Table 17.4 shows that the matrix $\Xi_a{}^b$ that transforms from ∂_b to ξ_a , equation (17.100), is an orthogonal matrix that is a product of three successive rotations by angles $2z$ in the 12-plane, $2y$ in the 31-plane, and $2x$ in the 23-plane, the factor of 2 in each angle coming from the adjusted Bianchi eigenvalue $n = 2$ compared to $n = 1$ in Table 17.4. The generator of each of the 3 rotations is a sum of right- and left-handed generators, and each of the 3 rotations is itself a product of (commuting) right- and left-handed rotations. The right-handed part of the rotation is generated by ξ_a , while the left-handed part is generated by ∂_a . The Cartesian 4-vector X^m is obtained by a left-handed transformation starting from the “north pole” $r\gamma_0$, with r the (arbitrary) radius of the 3-sphere,

$$X^m \gamma_m = r L_x L_y L_z \gamma_0 \tilde{L}_z \tilde{L}_y \tilde{L}_x , \quad (17.122)$$

where L_a are left-handed rotors

$$L_x = \exp\left(-\frac{1}{2}(1+I)\gamma_{10}x\right) , \quad L_y = \exp\left(-\frac{1}{2}(1+I)\gamma_{20}y\right) , \quad L_z = \exp\left(-\frac{1}{2}(1+I)\gamma_{30}z\right) . \quad (17.123)$$

Note that a factor of $\frac{1}{2}$ in the standard definition (13.50) of a rotor cancels against the factor of $n = 2$. Equation (17.122) yields the components X^m of the Euclidean 4-vector,

$$X^m = r \begin{pmatrix} \cos x \cos y \cos z - \sin x \sin y \sin z \\ \sin x \cos y \cos z + \cos x \sin y \sin z \\ \cos x \sin y \cos z - \sin x \cos y \sin z \\ \cos x \cos y \sin z + \sin x \sin y \cos z \end{pmatrix} . \quad (17.124)$$

The Euclidean line-element recast in terms of the radius r and the Bianchi IX coordinates x, y, z is

$$\sum_{m=0}^3 (dX^m)^2 = dr^2 + r^2(dx^2 + dy^2 + dz^2 - 2 \sin 2y dx dz) . \quad (17.125)$$

The angular part of the line-element (17.125) agrees with that of the Bianchi IX line-element with all equal axes, the factor of 2 in $\sin 2y$ coming from the eigenvalue $n = 2$.

17.4.4 Bianchi connections and curvatures

The formulae in this section are valid for homogeneous spacetimes regardless of whether the structure constants c_{ab}^c are spatially constant.

The non-vanishing tetrad-frame connections are, from equation (11.53),

$$\Gamma_{ab0} = \Gamma_{a0b} = -\Gamma_{0ab} = \frac{1}{2}\dot{\gamma}_{ab} , \quad \Gamma_{abc} = \frac{1}{2}(c_{cab} + c_{bac} - c_{abc}) , \quad (17.126)$$

where the overdot represents the time derivative, $\dot{\gamma}_{ab} \equiv d\gamma_{ab}/dt$ (an ordinary derivative because γ_{ab} varies only in time, not space), and $c_{cab} \equiv \gamma_{cd}c_{ab}^d$. The connections with one time 0 index are symmetric in their

spatial indices ab , while the purely spatial connections Γ_{abc} are antisymmetric in their first two indices ab . The tetrad frame is locally inertial (freely falling and non-rotating), as follows from the fact that the acceleration and precession both vanish, $\Gamma_{a00} = \Gamma_{[ab]0} = 0$. The vorticity, the antisymmetric part of Γ_{a0b} , vanishes. Altogether there are $6 + 9 = 15$ distinct non-vanishing connections. If the structure coefficients c_{ab}^c are spatially constant, then so are the spatial connections Γ_{abc} , but more generally the spatial connections can vary in space, as they do for example in the FLRW geometry discussed in Chapter 10 (although FLRW spacetimes can be realised as Bianchi spacetimes with constant structure coefficients — see §17.5). The spatial connections Γ_{abc} also vary in time because, whereas c_{ab}^c with one index raised is constant in time, the coefficients $c_{cab} \equiv \gamma_{cd}c_{ab}^d$ with all indices lowered depend on time through the time-dependent metric γ_{cd} . Explicitly,

$$\Gamma_{abc} = \frac{1}{2} (\gamma_{cd}c_{ab}^d + \gamma_{bd}c_{ac}^d - \gamma_{ad}c_{bc}^d) . \quad (17.127)$$

The unique non-vanishing contraction of the spatial connections Γ_{abc} is

$$\Gamma_{ba}^a = c_{ab}^a = \varepsilon_{abc}n^{ca} = 2k_b , \quad (17.128)$$

which is constant in time.

The extrinsic curvature K_{ab} , equation (17.21b), is by definition

$$K_{ab} \equiv \Gamma_{a0b} = \frac{1}{2}\dot{\gamma}_{ab} , \quad (17.129)$$

with trace

$$K \equiv K_a^a = \frac{1}{2}\gamma^{ab}\dot{\gamma}_{ab} = \frac{d \ln \sqrt{\gamma}}{dt} , \quad (17.130)$$

where $\gamma \equiv |\gamma_{ab}|$ is the determinant of the spatial tetrad metric. The last step of equations (17.130) is an application of equation (2.77). The proper spatial volume element is $d^3x = \sqrt{\gamma}d^3x^{123}$, so the trace K measures the logarithmic rate of change of a comoving volume element. Positive K means that the comoving volume element is expanding, while negative K means that the comoving volume element is contracting.

The tetrad-frame Riemann curvature tensor R_{klmn} , which homogeneity requires to be spatially constant regardless of whether the structure coefficients are spatially constant, is, from equations (17.40),

$$R_{a0b0} = -\dot{K}_{ab} + K_a^c K_{bc} , \quad (17.131a)$$

$$R_{abc0} = \Gamma_{cb}^d K_{da} - \Gamma_{ca}^d K_{db} + (\Gamma_{ab}^d - \Gamma_{ba}^d) K_{cd} , \quad (17.131b)$$

$$R_{abcd} = \hat{R}_{abcd} - K_{cb}K_{da} + K_{ca}K_{db} , \quad (17.131c)$$

where $\hat{R}_{abcd}$ is the restricted Riemann tensor,

$$\hat{R}_{abcd} = \partial_a \Gamma_{cdb} - \partial_b \Gamma_{cda} + \Gamma_{cb}^e \Gamma_{eda} - \Gamma_{ca}^e \Gamma_{edb} + (\Gamma_{ab}^e - \Gamma_{ba}^e) \Gamma_{cde} . \quad (17.132)$$

If the structure coefficients are spatially constant, then the two derivative terms on the right hand side of equation (17.132) can be dropped. For spatially constant structure coefficients, equations (17.131) and (17.132) along with equations (17.126) give the Riemann tensor in terms of the structure coefficients c_{ab}^c and the tetrad metric γ_{ab} , without the need for an explicit form for the spatial vierbein e^α_α . If the structure

coefficients were derived from an explicit vierbein, then the usual symmetries of the Riemann tensor (with vanishing torsion) would be guaranteed. But the symmetries are ensured in any case, since for constant structure coefficients the Jacobi identity (17.112) implies that the restricted Riemann tensor satisfies the cyclic symmetry $\varepsilon^{bcd}\hat{R}_{abcd} = -4\gamma_{ab}n^{bc}k_c = 0$, which in turn ensures that the restricted Riemann tensor $\hat{R}_{abcd}$ is symmetric in $ab \leftrightarrow cd$, Exercise 11.7.

Contracting the Riemann tensor yields the Ricci tensor R_{km} ,

$$R_{00} = -\dot{K} - K^{ab}K_{ab} , \quad (17.133a)$$

$$R_{a0} = \Gamma_{cb}^b K_a^c - \Gamma_{ab}^c K_c^b , \quad (17.133b)$$

$$R_{ab} = \hat{R}_{ab} + \dot{K}_{ab} - 2K_a^c K_{bc} + K_{ab}K , \quad (17.133c)$$

where $\hat{R}_{ab}$ is the restricted Ricci tensor,

$$\hat{R}_{ab} = -\partial_a \Gamma_{bc}^c + \partial_c \Gamma_{ba}^c + \Gamma_{ba}^e \Gamma_{ed}^d - \Gamma_{ad}^e \Gamma_{be}^d . \quad (17.134)$$

Again, if the structure coefficients are spatially constant, then the two derivative terms on the right hand side of equation (17.134) can be dropped. And again, for spatially constant structure coefficients, the Jacobi identity (17.112) ensures that the restricted Ricci tensor is symmetric, $\hat{R}_{[ab]} = 2\varepsilon_{abd}n^{dc}k_c = 0$. Contracting the Ricci tensor yields the Ricci scalar R ,

$$R = \hat{R} + 2\dot{K} + K^{ab}K_{ab} + K^2 , \quad (17.135)$$

where $\hat{R}$ is the restricted Ricci scalar.

17.4.5 Gravitational equations of motion for Bianchi spacetimes

The assumption of spatial homogeneity implies that the energy-momentum tensor of a Bianchi spacetime can vary in time but must be spatially constant. The components of the energy-momentum tensor T_{mn} are the energy density $\rho(t)$, the energy flux $f_a(t)$, and the pressure $p_{ab}(t)$,

$$T_{00} = \rho , \quad T_{0a} = -f_a , \quad T_{ab} = p_{ab} . \quad (17.136)$$

Note that ρ is the proper energy density, which is so because the time components of the tetrad metric are Minkowski, $\gamma_{00} = -1$ and $\gamma_{0a} = 0$. The trace of the energy-momentum tensor is $T = -\rho + 3p$ where $p \equiv \frac{1}{3}p_a^a$. In the special case of a perfect fluid at rest in the tetrad frame (which is not being assumed here), the energy flux f_a would vanish, and the pressure tensor would be proportional to the spatial metric tensor, $p_{ab} = p\gamma_{ab}$. The ADM equations of motion for a Bianchi spacetime are, equations (17.52) and (17.63),

$$\frac{d\gamma_{ab}}{dt} = 2K_{ab} , \quad (17.137a)$$

$$\frac{dK_{ab}}{dt} - 2K_a^c K_{bc} + K_{ab}K + \hat{R}_{ab} = R_{ab} = 4\pi [2p_{ab} + \gamma_{ab}(\rho - 3p)] . \quad (17.137b)$$

The Hamiltonian and momentum constraints (17.74) are

$$\frac{1}{2}(-K^{ab}K_{ab} + K^2 + \hat{R}) = 8\pi\rho , \quad (17.138a)$$

$$\Gamma_{cb}^b K_a^c - \Gamma_{ab}^c K_c^b = -8\pi f_a . \quad (17.138b)$$

Equations (17.137) combine to yield a second order ordinary differential equation for the spatial tetrad metric $\gamma_{ab}(t)$. The spatial tetrad metric can be thought of as an ellipsoid, described by the lengths of its 3 axes, and 3 rotation angles. The general solution to equations (17.137) is a tetrad ellipsoid that evolves in both size and rotation. Equation (17.133a) gives an equation for the evolution of the expansion rate K of the comoving volume element,

$$\dot{K} + K^{ab}K_{ab} = -R_{00} = -4\pi(\rho + 3p) , \quad (17.139)$$

which is the same as the trace of the equation of motion (17.137b) minus twice the Hamiltonian constraint (17.138a). Equation (17.139) is the Raychaudhuri equation (17.76) in a Bianchi spacetime. Since the spatial metric γ_{ab} is positive definite (all positive eigenvalues), $K^{ab}K_{ab}$ is positive.

17.5 Friedmann-Lemaître-Robertson-Walker spacetimes

Friedmann-Lemaître-Robertson-Walker spacetimes are isotropic in addition to being homogeneous. FLRW spacetimes form a subclass of Bianchi spacetimes for which the 3 scale factors a_a are all equal. Applying the dreibein from Table 17.2 with all three scale factors equal reveals that Type IX includes a strictly closed FLRW universe, while Types V and VII include an open FLRW universe. The special case $k = 0$, corresponding to Types I and VII₀, yields a flat FLRW universe. Exercise 17.4 derives explicitly the relation between Bianchi IX coordinates and Cartesian coordinates of the enclosing 4-dimensional Euclidean space.

The Friedmann equations can be cast into ADM form regardless of whether or not the structure coefficients are constant. For example, the isotropic (Poincaré) form (10.26) of the FLRW line-element is

$$ds^2 = -dt^2 + \frac{4a^2}{[1 + \kappa(x^2 + y^2 + z^2)]^2} (dx^2 + dy^2 + dz^2) , \quad (17.140)$$

which is in ADM form with unit lapse and zero shift. The line-element (17.140) takes ADM form (17.108) with spatial tetrad metric

$$\gamma_{ab} = a^2 \delta_{ab} , \quad (17.141)$$

and spatial vierbein

$$e^a{}_{\alpha} = \frac{2\delta_{\alpha}^a}{1 + \kappa(x^2 + y^2 + z^2)} . \quad (17.142)$$

The structure coefficients, equation (17.111), have zero symmetric part, and non-constant antisymmetric part given by

$$k_e = -\kappa x^e . \quad (17.143)$$

The extrinsic curvature K_{ab} is

$$K_{ab} = \frac{\dot{a}}{a} \gamma_{ab} , \quad (17.144)$$

and its trace K is 3 times the Hubble parameter,

$$K = \frac{3\dot{a}}{a} . \quad (17.145)$$

The restricted Ricci tensor $\hat{R}_{ab}$ is

$$\hat{R}_{ab} = \frac{2\kappa}{a^2} \gamma_{ab} , \quad (17.146)$$

and the restricted Ricci scalar $\hat{R}$ is

$$\hat{R} = \frac{6\kappa}{a^2} . \quad (17.147)$$

The Hamiltonian constraint (17.74a) is

$$\frac{3}{a^2} (\dot{a}^2 + \kappa) = 8\pi\rho , \quad (17.148)$$

which reproduces the first of the Friedmann equations (10.30). The equations of motion reduce to

$$\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + 2\frac{\kappa}{a^2} = 4\pi(\rho - p) . \quad (17.149)$$

With a factor of the Hamiltonian constraint (17.148) subtracted, the equation of motion (17.149) becomes

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p) , \quad (17.150)$$

which reproduces the second of the Friedmann equations (10.30). Equation (17.150) is the Raychaudhuri equation (17.76) for an FLRW spacetime.

17.6 BKL oscillatory collapse

An application of Bianchi spacetimes that is of particular relevance to black holes is the collapse of a Type VIII or IX Bianchi spacetime to a singularity, which shows a complicated oscillatory behaviour called Belinskii-Khalatnikov-Lifshitz (BKL) oscillations (Belinskii, Khalatnikov, and Lifshitz, 1970; Belinskii and Khalatnikov, 1971; Belinskii, Khalatnikov, and Lifshitz, 1972; Belinskii, Khalatnikov, and Lifshitz, 1982; Belinski, 2014). BKL oscillations are also called **mixmaster** oscillations. The prototypical BKL model is a Bianchi spacetime, which is spatially homogeneous, but Belinskii, Khalatnikov, and Lifshitz (1982) argue that oscillatory behaviour is generic for collapse to a singularity in general inhomogeneous spacetimes. See Berger (2002) and Belinski (2014) for reviews.

In BKL collapse, the comoving volume element decreases monotonically to zero in a finite proper time, but one spatial axis always expands while the other two collapse. When one of the collapsing axes becomes

sufficiently small, it “bounces” and starts expanding, while the previously expanding axis turns around and starts collapsing. Although the behaviour is deterministic, the sensitivity to initial conditions makes it look chaotic. Bounces occur irregularly in logarithmic time, and there can be an infinite number of bounces during the finite proper time that it takes to reach the singularity. Of course, this ignores quantum gravity, which presumably does something once either the density or the curvature reaches the Planck scale. Between BKL bounces, the three spatial axes expand or contract approximately as power laws $a_a \propto t^{q_a}$ in time t with different exponents q_a , following a behaviour discovered by Kasner (1921), Exercise 17.5. BKL call the phases between bounces Kasner epochs.

The simplest BKL model is where the tetrad metric γ_{ab} is diagonal, and remains so as it evolves. The axes of the tetrad ellipsoid $\gamma_{ab}(t)$ of the Bianchi spacetime in this case change in size, but they do not rotate. This non-rotating model is the one pursued in this section, and that you will explore in Exercise 17.8. Belinskii, Khalatnikov, and Lifshitz (1982) show that when rotation is included, each BKL bounce changes not only the expansion or contraction of each axis, but also changes its orientation. The behaviour between bounces remains Kasner.

The equations of motion (17.137) for a Bianchi spacetime show that the tetrad metric γ_{ab} remains diagonal if and only if the extrinsic curvature K_{ab} is diagonal in the same frame. The extrinsic curvature remains diagonal if the restricted Ricci tensor $\hat{R}_{ab}$ and the pressure tensor p_{ab} are diagonal in the same frame. The tetrad metric γ_{ab} in the non-rotating diagonal frame is conveniently written in terms of scale factors a_a along each of the three diagonal directions,

$$\gamma_{ab}(t) = a_a^2 \delta_{ab} . \quad (17.151)$$

The corresponding diagonal extrinsic curvature K_{ab} is then, from equation (17.137a),

$$K_{ab} = \frac{\dot{a}_a}{a_a} \gamma_{ab} , \quad (17.152)$$

whose trace is

$$K \equiv K_a^a = \sum_a \frac{\dot{a}_a}{a_a} . \quad (17.153)$$

The pressure is diagonal by assumption, with pressure p_a in the a th direction,

$$p_{ab} = p_a \gamma_{ab} , \quad (17.154)$$

whose trace is

$$p_a^a \equiv 3p = \sum_a p_a . \quad (17.155)$$

The equation of motion (17.137b) for the extrinsic curvature K_{ab} involves the restricted Ricci tensor $\hat{R}_{ab}$. A feature of Bianchi spacetimes (with spatially constant structure coefficients) is that the restricted Ricci tensor $\hat{R}_{ab}$, equation (17.134), is given in terms of the structure coefficients c_{ab}^c , equation (17.126), and the tetrad metric γ_{ab} without the need for an explicit expression for the vierbein. The structure coefficients can in turn be written in terms of the structure matrix n^{dc} , equation (17.110). The classification of Bianchi Types in

§17.4.3 used a frame in which the symmetric part of n^{dc} was diagonal, but that frame need not coincide with the frame where the tetrad metric γ_{ab} is diagonal. For a diagonal tetrad metric γ_{ab} but arbitrary structure matrix n^{dc} (subject to the Jacobi identity (17.112)), the components $\hat{R}_{ab}$ of the restricted Ricci tensor are

$$\hat{R}_{11} = a_1^2 \left(\frac{n^{22}n^{33} - (n^{23})^2 - (n^{32})^2}{a_1^2} + \frac{2n^{31}n^{[13]}}{a_2^2} + \frac{2n^{21}n^{[12]}}{a_3^2} + \frac{(n^{11})^2 a_1^2}{2a_2^2 a_3^2} - \frac{(n^{22})^2 a_2^2}{2a_3^2 a_1^2} - \frac{(n^{33})^2 a_3^2}{2a_1^2 a_2^2} \right), \quad (17.156a)$$

$$\hat{R}_{23} = n^{11}n^{(23)} + n^{31}n^{(12)} + n^{21}n^{(13)} + \frac{n^{22}n^{32}a_2^2 + n^{33}n^{23}a_3^2}{a_1^2}, \quad (17.156b)$$

and similarly with cyclically permuted indices for the other components. There are various solutions for which the off-diagonal components $\hat{R}_{23}$ and company, equation (17.156b), vanish; but the simplest condition yielding a diagonal restricted Ricci tensor $\hat{R}_{ab}$ is that the structure matrix n^{dc} be diagonal. If the structure matrix n^{dc} is diagonal (in which case the diagonal elements coincide with the eigenvalues, $n^{cc} = n_c$ for $c = 1, 2, 3$), then the restricted Ricci tensor (17.156) simplifies to

$$\hat{R}_{11} = a_1^2 \left(\frac{n_2 n_3}{a_1^2} + \frac{n_1^2 a_1^2}{2a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{2a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{2a_1^2 a_2^2} \right), \quad (17.157a)$$

$$\hat{R}_{23} = 0, \quad (17.157b)$$

and similarly with permuted indices for the other components. This is the solution pursued here, since it is the simplest that leads to BKL behaviour.

In this case, where the structure matrix n^{dc} is diagonal in the frame where the metric γ_{ab} is diagonal, the ADM equations of motion (17.137) imply that the equation of motion for the scale factor a_1 is

$$\frac{\ddot{a}_1}{a_1} + \frac{\dot{a}_1 \dot{a}_2}{a_1 a_2} + \frac{\dot{a}_1 \dot{a}_3}{a_1 a_3} + \frac{n_2 n_3}{a_1^2} + \frac{n_1^2 a_1^2}{2a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{2a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{2a_1^2 a_2^2} = \frac{R_{11}}{a_1^2} = 4\pi(2p_1 + \rho - 3p), \quad (17.158)$$

and like equations with permuted indices for a_2 and a_3 . The Hamiltonian constraint (17.74a) is

$$\frac{\dot{a}_2 \dot{a}_3}{a_2 a_3} + \frac{\dot{a}_3 \dot{a}_1}{a_3 a_1} + \frac{\dot{a}_1 \dot{a}_2}{a_1 a_2} + \frac{n_2 n_3}{2a_1^2} + \frac{n_3 n_1}{2a_2^2} + \frac{n_1 n_2}{2a_3^2} - \frac{n_1^2 a_1^2}{4a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{4a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{4a_1^2 a_2^2} = 8\pi\rho. \quad (17.159)$$

The momentum constraints (17.74b) vanish identically, enforcing zero energy-momentum flux, $f_a = 0$. You will explore how these equations lead to BKL oscillatory collapse to a singularity in Exercise 17.8.

A central part of the Belinskii, Khalatnikov, and Lifshitz (1982) argument that BKL oscillations are generic in gravitational collapse to a singularity, as opposed to an artefact of the assumption of spatial homogeneity, involves the dependence on time of the terms in the equations of motion (17.158) (which are really just the Einstein equations). The terms involving scale factors a_a but not their time derivatives act as “potentials” that are responsible for BKL bounces when one of the collapsing scale factors becomes sufficiently small. The potentials arise from the products of spatial connections in the restricted Ricci tensor $\hat{R}_{ab}$, equation (17.134). The form of the dependence of the restricted Ricci tensor on the scale factors follows from the fact that the restricted Ricci tensor (17.134) is proportional to two powers of the contravariant metric γ^{cd} , and two powers of the covariant metric γ_{cd} , and that one of the indices on one of the powers of the covariant metric must

be one of the indices a or b of the Ricci component R_{ab} . This form of the dependency of the Ricci tensor on the metric is generic.

The notable exception mentioned at the beginning of §17.4 of a homogeneous space that cannot be realised as a Bianchi space (the vierbein cannot be chosen such that structure coefficients c_{ab}^c are spatially constant), at least as long as the structure coefficients are taken to be real, is the closed ($\kappa > 0$) cylindrical space realised by the spatial vierbein

$$e^\alpha{}_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\sqrt{\kappa}x) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad e_a{}^\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sec(\sqrt{\kappa}x) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (17.160)$$

An open ($\kappa < 0$) cylindrical space on the other hand can be realised as a Bianchi space of Type III, which is the subcase of Type VI with $\kappa = k = 1$.

Exercise 17.5 Kasner spacetime. The Kasner (1921) line-element is

$$ds^2 = -dt^2 + a_x^2 dx^2 + a_y^2 dy^2 + a_z^2 dz^2, \quad (17.161)$$

where $a_\alpha(t)$ are functions only of time t . What Bianchi type is the Kasner line-element (17.161)? Show that the Kasner line-element (17.161) solves the vacuum Einstein equations if

$$a_\alpha = |t|^{q_\alpha} \quad (17.162)$$

with either $q_x = q_y = q_z = 0$ (Minkowski spacetime), or

$$q_x + q_y + q_z = 1, \quad q_x^2 + q_y^2 + q_z^2 = 1. \quad (17.163)$$

Show that a parametric solution of equations (17.163) is

$$q_x = \frac{-u}{1+u+u^2}, \quad q_y = \frac{1+u}{1+u+u^2}, \quad q_z = \frac{u(1+u)}{1+u+u^2}. \quad (17.164)$$

Plot the q_a versus u . Show that, if q_a are ordered such that $q_1 \leq q_2 \leq q_3$, then

$$-\frac{1}{3} \leq q_1 \leq 0 \leq q_2 \leq \frac{2}{3} \leq q_3 \leq 1. \quad (17.165)$$

Solution. Type I. See Figure 17.1.

Exercise 17.6 Schwarzschild interior. Inside the horizon of the Schwarzschild geometry, where the horizon function Δ is negative, the Killing vector associated with time translation symmetry becomes spacelike, so the spacetime has three spacelike Killing vectors, and is therefore spatially homogeneous. The line-element inside the horizon is

$$ds^2 = -dR^2 + |\Delta|dt^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad (17.166)$$

where $dR \equiv dr/\sqrt{|\Delta|}$. The line-element (17.166) is in the form (17.108) with time coordinate R , spatial coordinates t, θ, ϕ , spatial tetrad metric

$$\gamma_{ab} = \text{diag}(|\Delta|, r^2, r^2), \quad (17.167)$$

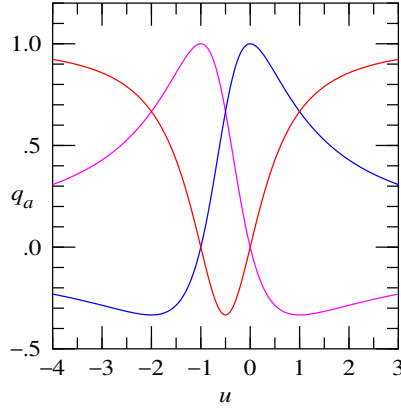


Figure 17.1 Power law indices q_a , equation (17.164), of the scale factors a_α in the vacuum Kasner solution.

and spatial vierbein and inverse vierbein

$$e^a{}_\alpha = \text{diag}(1, 1, \sin \theta) , \quad e_a{}^\alpha = \text{diag}(1, 1, 1/\sin \theta) . \quad (17.168)$$

Is the Schwarzschild line-element (17.166) a Bianchi spacetime? Argue that the Schwarzschild interior resembles a Kasner geometry near the singularity.

Solution. No, the Schwarzschild line-element (17.166) is not a Bianchi spacetime, because its structure coefficients are not independent of the spatial coordinates: the one non-vanishing structure coefficient, $c_{23}^3 = -\cot \theta$, depends on the spatial coordinate θ . However, the causal patch of an infalling object near the singularity, $r \rightarrow 0$, extends over an ever-narrowing angular region, so θ can be treated as effectively constant near the singularity. To the extent that θ is essentially constant, the Schwarzschild geometry approximates a Bianchi spacetime of Type I. The interior near the singularity approximates vacuum Kasner (17.161) with $t \propto r^{3/2}$, and $q_1 = -\frac{1}{3}$, $q_2 = q_3 = \frac{2}{3}$.

Exercise 17.7 Kasner spacetime for a perfect fluid. A generalization of the Kasner line-element (17.161) is

$$ds^2 = -dt^2 + \sum_{\alpha} a_{\alpha}^2 dx_{\alpha}^2 , \quad (17.169)$$

with scale factors

$$a_{\alpha} = a T^{q_{\alpha}-1/3} , \quad (17.170)$$

where the overall scale factor $a(t)$ and time coordinate $T(t)$ are functions of time t , and the constants q_a are Kasner coefficients satisfying equation (17.163). The overall scale factor a satisfies

$$a \equiv (a_1 a_2 a_3)^{1/3} . \quad (17.171)$$

1. Show that the Einstein tensor corresponding to the Kasner line-element (17.169) is diagonal.
2. Show that the energy-momentum is that of a perfect fluid (i.e. the pressure is isotropic, with tetrad-frame pressures $p_a \equiv T_{aa} = p$ all equal) provided that the scale factor a and time coordinate T are related by

$$a = \left(3\kappa \frac{dt}{d \ln T} \right)^{1/3}, \quad (17.172)$$

where κ is a real constant. Notice that the Kasner spacetime is not isotropic even though the energy-momentum is isotropic.

3. Show that in this case of a perfect fluid the tetrad-frame Einstein equations are

$$G_{00} = 3 \left(\frac{\dot{a}^2}{a^2} - \frac{\kappa^2}{a^6} \right) = 8\pi\rho, \quad (17.173a)$$

$$G_{aa} = -\frac{2\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{3\kappa^2}{a^6} = 8\pi p. \quad (17.173b)$$

The Einstein equations (17.173) resemble those (10.29) of the FLRW geometry except that the curvature terms κ/a^2 in FLRW are replaced by terms proportional to $-\kappa^2/a^6$.

4. The Hubble parameter is defined by $H \equiv \dot{a}/a$ as in FLRW. Conclude that the evolution of the scale factor $a(t)$ with time t is determined by the same equation (10.71) as for FLRW,

$$t = \int \frac{da}{aH}. \quad (17.174)$$

5. Show that the Einstein equations (17.173) enforce that the energy-momentum of the perfect fluid satisfies the first law of thermodynamics, similarly to FLRW, §10.9.2,

$$\frac{d\rho a^3}{dt} + p \frac{da^3}{dt} = 0. \quad (17.175)$$

6. From the first law of thermodynamics, show that for a perfect fluid with equation of state $p/\rho = w = \text{constant}$, the density ρ is related to scale factor a by, as in FLRW,

$$\rho \propto a^{-3(1+w)}. \quad (17.176)$$

7. More generally, as in FLRW, the energy-momentum may comprise multiple perfect fluid components x satisfying the first law (17.175). The critical density ρ_{crit} is defined in terms of the Hubble parameter H in the usual way by equation (10.47). Argue that the Kasner Einstein equation (17.173a) implies that

$$\frac{3H^2}{8\pi} \equiv \rho_{\text{crit}} = \rho_\kappa + \sum_{\text{species } x} \rho_x, \quad (17.177)$$

which differs from FLRW, equation (10.73), in that the FLRW curvature density $\rho_\kappa \propto a^{-2}$, equation (10.49), is replaced by the Kasner curvature density $\rho_\kappa \propto a^{-6}$,

$$\rho_\kappa \equiv \frac{3\kappa^2}{8\pi a^6}. \quad (17.178)$$

The Kasner curvature density ρ_κ behaves like a perfect fluid with positive energy and an ultra-hard equation of state, $w = 1$.

8. Define a_κ and H_κ to be the cosmic scale factor and Hubble parameter at density-curvature equality, where $\rho = \rho_\kappa = \frac{1}{2}\rho_{\text{crit}}$. Show that

$$\kappa = \frac{a_\kappa^3 H_\kappa}{\sqrt{2}} . \quad (17.179)$$

9. From equation (17.172) conclude that T equals an integral over scale factor a ,

$$\ln T = 3\kappa \int \frac{da}{a^4 H} . \quad (17.180)$$

Conclude that for a single perfect fluid with $p/\rho = w = \text{constant}$,

$$T/T_\kappa = \frac{(a/a_\kappa)^3}{[1 + \sqrt{1 + (a/a_\kappa)^{3(1-w)}}]^{2/(1-w)}} . \quad (17.181)$$

Conclude that the small and large a limits of T are, for $w \leq 1$,

$$T/T_\kappa \rightarrow \begin{cases} (a/a_\kappa)^3 & a \ll a_\kappa , \\ 1 & a \gg a_\kappa . \end{cases} \quad (17.182)$$

Hence conclude that the perfect fluid Kasner solution goes over to vacuum Kasner for small a and to FLRW for large a . The solution approximates vacuum Kasner at small a not because physical densities are going to zero, but rather because the density becomes dominated by the Kasner curvature density (17.178).

10. For the particular case of a cosmological constant, $w = -1$, show that $\kappa = \sqrt{\Lambda/3}$, and that

$$a/a_\kappa = \sinh^{1/3}(\sqrt{3\Lambda} t) , \quad T/T_\kappa = \tanh(\sqrt{3\Lambda} t/2) . \quad (17.183)$$

Exercise 17.8 Oscillatory Belinskii-Khalatnikov-Lifshitz (BKL) instability. The contracting phase of a Type VIII or IX Bianchi spacetime provides a model of collapse to a singularity that illustrates how complicated such a collapse can be (Belinskii, Khalatnikov, and Lifshitz, 1982). Type VIII and IX Bianchi spacetimes have all three eigenvalues n_a non-zero, and k_a therefore necessarily all zero.

1. Define q_a by

$$q_a \equiv \frac{d \ln a_a}{d \ln |t|} , \quad (17.184)$$

and let q be their sum,

$$q \equiv \sum q_a = \frac{d \ln(a_1 a_2 a_3)}{d \ln |t|} . \quad (17.185)$$

Note that in a collapsing spacetime, t is negative and tending to zero, and $\ln |t| \rightarrow -\infty$ as $|t| \rightarrow 0$, so q_a is positive for a collapsing scale factor a_a . Define further

$$A_a \equiv n_a a_a^2 . \quad (17.186)$$

Show that the equations of motion (17.158) are

$$\frac{dq_1}{d \ln |t|} + q_1(q-1) = \frac{1}{2} \left(\frac{t}{a_1 a_2 a_3} \right)^2 [(A_2 - A_3)^2 - A_1^2] + 4\pi t^2 (2p_1 + \rho - 3p) , \quad (17.187a)$$

$$\frac{dq_2}{d \ln |t|} + q_2(q-1) = \frac{1}{2} \left(\frac{t}{a_1 a_2 a_3} \right)^2 [(A_1 - A_3)^2 - A_2^2] + 4\pi t^2 (2p_2 + \rho - 3p) , \quad (17.187b)$$

$$\frac{dq_3}{d \ln |t|} + q_3(q-1) = \frac{1}{2} \left(\frac{t}{a_1 a_2 a_3} \right)^2 [(A_1 - A_2)^2 - A_3^2] + 4\pi t^2 (2p_3 + \rho - 3p) , \quad (17.187c)$$

and that the Hamiltonian constraint (17.159) is

$$q^2 - \sum q_a^2 = \frac{1}{4} \left(\frac{t}{a_1 a_2 a_3} \right)^2 [2(A_1^2 + A_2^2 + A_3^2) - (A_1 + A_2 + A_3)^2] = 16\pi t^2 \rho . \quad (17.188)$$

2. In gravitational collapse as $t \rightarrow -0$, the scale factors a_a might be expected to become small. Argue that if the right hand sides of equations (17.187) and (17.188) are neglected, then the solution is the Kasner solution, with q_a constant, satisfying equation (17.163).
3. In the Kasner solution, the q_a satisfy the inequalities (17.165). Argue that if q_a are ordered $q_1 < q_2 < q_3$, then Kasner evolution tends to drive the A_a so that $|A_1| > |A_2| > |A_3|$. Then argue from equations (17.187) that the effect of the A_a terms on the right hand sides is to drive smaller q_a to increase, and larger q_a to decrease.
4. Explore the evolution of the scale factors a_a numerically in the case of vanishing energy-momentum, $\rho = p_a = 0$. Choose either Type VIII or Type IX: they are equally fun. You will find better numerical behaviour by transforming to a time variable τ defined by

$$\frac{d}{d\tau} \equiv a_1 a_2 a_3 \frac{d}{dt} = \frac{a_1 a_2 a_3}{t} \frac{d}{d \ln |t|} , \quad (17.189)$$

which increases as (negative) t increases to -0 , and $\ln |t|$ decreases. Define

$$Q_a \equiv -\frac{1}{2} \frac{d \ln |A_a|}{d\tau} = \frac{a_1 a_2 a_3}{-t} q_a , \quad (17.190)$$

which has the same sign as q_a . Show that the equation of motion (17.187a) for A_1 is

$$\frac{dQ_1}{d\tau} = \frac{1}{2} [A_1^2 - (A_2 - A_3)^2] - 4\pi (a_1 a_2 a_3)^2 (2p_1 + \rho - 3p) , \quad (17.191)$$

and similarly for A_2 and A_3 . Show that the Hamiltonian constraint (17.188) is

$$Q_2 Q_3 + Q_3 Q_1 + Q_1 Q_2 = \frac{1}{2} (A_1^2 + A_2^2 + A_3^2) - \frac{1}{4} (A_1 + A_2 + A_3)^2 + 8\pi (a_1 a_2 a_3)^2 \rho . \quad (17.192)$$

The equation of motion for $t/(a_1 a_2 a_3)$ tends to become unstable when $a_1 a_2 a_3$ is small. These circumstances are precisely those where $q = 1$ to good accuracy. Thus when instability arises for small $a_1 a_2 a_3$, it can be worked around by enforcing $q = 1$.

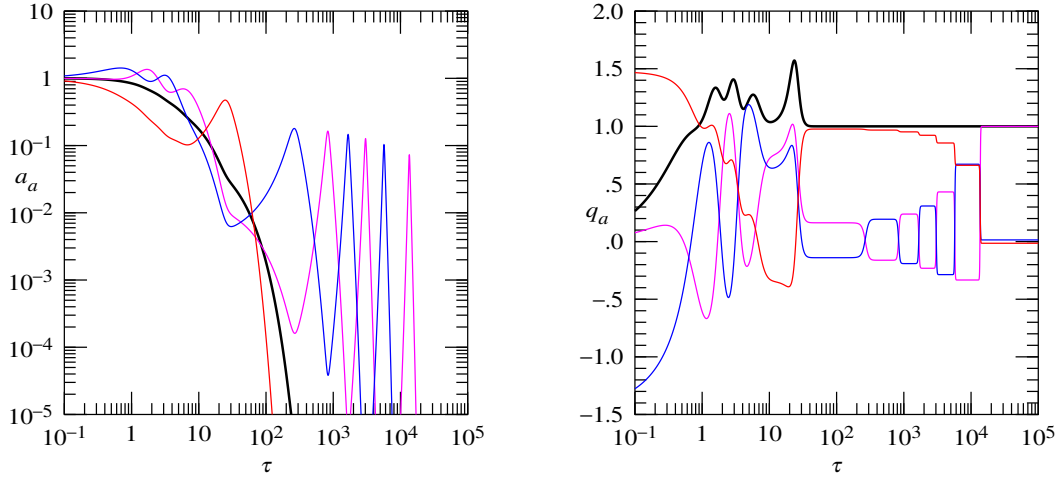


Figure 17.2 Left panel: Cosmic scale factors a_a in BKL collapse of a vacuum Bianchi Type IX spacetime (with eigenvalues normalized to $n_a = 1$). The thick (black) line is the geometric average $(a_1 a_2 a_3)^{1/3}$ of the scale factors, which is proportional to the cube root of the comoving volume element. Right panel: Logarithmic derivatives q_a of the scale factors, equation (17.184). The thick (black) line is the sum $q \equiv q_1 + q_2 + q_3$ of the logarithmic derivatives, which asymptotes to 1 as collapse proceeds. The initial conditions were $a_1 = a_2 = a_3 = 1$ and such that the comoving volume element was initially barely collapsing, $Q_1 = -\frac{6}{7}$, $Q_2 = 0$, $Q_3 = \frac{7}{8}$, whence $\sum Q_a = \frac{1}{56}$. In the initial conditions, the Hamiltonian constraint (17.192) determines the third Q_a in terms of the other two. Integration established a posteriori that the initial time was $t_0 = -1.6859987$. By the end of the plotted era, where $\tau = 10^5$, the comoving volume element had shrunk to $a_1 a_2 a_3 \approx 10^{-230}$.

5. Show that for energy-momentum with equation of state $p_a/\rho = w_a$ with constant w_a , the proper energy density ρ varies as

$$\rho \propto a_1^{-(1+w_1)} a_2^{-(1+w_2)} a_3^{-(1+w_3)}. \quad (17.193)$$

By comparing energy-momentum terms to the largest A_a terms on the right hand sides of equations (17.191), conclude that the influence of energy-momentum is sub-dominant as $t \rightarrow -0$. [Hint: If $q_1 \leq 0 \leq q_2 \leq q_3$ as in the Kasner solution, equation (17.165), then the largest A_a is A_1 as $t \rightarrow -0$. Subject to $-1 \leq w_a \leq 1$, the energy-momentum takes its largest value when $w_1 = -1$ and $w_2 = w_3 = 1$, in which case $(a_1 a_2 a_3)^2 \rho \propto a_1^2 \propto A_1$, which is dominated by A_1^2 as $t \rightarrow -0$.]

6. Following Belinskii, Khalatnikov, and Lifshitz (1970), show that as $t \rightarrow -0$ the collapse may be described as a sequence of Kasner epochs punctuated by bounces. The Kasner exponents q_a before a bounce are given by equation (17.164) for some $u \geq 1$. After the bounce, the exponents q_a satisfy the same equation with u flipped,

$$u \rightarrow -u. \quad (17.194)$$

For $u \geq 2$ the flip reorders the smaller pair of q_a while the largest q_a remains the largest. For $1 \leq u \leq 2$ the flip takes the smallest q_a to the largest, leaving the other pair in original order. To prepare for the

next bounce, reset $u \geq 1$ by transforming

$$u \rightarrow \begin{cases} (u-1)^{-1} & 1 \leq u \leq 2, \\ u-1 & u \geq 2. \end{cases} \quad (17.195)$$

Solution. Figure 17.2 illustrates an example computation. To avoid premature overflow, the computation used logarithmic quantities $\ln a_a$ and $\ln [|t|/(a_1 a_2 a_3)]$ as variables.

17.7 Numerical considerations

Numerical experiments during the 1990s established that the ADM equations, whether in the original form with momenta π_{ab} , or in the York-modified form with momenta K_{ab} , are numerically unstable.

The most popular formalism for long-term evolution of spacetimes is the Baumgarte-Shapiro-Shibata-Nakamura (BSSN) formalism (Shibata and Nakamura, 1995; Baumgarte and Shapiro, 1998), and variants thereof (Shinkai, 2009; Baumgarte and Shapiro, 2010; Brown et al., 2012). The BSSN formalism differs from ADM in that it adjoins equations of motion (17.209) for a vector set of 3 BSSN momentum variables $\hat{H}_\alpha$, and treats the definition (17.208) of $\hat{H}_\alpha$ in terms of derivatives of the metric as a constraint equation. The BSSN equation was discussed in the language of multivector-valued differential forms in §16.16.2.

The superior numerical stability of the BSSN over the ADM formalism can be attributed to the fact that BSSN is **strongly hyperbolic**, §17.7.1, whereas ADM is only weakly hyperbolic (Kreiss and Ortiz, 2002; Nagy, Ortiz, and Reula, 2004).

17.7.1 Strong hyperbolicity

For numerical work, it is not sufficient to have an integrable set of equations. Integrability does not guarantee good numerical behavior, if small errors in the initial conditions blow up exponentially. A condition that guarantees good numerical behavior is that the system be strongly hyperbolic (Kreiss and Ortiz, 2002; Nagy, Ortiz, and Reula, 2004; Hilditch, 2013). Loosely speaking, strong hyperbolicity requires that perturbations to initial conditions propagate as waves $\sim e^{i\omega t}$ rather than growing exponentially $\sim e^{\omega t}$.

Strong hyperbolicity for a first-order system of partial differential equations is defined as follows (Hilditch, 2013). Let u_i denote a set of variables satisfying the first-order system

$$\frac{\partial u_i}{\partial t} + A_{ij}^\alpha \frac{\partial u_j}{\partial x^\alpha} + \dots = 0, \quad (17.196)$$

where $\dots$ does not involve derivatives of the variables. The matrix A_{ij}^α for each spatial coordinate direction α is called the principal symbol of the system. The system is called weakly hyperbolic if, for every direction α , all the eigenvalues of the principal symbol are real. The system is called strongly hyperbolic if in addition, for every α , the eigenvectors of the principal symbol form a complete set, and the eigenvector matrix and its inverse are uniformly bounded.

17.8 BSSN formalism

BSSN reorganizes the second derivative structure of the spatial Einstein equations so that their behaviour as wave equations for the spatial metric $g_{\alpha\beta}$ is manifest, equation (17.211). Only the 5 trace-free spatial Einstein equations are genuine wave equations. The spatial trace of the Einstein equations is a non-wave equation, the Raychaudhuri equation (17.76).

The Hamiltonian structure of the BSSN formalism was explored previously, in the language of multivector-valued differential forms, in §16.16.2.

17.8.1 BSSN momentum equation

In the BSSN formalism, the momentum equation is treated as an equation of motion for the evolution with time t of a momentum variable $\hat{H}_\alpha$. To identify what this momentum variable $\hat{H}_\alpha$ is, it is most straightforward to start not with equation (17.40b) for the Riemann components R_{abc0} , as does ADM, but rather with equation (17.40c) for R_{c0ab} . The $ab \leftrightarrow c0$ symmetry of the Riemann tensor R_{abc0} means that the two expressions are identical when expanded in terms of vierbein derivatives, but the two expressions package the connections and their derivatives in different ways. The restricted contribution $\hat{R}_{c0ab}$ to the Riemann tensor, equation (17.42), involves $\partial_0 \hat{\Gamma}_{abc}$, which is a time derivative of an expression $\hat{\Gamma}_{abc}$ involving spatial derivatives of the vierbein, which looks promising as a precursor of an object whose time evolution might be governed by a momentum equation. However, the other derivative $\partial_c \hat{\Gamma}_{ab0}$ in $\hat{R}_{c0ab}$ also includes mixed time-space second derivatives of the vierbein.

As with the earlier ADM equations of motion (17.55) for $g_{\alpha\beta}$ and (17.68) for $K_{\alpha\beta}$, the identity of the object whose time evolution is being governed becomes manifest in the coordinate frame, where the spatial tetrad is set equal to the spatial coordinate tangent axes, equation (17.14). The desired equation for the coordinate-frame Riemann components $R_{\gamma t \alpha \beta}$ can be derived from a combination of equations (17.40c) and (17.40d), but is obtained more directly from the general equation (17.34), with the restricted Riemann components from equation (17.35),

$$R_{\gamma t \alpha \beta} = \hat{R}_{\gamma t \alpha \beta} + K_{\beta t} K_{\alpha \gamma} - K_{\alpha t} K_{\beta \gamma} , \quad (17.197a)$$

$$\hat{R}_{\gamma t \alpha \beta} = \frac{\partial \hat{\Gamma}_{\alpha \beta t}}{\partial x^\gamma} - \frac{\partial \hat{\Gamma}_{\alpha \beta \gamma}}{\partial t} + \hat{\Gamma}_{\delta \alpha t} \hat{\Gamma}_{\beta \gamma}^\delta - \hat{\Gamma}_{\delta \beta t} \hat{\Gamma}_{\alpha \gamma}^\delta , \quad (17.197b)$$

where the greek indices are a reminder that this is a coordinate-frame expression, and where the final terms in equations (17.34) and (17.35) vanish because of the symmetry $\Gamma_{\kappa\lambda}^\mu = \Gamma_{\lambda\kappa}^\mu$ of coordinate connections (Christoffel symbols), for vanishing torsion. As shown below, equation (17.202), the index on the restricted coordinate connections in equation (17.197b) is raised with the spatial coordinate metric, not with the full metric,

$$\hat{\Gamma}_{\beta\mu}^\alpha \equiv g^{\alpha\gamma} \hat{\Gamma}_{\gamma\beta\mu} . \quad (17.198)$$

Thanks to the ADM gauge condition $e^0_\alpha = 0$, the non-vanishing components of the coordinate-frame generalized extrinsic curvature $K_{\lambda\mu\nu} \equiv e^l_\lambda e^m_\mu e^n_\nu K_{lmn}$ are, similarly to the tetrad-frame generalized extrinsic

curvature K_{lmn} , those whose first two indices are one spatial α and one time t index,

$$K_{\alpha t \nu} = \alpha K_{\alpha 0 \nu} , \quad (17.199)$$

which like the tetrad-frame generalized extrinsic curvature is antisymmetric in its first two indices αt . The extrinsic curvature is as usual $K_{\alpha\beta} \equiv K_{\alpha 0 \beta} \equiv e^a{}_{\alpha} e^b{}_{\beta} K_{a0b}$, which is symmetric in $\alpha\beta$, while the acceleration is as usual $K_{\alpha} \equiv K_{\alpha 0 0} \equiv e^a{}_{\alpha} K_{a00}$. The tensor $K_{\alpha t}$ in equation (17.197a) is

$$K_{\alpha t} \equiv K_{\alpha 0 t} = e^m{}_t K_{\alpha 0 m} = \alpha K_{\alpha} - \beta^{\delta} K_{\alpha \delta} . \quad (17.200)$$

The decomposition $\Gamma_{\lambda\mu\nu} = \hat{\Gamma}_{\lambda\mu\nu} + K_{\lambda\mu\nu}$, equation (17.27), holds for coordinate connections, but the coordinate connections differ from the tetrad connections by a vierbein derivative, equation (11.44). Thus the restricted coordinate-frame connections $\hat{\Gamma}_{\lambda\mu\nu}$ are related to the restricted tetrad-frame connections $\hat{\Gamma}_{lmn}$ by

$$\hat{\Gamma}_{\lambda\mu\nu} = e^l{}_{\lambda} e^m{}_{\mu} e^n{}_{\nu} (d_{lmn} + \hat{\Gamma}_{lmn}) . \quad (17.201)$$

The vierbein derivative d_{0an} with first index 0 and second index a spatial vanishes because of the ADM gauge condition $e^0{}_{\alpha} = 0$. For convenience, define the restricted coordinate connection with first index a tetrad index k by $\hat{\Gamma}_{k\mu\nu} \equiv e_k{}^{\lambda} \hat{\Gamma}_{\lambda\mu\nu}$. Since $d_{0an} = 0$ it follows that the coordinate-frame connection $\hat{\Gamma}_{0\alpha\nu}$ vanishes like its tetrad-frame counterpart. Consequently the product of coordinate connections $\hat{\Gamma}_{\pi\alpha t} \hat{\Gamma}_{\beta\gamma}^{\pi}$ contracted with the full coordinate metric $g^{\pi\rho}$ equals the product $\hat{\Gamma}_{\delta\alpha t} \hat{\Gamma}_{\beta\gamma}^{\delta}$ contracted with the spatial metric $g^{\delta\epsilon}$,

$$\hat{\Gamma}_{\pi\alpha t} \hat{\Gamma}_{\beta\gamma}^{\pi} = \hat{\Gamma}_{p\alpha t} \hat{\Gamma}_{\beta\gamma}^p = \hat{\Gamma}_{d\alpha t} \hat{\Gamma}_{\beta\gamma}^d = \hat{\Gamma}_{\delta\alpha t} \hat{\Gamma}_{\beta\gamma}^{\delta} , \quad (17.202)$$

which justifies equation (17.198).

The coordinate connection $\hat{\Gamma}_{[\alpha\beta]t}$ antisymmetrized over its spatial indices $\alpha\beta$ is an antisymmetric spatial tensor, which can be denoted $F_{\alpha\beta}$,

$$F_{\alpha\beta} \equiv \hat{\Gamma}_{[\alpha\beta]t} = \frac{1}{2} \left(\frac{\partial \beta_{\beta}}{\partial x^{\alpha}} - \frac{\partial \beta_{\alpha}}{\partial x^{\beta}} \right) . \quad (17.203)$$

The tensorial nature of $F_{\alpha\beta}$ follows from the fact that the coordinate-frame curl of a vector is a tensor, Exercise 2.7. Expression (17.197b) for the restricted Riemann tensor can thus be written

$$\hat{R}_{\gamma t \alpha \beta} = \hat{D}_{\gamma} F_{\alpha \beta} - \frac{\partial \hat{\Gamma}_{[\alpha\beta]\gamma}}{\partial t} + \hat{\Gamma}_{(\delta\alpha)t} \hat{\Gamma}_{\beta\gamma}^{\delta} - \hat{\Gamma}_{(\delta\beta)t} \hat{\Gamma}_{\alpha\gamma}^{\delta} , \quad (17.204)$$

in which the only term containing mixed time-space second derivatives is $\partial \hat{\Gamma}_{[\alpha\beta]\gamma} / \partial t$. In equation (17.204), the coordinate connection $\hat{\Gamma}_{(\alpha\beta)t}$ symmetrized over its spatial indices is

$$\hat{\Gamma}_{(\alpha\beta)t} = \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial t} . \quad (17.205)$$

Contracting the Riemann tensor $R_{\gamma t \alpha \beta}$ yields the time-space components $R_{t\alpha}$ of the Ricci tensor,

$$R_{t\alpha} = \hat{R}_{t\alpha} - K_t{}^{\beta} K_{\alpha\beta} + K_{\alpha t} K , \quad (17.206a)$$

$$\hat{R}_{t\alpha} = \hat{D}^{\beta} F_{\beta\alpha} + \frac{\partial \hat{\Gamma}_{[\alpha\beta]}^{\beta}}{\partial t} - \hat{\Gamma}_{(\delta\alpha)t} \hat{\Gamma}^{\delta\beta}{}_{\beta} + \hat{\Gamma}^{(\delta\beta)}{}_{t\beta} \hat{\Gamma}_{\alpha\delta\beta} , \quad (17.206b)$$

in which the only term containing mixed time-space derivatives is $\partial \hat{\Gamma}_{[\alpha\beta]}^{\beta} / \partial t$. In terms of derivatives of the metric, $\hat{\Gamma}_{[\alpha\beta]}^{\beta}$ is

$$\hat{\Gamma}_{[\alpha\beta]}^{\beta} = \frac{1}{2} g^{\beta\gamma} \left(\frac{\partial g_{\alpha\gamma}}{\partial x^{\beta}} - \frac{\partial g_{\beta\gamma}}{\partial x^{\alpha}} \right) = -\frac{1}{2} \frac{g_{\alpha\gamma}}{g} \frac{\partial (g g^{\beta\gamma})}{\partial x^{\beta}}, \quad (17.207)$$

where $g \equiv |g_{\alpha\beta}|$ is the determinant of the spatial metric. Equations (17.206) show that the variable $\hat{\Gamma}_{[\alpha\beta]}^{\beta}$ appears to be the desired BSSN momentum variable. However, it is common to use a variant BSSN momentum variable $\hat{H}_{\alpha}$ in which the spatial metric is scaled by some power of the spatial metric determinant g ,

$$\hat{H}_{\alpha} \equiv \hat{\Gamma}_{\alpha\beta}^{\beta} + \frac{p}{2} \frac{\partial \ln g}{\partial x^{\alpha}} = 2\hat{\Gamma}_{[\alpha\beta]}^{\beta} + \frac{(1+p)}{2} \frac{\partial \ln g}{\partial x^{\alpha}} = -\frac{g_{\alpha\gamma}}{g^{(1-p)/2}} \frac{\partial (g^{(1-p)/2} g^{\beta\gamma})}{\partial x^{\beta}}, \quad (17.208)$$

with p an adjustable constant. For example, the choice $p = -1$ recovers (twice) the original momentum variable $\hat{\Gamma}_{[\alpha\beta]}^{\beta}$, the choice $p = 0$ yields a spatial Ricci tensor (17.210) whose only explicit second spatial derivatives are a Laplacian of the spatial metric, and the choice $p = 1/3$ gives an $\hat{H}_{\alpha}$ that depends only on the scaled spatial metric $g^{-1/3} g_{\alpha\gamma}$ with unit determinant (and its inverse $g^{1/3} g^{\beta\gamma}$). In the BSSN formalism, the evolution of the momentum variable $\hat{H}_{\alpha}$ is governed by the momentum equation

$$\left[\frac{1}{2} \frac{\partial \hat{H}_{\alpha}}{\partial t} = -\frac{(1+p)}{4} \frac{\partial^2 \ln g}{\partial t \partial x^{\alpha}} + \hat{\Gamma}_{(\delta\alpha)t} \hat{\Gamma}^{[\delta\beta]}_{\beta} - \hat{\Gamma}^{(\delta\beta)}_{\beta} \hat{\Gamma}_{\alpha\delta\beta} + \hat{D}^{\beta} F_{\alpha\beta} + K_t^{\beta} K_{\alpha\beta} - K_{\alpha t} K + 8\pi T_{t\alpha} \right]. \quad (17.209)$$

In the BSSN formalism, equation (17.208) is a constraint equation, which must be imposed in the initial conditions, but which is satisfied automatically thereafter.

17.8.2 BSSN spatial Ricci tensor

In the BSSN formalism, the spatial components $\hat{R}_{\alpha\beta}$ of the restricted Ricci tensor are recast in terms of the BSSN variable $\hat{H}_{\alpha}$,

$$\hat{R}_{\alpha\beta} = -\frac{p}{2} \frac{\partial^2 \ln g}{\partial x^{\alpha} \partial x^{\beta}} - \frac{g^{\gamma\delta}}{2} \frac{\partial^2 g_{\alpha\beta}}{\partial x^{\gamma} \partial x^{\delta}} + \frac{1}{2} \frac{\partial \hat{H}_{\beta}}{\partial x^{\alpha}} + \frac{1}{2} \frac{\partial \hat{H}_{\alpha}}{\partial x^{\beta}} - \hat{\Gamma}_{\alpha\beta}^{\delta} \hat{\Gamma}_{\delta\gamma}^{\gamma} + \hat{\Gamma}^{\gamma\delta}_{\alpha} \hat{\Gamma}_{\beta\gamma\delta} + \hat{\Gamma}^{\gamma\delta}_{\beta} \hat{\Gamma}_{\alpha\gamma\delta} + \hat{\Gamma}^{\gamma\delta}_{\alpha} \hat{\Gamma}_{\gamma\delta\beta}. \quad (17.210)$$

The only explicit second spatial derivatives in the expression (17.210) for $\hat{R}_{\alpha\beta}$ are a double gradient of the spatial metric determinant g , and a spatial Laplacian of the spatial metric $g_{\alpha\beta}$, the remaining second derivatives having been absorbed into first spatial derivatives of the BSSN momentum variable $\hat{H}_{\alpha}$.

When the restricted spatial Ricci tensor (17.210) is inserted into the equation of motion (17.68) for the extrinsic curvature $K_{\alpha\beta}$, the spatial Laplacian combines with a second time derivative coming from $\partial K_{\alpha\beta} / \partial t$ to form a 4-dimensional wave equation for the spatial metric $g_{\alpha\beta}$. Thus the character of the spatial Einstein equations as wave equations for the spatial metric $g_{\alpha\beta}$ is manifest in the BSSN formalism. Explicitly, the spatial Einstein equations, which are just the equations of motion (17.68) for the spatial extrinsic curvature

$K_{\alpha\beta}$, are

$$\frac{1}{2} \left[\left(\frac{\partial}{\alpha \partial t} \right)^2 - g^{\gamma\delta} \frac{\partial^2}{\partial x^\gamma \partial x^\delta} \right] g_{\alpha\beta} + \frac{\partial^2 \ln(\alpha g^{-p/2})}{\partial x^\alpha \partial x^\beta} + \dots = 8\pi \left(T_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} T \right), \quad (17.211)$$

where ... signifies terms involving no higher than first time or space derivatives of the lapse α , the shift β^α , the spatial coordinate metric $g_{\alpha\beta}$, the extrinsic curvatures $K_{\alpha\beta}$, or the BSSN variable $\hat{H}_\alpha$.

Commonly, only the 5 trace-free equations of motion for $K_{\alpha\beta}$ are used in the BSSN formalism, the trace equation being replaced by the Raychaudhuri equation (17.76).

17.8.3 BSSN summary

To summarize, the dynamical variables in the BSSN formalism are the spatial metric $g_{\alpha\beta}$, the spatial extrinsic curvature $K_{\alpha\beta}$, and the spatial BSSN variable $\hat{H}_\alpha$. The equations of motion for the dynamical variables are:

1. the 6 equations (17.55) for the spatial metric $g_{\alpha\beta}$;
2. the 5 equations constituting the trace-free part of the 6 equations (17.68) for the spatial extrinsic curvature $K_{\alpha\beta}$;
3. the 1 Raychaudhuri equation (17.76) for the trace K of the extrinsic curvature;
4. the 3 equations (17.209) for the BSSN variable $\hat{H}_\alpha$.

The constraint equations, which must be arranged to be satisfied on the initial hypersurface, but which are thereafter satisfied automatically are:

1. the 1 Hamiltonian constraint (17.74a);
2. the 3 momentum constraints (17.74b);
3. the 3 constraints (17.208) on the BSSN variable $\hat{H}_\alpha$.

The Hamiltonian and momentum constraints are differential constraints, elliptic partial differential equations of second order in the spatial coordinates, which are in general non-trivial to set up. The constraints on $\hat{H}_\alpha$ on the other hand are algebraic constraints, which are straightforward to impose once the differential constraints are solved.

17.9 Pretorius formalism

Pretorius (2005) proposed an elegant 4-dimensional version of the BSSN formalism. A natural 4-dimensional generalization of the BSSN momentum variable $\hat{H}_\alpha$ defined by equation (17.208) is (with $p = 0$)

$$H_\kappa \equiv \Gamma_{\kappa\lambda}{}^\lambda = 2\Gamma_{[\kappa\lambda]}{}^\lambda + \frac{\partial \ln \sqrt{-g}}{\partial x^\kappa} = -\frac{g_{\kappa\mu}}{\sqrt{-g}} \frac{\partial(\sqrt{-g} g^{\lambda\mu})}{\partial x^\lambda}, \quad (17.212)$$

in which $g \equiv |g_{\kappa\lambda}|$ is the determinant of the full 4-dimensional metric. If the coordinates x^κ are treated as four scalars (they are not; and neither do they form a 4-vector), then the contravariant components H^κ can

be written as minus the (torsion-free) d'Alembertian $\square \equiv D_\lambda D^\lambda$ of the coordinates,

$$H^\kappa = -\frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} g^{\lambda\kappa})}{\partial x^\lambda} = -\frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^\lambda} \left(\sqrt{-g} g^{\lambda\mu} \frac{\partial x^\kappa}{\partial x^\mu} \right) = -\square x^\kappa, \quad (17.213)$$

which motivates calling H^κ the **harmonic function**. The coordinates x^κ are not scalars, and neither is the harmonic function H^κ a tensor. In the Pretorius formalism, the Ricci tensor takes the form

$$R_{\kappa\lambda} = -\frac{1}{2} g^{\mu\nu} \frac{\partial^2 g_{\kappa\lambda}}{\partial x^\mu \partial x^\nu} + \frac{1}{2} \frac{\partial H_\lambda}{\partial x^\kappa} + \frac{1}{2} \frac{\partial H_\kappa}{\partial x^\lambda} - \Gamma_{\kappa\lambda}^\nu H_\nu + \Gamma^{\mu\nu}{}_\kappa \Gamma_{\lambda\mu\nu} + \Gamma^{\mu\nu}{}_\lambda \Gamma_{\kappa\mu\nu} + \Gamma^{\mu\nu}{}_\kappa \Gamma_{\mu\nu\lambda}, \quad (17.214)$$

in which the only explicit second derivatives are those in the $g^{\mu\nu} \partial^2 g_{\kappa\lambda} / \partial x^\mu \partial x^\nu$ term. This second derivative term has the form of a 4-dimensional coordinate wave operator acting on the 4-dimensional coordinate metric $g_{\kappa\lambda}$. The Einstein equations are as usual

$$R_{\kappa\lambda} = 8\pi (T_{\kappa\lambda} - \frac{1}{2} g_{\kappa\lambda} T). \quad (17.215)$$

Despite the covariant 4-dimensional character of the Pretorius formalism, it is still possible to make ADM gauge choices, §17.1, that is, to foliate the spacetime into hypersurfaces of constant time t , and to work in an ADM tetrad whose time axis γ_0 is the future-pointing unit normal to hypersurfaces of constant time t . In the ADM tetrad, the tetrad-frame harmonic function $H_k \equiv e_k{}^\kappa H_\kappa$ with H_κ defined by equation (17.212) is, in terms of the vierbein derivatives d_{klm} defined by equation (11.33), the tetrad-frame restricted connections $\hat{\Gamma}_{klm}$, and the generalized extrinsic curvature K_{lmn} ,

$$H_k \equiv e_k{}^\kappa H_\kappa = d_{km}{}^m + \Gamma_{km}{}^m = d_{km}{}^m + \hat{\Gamma}_{km}{}^m + K_{km}{}^m = d_{k0}{}^0 + \hat{H}_k - K_k, \quad (17.216)$$

where $\hat{H}_k \equiv d_{ka}{}^a + \hat{\Gamma}_{ka}{}^a = \{0, \hat{H}_a\} = \{0, e_a{}^\alpha \hat{H}_\alpha\}$, and $\hat{H}_\alpha$ is the BSSN momentum variable defined by equation (17.208) with $p = 0$. The tetrad-frame components H_k of the harmonic function are

$$H_0 = \frac{1}{\alpha} \partial_0 \alpha - K, \quad (17.217a)$$

$$H_a = \frac{1}{\alpha} e_{a\alpha} \partial_0 \beta^\alpha + \hat{H}_a - K_a. \quad (17.217b)$$

Pretorius (2005) points out that the arbitrariness of the choice of coordinates x^κ translates into an arbitrariness in the choice of the 4 components H_κ of the harmonic function. Thus instead of treating the lapse and shift as arbitrarily adjustable functions, the harmonic functions H_κ can be adjusted arbitrarily. For example, the harmonic function can be chosen to vanish identically, $H_\kappa = 0$, a coordinate condition first proposed by Fock (1957). Equations (17.217) can then be interpreted as evolution equations for the lapse α and the shift β^α . In this case the 4 Einstein equations with at least one temporal index are not used as evolution equations.

However, it is also possible (Bona et al., 2003) to follow the BSSN strategy of choosing the lapse and shift arbitrarily, in which case the 4 Einstein equations (17.214) with at least one temporal index provide evolution equations for the harmonic function H_κ , and equations (17.217) are constraint equations that must be imposed on the initial hypersurface, but which are guaranteed thereafter.

As in ADM and BSSN, the Hamiltonian and momentum constraints, along with the conditions (17.213), must be arranged to be satisfied on the initial hypersurface.

17.10 $M+N$ split

In situations where fields are highly relativistic, such as inside black holes, or when following gravitational waves, it can be natural to work in a frame where some of the tetrad axes are null. A null direction γ_v is orthogonal to itself, $\gamma_v \cdot \gamma_v = 0$, so it is not possible to carry out a 3+1 split of spacetime into a 1-dimensional space aligned with γ_v and a 3-dimensional space orthogonal to it. It is however possible, as in the Newman-Penrose formalism, to carry out a 2+2 split of spacetime into a 2-dimensional space spanned by two null directions γ_v and γ_u , and a 2-dimensional space orthogonal to the null directions.

This section 17.10 considers the general case of an $M+N$ split of an $M+N$ -dimensional spacetime.

17.10.1 $M+N$ tetrad and extrinsic curvature

In an $M+N$ split of spacetime, the tetrad-frame axes γ_m at each point are split into two orthogonal sets, of dimensions respectively N and M . Label the N tetrad axes γ_z of the first set with late letters z , and the M tetrad axes γ_a of the second set with early letters a , and let mid letters $kl\dots$ run over all indices. The orthogonality of the tetrad axes from opposite sets is expressed by the MN conditions

$$\gamma_a \cdot \gamma_z = 0 . \quad (17.218)$$

The spaces remain unmixed under raising and lowering indices. (the non-vanishing components of the metric are γ_{ab} and γ_{yz}). In the $M+N$ split, the two orthogonal subspaces at each point are fixed a priori, which amounts to making a specific choice of gauge of the tetrad. The gauge-fixing fixes the two subspaces, but allows tetrad transformations within each subspace. Under this restricted group of tetrad transformations, the tetrad connections Γ_{azm} with first two indices az from opposite subspaces form a tensor, the generalized extrinsic curvature K_{azm} ,

$$K_{azm} \equiv \Gamma_{azm} = \gamma_a \cdot \partial_m \gamma_z . \quad (17.219)$$

These connections form a tensor under the restricted group because the only potentially non-tensorial contribution to $\gamma_a \cdot \partial_m \gamma_z$ under a restricted tetrad transformation $\gamma_z \rightarrow L_z^y \gamma_y$ is

$$\gamma_a \cdot \gamma_y \partial_m L_z^y = 0 , \quad (17.220)$$

which vanishes because γ_a and γ_y are orthogonal. There are $MN(M+N)$ non-vanishing components of the extrinsic curvature K_{azm} (hence 12 if $M=3$ and $N=1$, or 16 if $M=N=2$). The remaining tetrad connections Γ_{mnl} , namely those with first two indices mn from the same subspace, constitute the restricted connections $\hat{\Gamma}_{mnl}$,

$$\hat{\Gamma}_{mnl} \equiv \Gamma_{mnl} \quad \text{for } mn = yz \text{ or } mn = ab . \quad (17.221)$$

The vanishing of the mixed components γ_{az} of the tetrad metric implies that the generalized extrinsic curvature is antisymmetric in its first two indices,

$$K_{zal} = -K_{azl} . \quad (17.222)$$

The vanishing components of K_{mnl} and $\hat{\Gamma}_{mnl}$ are

$$K_{abl} = K_{yzl} = 0 , \quad \hat{\Gamma}_{azl} = \hat{\Gamma}_{zal} = 0 . \quad (17.223)$$

The contraction K_m of the generalized extrinsic curvature is (this repeats equation (17.23))

$$K_m \equiv K_{mn}^n = \{K_{za}^a, K_{az}^z\} = \{K_z, K_a\} . \quad (17.224)$$

17.10.2 $M+N$ Riemann and Ricci tensors

The extrinsic curvature K_{mnl} is a tensor under the restricted group of tetrad transformations. The restricted Riemann curvature tensor $\hat{R}_{klaz}$ with its last two indices from opposite subspaces vanishes since $\hat{\Gamma}_{azk}$ vanishes,

$$\hat{R}_{klaz} = \partial_k \hat{\Gamma}_{azl} - \partial_l \hat{\Gamma}_{azk} + \hat{\Gamma}_{al}^p \hat{\Gamma}_{pzk} - \hat{\Gamma}_{ak}^p \hat{\Gamma}_{pzl} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \hat{\Gamma}_{azp} = 0 . \quad (17.225)$$

For vanishing torsion, the full Riemann curvature tensor R_{klmn} is as usual antisymmetric in each of kl and mn , and symmetric in $kl \leftrightarrow mn$. As remarked after equation (17.35), the restricted Riemann tensor $\hat{R}_{klmn}$ is likewise antisymmetric in each of kl and mn , but it is not symmetric under $kl \leftrightarrow mn$. Thus the components $\hat{R}_{azkl}$ of the restricted Riemann curvature do not vanish even though the components $\hat{R}_{klaz}$ do vanish.

In the $M+N$ split, the expression (17.34) for the Riemann curvature tensor becomes

$$R_{wxyz} = \hat{R}_{wxyz} + K_{yx}^a K_{azw} - K_{yw}^a K_{azx} , \quad (17.226a)$$

$$R_{xyaz} = \hat{D}_x K_{azy} - \hat{D}_y K_{azx} + (K_{xy}^c - K_{yx}^c) K_{azc} \quad (17.226b)$$

$$= R_{azxy} = \hat{R}_{azxy} + K_{xz}^c K_{cya} - K_{xa}^c K_{cyz} , \quad (17.226c)$$

$$R_{byaz} = \hat{D}_b K_{azy} - \hat{D}_y K_{azb} + K_{by}^x K_{azx} - K_{yb}^c K_{azc} , \quad (17.226d)$$

$$R_{bc az} = \hat{D}_b K_{azc} - \hat{D}_c K_{azb} + (K_{bc}^x - K_{cb}^x) K_{azx} \quad (17.226e)$$

$$= R_{azbc} = \hat{R}_{azbc} + K_{bz}^x K_{xca} - K_{ba}^x K_{xcz} , \quad (17.226f)$$

$$R_{abcd} = \hat{R}_{abcd} + K_{cb}^z K_{zda} - K_{ca}^z K_{zdb} . \quad (17.226g)$$

In an $M+1$ split, equations (17.226a)–(17.226c) vanish identically, and equations (17.226) reduce to the four equations (17.40). If the tetrad connections are replaced by their torsion-free expressions in terms of derivatives of the vierbein, then the various alternative expressions for the Riemann tensor in equations (17.226)

become identities. The Ricci tensor R_{km} is

$$R_{yz} = \hat{R}_{yz} + (\hat{D}_a + K_a)K_{zy}^a - \hat{D}_y K_z - K_{ya}^b K_{zb}^a, \quad (17.227a)$$

$$R_{za} = \hat{R}_{za} + (\hat{D}_y + K_y)K_{az}^y - \hat{D}_z K_a - K_{ab}^y K_{zy}^b \quad (17.227b)$$

$$= R_{az} = \hat{R}_{az} + (\hat{D}_b + K_b)K_{za}^b - \hat{D}_a K_z - K_{ab}^y K_{zy}^b \quad (17.227c)$$

$$= \hat{D}_y K_{az}^y - \hat{D}_z K_a + \hat{D}_b K_{za}^b - \hat{D}_a K_z - 2K_{ab}^y K_{zy}^b + K_{ab}^y K_{yz}^b + K_{ba}^y K_{zy}^b \quad (17.227d)$$

$$R_{ab} = \hat{R}_{ab} + (\hat{D}_z + K_z)K_{ba}^z - \hat{D}_a K_b - K_{ay}^z K_{bz}^y, \quad (17.227e)$$

with $\hat{R}_{za} = \hat{R}_{zba}{}^b$ and $\hat{R}_{az} = \hat{R}_{ayz}{}^y$. For vanishing torsion, the full Ricci tensor R_{km} is symmetric, but the restricted Ricci tensor $\hat{R}_{km}$ is not symmetric. Contracting the Ricci tensor yields the Ricci scalar R ,

$$R = \hat{R} - 2\hat{D}_z K^z - 2\hat{D}_a K^a - K^{bza} K_{azb} - K^{zay} K_{yaz} - K^z K_z - K^a K_a. \quad (17.228)$$

Singularity theorems

Singularity theorems prove that, given a number of plausible assumptions, general relativity commits suicide inside black holes. The conclusion that there are places, called singularities, inside black holes where the general relativistic description of spacetime fails is profound. It means that new physics, presumably quantum gravity in some form, *must* replace general relativity at singularities. Any viable theory of quantum gravity must be able to resolve the problem of singularities.

The first singularity theorem was proved by Penrose (1965). The classic book by Hawking and Ellis (1973) lays out a variety of singularity theorems. As reviewed by Senovilla (1998), singularity theorems state that given:

1. a trapped surface condition,
2. a positive energy condition,
3. a causality condition,

then there exist geodesics that are incomplete, in the sense that the geodesics reach a point beyond which they cannot be continued. The power of singularity theorems is that they show that general relativity fails inside black holes. The weakness of singularity theorems is that they are quite unspecific about the nature or location of a “singularity.”

This Chapter focuses on the principal ingredients of the singularity theorems, namely the Raychaudhuri equations, §18.2, and the construction of hypersurface-orthogonal congruences of geodesics, §§18.7 and 18.8. The Chapter concludes, §18.10, with a brief exposition of the original singularity theorem discovered by Penrose (1965).

Whereas in the previous Chapter 17 on the ADM formalism the tetrad metric γ_{kl} was permitted to be non-constant, the present Chapter takes the tetrad metric to be constant

$$\gamma_{kl} = \text{constant} . \tag{18.1}$$

Without loss of generality, for a timelike congruence, §18.3, the tetrad can be taken to be orthonormal, while for a null congruence, §18.4, the tetrad can be taken to be a double-null Newman-Penrose tetrad.

18.1 Congruences

The Raychaudhuri equations govern the evolution of the extrinsic curvature along systems of paths called **congruences**, which fill, and do not cross or overlap in, at least some connected region of spacetime. Congruences may be timelike or null, and they may be geodesic or otherwise. Congruences are often defined with the restriction that the paths do not cross or overlap anywhere in spacetime, but in this book the more relaxed condition is imposed, that paths do not cross or overlap over some connected region.

A path is specified by its coordinates $x^\mu(\lambda)$ as a function of some parameter λ along the path. The derivative of the path defines the 4-velocity u^μ along the path,

$$u^\mu \equiv \frac{dx^\mu}{d\lambda} . \quad (18.2)$$

If the congruence of paths is timelike, then the parameter λ may be taken equal to the proper time τ along the path. The 4-velocity $u^\mu \equiv dx^\mu/d\tau$ then satisfies the normalization condition $u_\mu u^\mu = -1$. The 4-velocity vector $\mathbf{u} \equiv e_\mu u^\mu$ defines the tetrad time vector γ_0 ,

$$\gamma_0 = \mathbf{u} . \quad (18.3)$$

The tetrad time vector γ_0 is the unique future-pointing vector that is tangent to the timelike path and normalized to $\gamma_0 \cdot \gamma_0 = -1$.

If the congruence of paths is null, then λ may be any arbitrary parameter, not necessarily an affine parameter. If the parameter λ is an affine parameter, then the path is said to be affinely parameterized. The 4-velocity $u^\mu \equiv dx^\mu/d\lambda$ satisfies the normalization condition $u_\mu u^\mu = 0$ regardless of whether the parameter λ is affine. The 4-velocity vector $\mathbf{u} \equiv e_\mu u^\mu$ defines the tetrad null vector γ_v (say),

$$\gamma_v = \mathbf{u} . \quad (18.4)$$

Unlike the timelike case, the normalization condition $\gamma_v \cdot \gamma_v = 0$ does not determine uniquely the null vector γ_v .

For either a timelike or a null path, the 4-velocity $\mathbf{u} = u^m \gamma_m$ has tetrad-frame components

$$u^m = \{1, 0, 0, 0\} , \quad (18.5)$$

whose only non-vanishing component is $u^z = 1$, with index $z = 0$ for a timelike path, $z = v$ for a null path. The covariant derivative of the 4-velocity along the path is

$$D_n u_m = \gamma_{mk} D_n u^k = \gamma_{mk} (\partial_n u^k + \Gamma_{ln}^k u^l) = \Gamma_{mzn} . \quad (18.6)$$

The components for spatial $m = a$ constitute by definition the generalized extrinsic curvature K_{azn} , equation (17.219),

$$D_n u_a = K_{azn} . \quad (18.7)$$

The 4-velocity along the path evolves as

$$\frac{Du^k}{D\lambda} = u^n \partial_n u^k + \Gamma_{mn}^k u^m u^n = \Gamma_{zz}^k , \quad (18.8)$$

whose spatial components constitute the acceleration K_{zz}^a ,

$$\frac{Du^a}{D\lambda} = K_{zz}^a . \quad (18.9)$$

For a timelike geodesic ($z = 0$), the time component of the acceleration vanishes automatically, $Du^0/D\lambda = \Gamma_{00}^0 = 0$. For a null geodesic ($z = v$), the v -component of the acceleration $Du^v/D\lambda = \Gamma_{vv}^v = -\Gamma_{uvv}$ vanishes if the path is affinely parameterized, but not in general. If the null path is affinely parameterized, then the 4-velocity u^m coincides (up to a constant factor) with the momentum p^m along the path. Choosing the path to be affinely parameterized amounts to choosing the null vector γ_v such that the momentum p^v is constant along the null geodesic, that is, a light ray is neither redshifted nor blueshifted as it propagates along the affinely parameterized path.

The covariant divergence of the 4-velocity is

$$D_m u^m = \Gamma_{zz}^z + K_z . \quad (18.10)$$

For a timelike congruence, the covariant divergence is just the trace $K \equiv K_0 \equiv K_{0a}^a$ of the extrinsic curvature. For a null congruence, the covariant divergence is the acceleration Γ_{vv}^v plus the trace $K_v \equiv K_{va}^a$ of the extrinsic curvature. If the null path is affinely parameterized, then the covariant divergence is just the trace K_v .

18.2 Raychaudhuri equations

The **Raychaudhuri equations**, which in their most general form are equations (18.11), govern the evolution of the extrinsic curvature along arbitrary timelike or null congruences. Actually, the equation traditionally named after Raychaudhuri (1955) is the equation for the evolution of the trace of the extrinsic curvature. Here however the full suite of equations for the components of the extrinsic curvature are called Raychaudhuri equations.

The Raychaudhuri equations come in various flavours, depending on whether the congruence is timelike or null, whether the congruence is geodesic, and what additional gauge conditions are imposed on the tetrad. If the congruence is timelike, it is convenient to take the tetrad to be orthonormal, with the time axis γ_0 tangent to the timelike paths, equation (18.3). If the congruence is null, it is convenient to take the tetrad to be Newman-Penrose, that is, a double-null tetrad, with the null axis γ_v tangent to the null paths. To cover both timelike and null cases at the same time, denote the tangent axis by γ_z , with index $z = 0$ in the timelike case, and $z = v$ in the null case.

The Raychaudhuri equations are just a subset of the equations (17.226) for the Riemann tensor in an $M+N$ split of spacetime, §17.10. In 4 spacetime dimensions, the split is 3+1 for a timelike congruence, and 2+2 for a null congruence. In an $M+N$ split of spacetime, the Raychaudhuri equations are the equations for the components $R_{bza z}$ of the Riemann tensor, equation (17.226d),

$$\hat{D}_z K_{azb} - \hat{D}_b K_{azz} - K_{bz}^y K_{azy} + K_{zb}^c K_{azc} = -R_{bza z} \quad (\text{no sum over } z) , \quad (18.11)$$

with $z = 0$ for a timelike congruence, or $z = v$ for a null congruence. Equation (18.11) is to be interpreted

as an equation governing the evolution of the extrinsic curvature K_{azb} along any path of the congruence, that is, along the z -direction. The evolution depends on the Riemann curvature R_{bzaa} encountered along the path.

The left hand side of equation (18.11) also depends on a derivative of the spatial acceleration K_{azz} . A necessary and sufficient condition for the congruence to be geodesic is that the spatial acceleration vanishes

$$K_{azz} = 0 . \quad (18.12)$$

For a geodesic congruence (not necessarily affinely parameterized), the Raychaudhuri equation (18.11) becomes

$$\hat{D}_z K_{azb} - K_{bz}^y K_{azy} + K_{zb}^c K_{azc} = -R_{bzaz} \quad (\text{no sum over } z) . \quad (18.13)$$

If the congruence is geodesic, then the tetrad can be chosen to be parallel-transported along each path of the congruence. In this case all the components of the tetrad-frame connection with final index z vanish

$$\Gamma_{klz} = 0 . \quad (18.14)$$

The conditions (18.14) exhaust all the 6 degrees of freedom of Lorentz transformations of the tetrad. In this case the restricted covariant derivative $\hat{D}_z$ in the Raychaudhuri equation (18.13) reduces to the directed derivative ∂_z , and the equation becomes

$$\partial_z K_{azb} + K_{zb}^c K_{azc} = -R_{bzaz} \quad (\text{no sum over } z) . \quad (18.15)$$

18.3 Raychaudhuri equations for a timelike geodesic congruence

For a congruence of timelike paths, the extrinsic curvature is the spatial tensor $K_{ab} \equiv K_{a0b} \equiv \Gamma_{a0b}$. If the timelike paths are geodesic, then the acceleration $K_a \equiv K_{a00}$ vanishes. Along a timelike geodesic congruence, the Raychaudhuri equations (18.13) become

$$\hat{D}_0 K_{ab} + K_{ab}^c K_{ac} = -R_{b0a0} . \quad (18.16)$$

In 4-dimensional spacetime, the 9 components of the extrinsic curvature K_{ab} are commonly resolved into an **expansion** scalar ϑ , a 3-component antisymmetric **vorticity** tensor ϖ_{ab} , and a 5-component traceless symmetric **shear** tensor σ_{ab} ,

$$K_{ab} = \delta_{ab}\vartheta + \varpi_{ab} + \sigma_{ab} . \quad (18.17)$$

Like the extrinsic curvature, the expansion, vorticity, and shear are restricted tensors, that is, tensors with respect to the restricted group of spatial Lorentz transformations. The trace of the extrinsic curvature is three times the expansion, $K \equiv K_a^a = 3\vartheta$. The vorticity is sometimes referred to alternatively as the **rotation**, or the **twist**. If desired, the vorticity can be written $\varpi_{ab} = \varepsilon_{abc}\varpi^c$.

If one imagines comoving coordinates attached to the congruence of paths, then the extrinsic curvature describes the rate at which the comoving volume element distorts, equation (18.6). The expansion ϑ equals one third the logarithmic rate of change of the volume of the comoving volume element, the vorticity is

the rate at which the comoving volume element rotates (see §18.7), and the shear is the rate at which the comoving volume element distorts tidally.

To see that the expansion measures the logarithmic rate of change of the volume, choose comoving coordinates consisting of the proper time τ along with 3 spatial coordinates x^α that remain constant along the geodesics of the congruence. The comoving coordinate 4-velocity along geodesics is $u^\mu = \{1, 0, 0, 0\}$. The inverse vierbein satisfies $e_0^\mu = u^\mu = \{1, 0, 0, 0\}$, so the determinant e of the full vierbein reduces to the determinant of its spatial part, $e \equiv |e^m_\mu| = |e^a_\alpha|$. The trace $K \equiv K_{0a}^a \equiv K_0$ equals the covariant divergence $D_m u^m$, equation (18.10). The expansion ϑ thus satisfies

$$3\vartheta = K = D_m u^m = D_\mu u^\mu = \frac{1}{\sqrt{g}} \frac{\partial(\sqrt{g}u^\mu)}{\partial x^\mu} = u^\mu \frac{\partial \ln \sqrt{g}}{\partial x^\mu} = \frac{d \ln \sqrt{g}}{d\tau}, \quad (18.18)$$

where $\sqrt{g} = e$ is the square root of the determinant of the spatial metric of the comoving line-element, which is the same as the determinant e of the vierbein.

In the ADM formalism, the tetrad time vector γ_0 is chosen to be orthogonal to hypersurfaces of constant time t . If γ_0 is so chosen, and if torsion vanishes as general relativity assumes, then vorticity ϖ_{ab} vanishes, as shown in §18.7, equation (18.42). This explains why in the ADM formalism the extrinsic curvature K_{ab} is symmetric in ab . The paths of an ADM congruence are vorticity-free, but not necessarily geodesic. They are geodesic if and only if the lapse α is constant, equation (18.42). In the ADM formalism, the expansion satisfies equation (17.60), which reduces to equation (18.18) if the lapse is unity and the shift vanishes, that is, if the spatial coordinates are comoving and the time coordinate t is the proper time τ .

Not all congruences are hypersurface-orthogonal, so vorticity does not vanish in general. For example, if a congruence is chosen to follow the worldlines of a system of dust particles (dust particles being neutral and collisionless, to ensure that they follow geodesics), then the vorticity, which is related to the angular momentum of the system of particles, will generically be non-zero.

The vorticity $\varpi_{ab} \equiv K_{[ab]}$, the antisymmetric part of the extrinsic curvature Γ_{a0b} , should be distinguished from the **precession** $\Gamma_{[ab]0}$ (if the tetrad metric γ_{ab} is constant, as here, then Γ_{ab0} is automatically antisymmetric in ab ; in the more general case where the tetrad metric is non-constant, as in ADM, §17.2.1, the precession equals the antisymmetric part of Γ_{ab0}). The condition for the tetrad frame to be locally inertial, that is, freely falling and non-rotating, is that the acceleration and precession vanish, $\Gamma_{a00} = \Gamma_{[ab]0} = 0$. By a suitable spatial rotation of the tetrad (which rotates the spatial axes γ_a while leaving the time axis γ_0 unchanged) the precession $\Gamma_{[ab]0}$ can be arranged to vanish along a congruence. Whereas the precession describes the spatial rotation of the tetrad frame with respect to locally inertial, the vorticity is related to the angular momentum of particles following the congruence. Since the extrinsic curvature is a spatial tensor, if the vorticity vanishes in one frame, then it vanishes in any spatially rotated frame; and conversely if the vorticity is non-vanishing in one frame, then it is non-vanishing in any spatially rotated frame.

The Raychaudhuri equations (18.16) for the expansion, vorticity, and shear along a timelike geodesic

congruence are

$$\hat{D}_0\vartheta + \vartheta^2 + \frac{1}{3}\sigma^{ab}\sigma_{ab} - \frac{1}{3}\varpi^{ab}\varpi_{ab} = -\frac{1}{3}R_{00} , \quad (18.19a)$$

$$(\hat{D}_0 + 2\vartheta)\varpi_{ab} + \sigma^c{}_a\varpi_{cb} - \sigma^c{}_b\varpi_{ca} = \frac{1}{2}(R_{a0b0} - R_{b0a0}) , \quad (18.19b)$$

$$(\hat{D}_0 + 2\vartheta)\sigma_{ab} + (\sigma^c{}_a\sigma_{cb} - \frac{1}{3}\delta_{ab}\sigma^{cd}\sigma_{cd}) - (\varpi^c{}_a\varpi_{cb} - \frac{1}{3}\delta_{ab}\varpi^{cd}\varpi_{cd}) = -C_{0a0b} , \quad (18.19c)$$

where C_{klmn} is the Weyl tensor, the traceless part of the Riemann tensor. If torsion vanishes as general relativity assumes, then the antisymmetric part of the Riemann tensor vanishes, equation (11.69), and the right hand side of the vorticity equation (18.19b) vanishes. The restricted derivatives in equations (18.19) are

$$\hat{D}_0\vartheta = \partial_0\vartheta , \quad (18.20a)$$

$$\hat{D}_0\varpi_{ab} = \partial_0\varpi_{ab} - \Gamma_{a0}^c\varpi_{cb} - \Gamma_{b0}^c\varpi_{ac} , \quad (18.20b)$$

$$\hat{D}_0\sigma_{ab} = \partial_0\sigma_{ab} - \Gamma_{a0}^c\sigma_{cb} - \Gamma_{b0}^c\sigma_{ac} . \quad (18.20c)$$

If the tetrad is chosen to be parallel-transported along the geodesic, then all 6 of the tetrad connections with final index 0 vanish,

$$\Gamma_{kl0} = 0 , \quad (18.21)$$

including not only the 3 components $K_a \equiv K_{a00}$ of the acceleration, but also the 3 components Γ_{ab0} of the precession. In this case, the restricted covariant time derivative simplifies to the directed time derivative, which is the same as the proper time derivative $d/d\tau$ in the parallel-transported frame,

$$\hat{D}_0 = \partial_0 = \frac{d}{d\tau} . \quad (18.22)$$

Exercise 18.1 Raychaudhuri equations for a non-geodesic timelike congruence. Derive the Raychaudhuri equations for a timelike congruence that is not geodesic.

Solution. The Raychaudhuri equations for a timelike congruence including non-vanishing acceleration K_a are

$$\hat{D}_0\vartheta + \vartheta^2 + \frac{1}{3}\sigma^{ab}\sigma_{ab} - \frac{1}{3}\varpi^{ab}\varpi_{ab} - \frac{1}{3}\hat{D}^a K_a - \frac{1}{3}K^a K_a = -\frac{1}{3}R_{00} , \quad (18.23a)$$

$$(\hat{D}_0 + 2\vartheta)\varpi_{ab} + \sigma^c{}_a\varpi_{cb} - \sigma^c{}_b\varpi_{ca} + \frac{1}{2}(\hat{D}_a K_b - \hat{D}_b K_a) = \frac{1}{2}(R_{a0b0} - R_{b0a0}) , \quad (18.23b)$$

$$\begin{aligned} (\hat{D}_0 + 2\vartheta)\sigma_{ab} + \sigma^c{}_a\sigma_{cb} - \varpi^c{}_a\varpi_{cb} - \frac{1}{2}(\hat{D}_a K_b + \hat{D}_b K_a) - K_a K_b \\ - \frac{1}{3}\delta_{ab}(\sigma^{cd}\sigma_{cd} - \varpi^{cd}\varpi_{cd} - \hat{D}^c K_c - K^c K_c) = -C_{0a0b} , \end{aligned} \quad (18.23c)$$

with the restricted covariant derivatives given by equations (18.20). If torsion vanishes as general relativity assumes, then the antisymmetric part of the Riemann tensor vanishes, and the right hand side of the vorticity equation (18.19b) vanishes. If the acceleration is the gradient of a potential, $K_a = \partial_a \ln \alpha$, and if torsion vanishes, then $\hat{D}_a K_b - \hat{D}_b K_a = 0$, and vorticity vanishes if it vanishes initially. This is the situation imposed

in the ADM formalism. If on the other hand the acceleration takes a more general form, then vorticity may be generated along the path.

18.4 Raychaudhuri equations for a null geodesic congruence

For a null congruence in 4-dimensional spacetime, it is convenient to work with a Newman-Penrose double-null tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$, with two null directions at each point, an “outgoing” direction γ_v , and an “ingoing” direction γ_u . The spin axes γ_+ and γ_- span the two-dimensional spatial plane orthogonal to the null directions. Late latin indices $z, y, \dots$ run over null indices v, u , early latin indices $a, b, \dots$ run over spin indices $+$, $-$, and mid latin indices $k, l, \dots$ run over all four indices.

The extrinsic curvature constitutes the components $K_{azk} \equiv \Gamma_{azk}$ of the tetrad-frame connections with first two indices az from opposite subspaces, equation (17.219). If the null congruence along the outgoing v -direction is geodesic, then the acceleration K_{avv} vanishes. Along outgoing null geodesics, the Raychaudhuri equations (18.13) are

$$\hat{D}_v K_{avb} + K_{vb}^c K_{avc} = -R_{bvav} . \quad (18.24)$$

The condition $K_{avv} = 0$ that the outgoing null directions of the congruence be geodesic fixes 2 of the 6 degrees of freedom of Lorentz transformations of the tetrad. Additional convenient gauge choices can be imposed. A common choice is to impose sufficient conditions that the restricted covariant derivative $\hat{D}_v$ in the Raychaudhuri equation (18.24) reduces to the directed derivative ∂_v . This requires that the null axis γ_v and the 2 spatial axes $\gamma_\pm$ (but not the null axis γ_u) of the tetrad be parallel-transported along the null geodesic congruence. Parallel-transport of γ_v and $\gamma_\pm$ amounts to imposing that 4 of the 6 tetrad connections vanish,

$$\Gamma_{uvv} = \Gamma_{+-v} = K_{+vv} = K_{-vv} = 0 . \quad (18.25)$$

The condition $\Gamma_{uvv} = 0$ is the condition that the geodesics along γ_v be affinely parameterized, while the condition $\Gamma_{+-v} = 0$ is the condition that the spatial axes $\gamma_\pm$ do not rotate in the parallel-transported frame. Under the conditions (18.25), the restricted covariant derivative in the Raychaudhuri equation (18.24) equals a derivative with respect to an affine parameter λ along the null geodesic,

$$\hat{D}_v = \partial_v = \gamma_v \cdot \partial = u \cdot \partial = \frac{dx^\mu}{d\lambda} \frac{\partial}{\partial x^\mu} = \frac{d}{d\lambda} . \quad (18.26)$$

Other gauge choices can be made. A natural choice is to choose the tetrad so that both outgoing and ingoing null directions are geodesic. For example, the principal null directions of an ideal black hole are geodesic (the tetrad that aligns with the principal null directions is the Boyer-Lindquist tetrad). The condition that the outgoing and ingoing null directions be geodesic translates into the condition that $K_{azz} = 0$, or explicitly the 4 conditions

$$K_{+vv} = K_{-vv} = K_{+uu} = K_{-uu} = 0 . \quad (18.27)$$

If the ingoing null direction is geodesic, then the Raychaudhuri equations along the ingoing null geodesic

are the same as equations (18.24) with null indices swapped, $v \leftrightarrow u$. By a suitable Lorentz boost in the $\gamma_v \gamma_u$ -plane, it is always possible to arrange that the tetrad frame is affinely parameterized in either the γ_v or the γ_u direction (that is, either Γ_{uvv} or $\Gamma_{vu u}$ vanishes), but in general it is not possible to arrange that both null directions are affinely parameterized. Similarly, by a suitable spatial rotation in the $\gamma_+ \gamma_-$ -plane, it is always possible to arrange that the spatial axes are parallel-transported along either the γ_v or the γ_u direction (that is, either Γ_{+-v} or Γ_{+-u} vanishes), but in general it is not possible to arrange that the spatial axes are parallel-transported along both null directions.

The Raychaudhuri equations (18.24) are equations governing the evolution of the extrinsic curvatures K_{avb} with middle index the null direction v , and outer indices ab spin indices. Analogously to the 3+1 decomposition (18.17), these 4 components are commonly decomposed into an **expansion** scalar ϑ , an antisymmetric **vorticity** tensor $\varpi_{ab} \equiv \varepsilon_{ab} \varpi$, and a traceless symmetric **shear** tensor σ_{ab} ,

$$K_{avb} = \gamma_{ab} \vartheta + \varepsilon_{ab} \varpi + \sigma_{ab} . \quad (18.28)$$

Like the extrinsic curvature, the expansion, vorticity, and shear are restricted tensors. As usual in the Newman-Penrose formalism, complex conjugation flips the spin indices on any tensor, $+\leftrightarrow-$, a consequence of the fact that the Newman-Penrose spin axes γ_+ and γ_- are complex conjugates of each other. The totally antisymmetric tensor ε_{ab} in 2-dimensional spin space flips sign under complex conjugation, so is purely imaginary, $\varepsilon_{+-} = i$. The expansion and vorticity scalars ϑ and ϖ are both real. The shear is complex, with two components that are complex conjugates of each other, $\sigma_{--} = \sigma_{++}^*$.

Just as the timelike expansion equals one third the logarithmic rate of change of the comoving volume element along a timelike congruence, equation (18.18), so also the null expansion equals one half the logarithmic rate of change of the comoving area element along a null congruence. First, notice that along an outgoing null congruence, the ingoing γ_u component of the tetrad-frame covariant divergence $D_m u^m$ vanishes, $D_u u^u = \partial_u u^u + \Gamma_{mu}^u u^m = -\Gamma_{vvu} = 0$ (no sum over u or v). Therefore the covariant divergence equals the tetrad-frame covariant divergence restricted to the 3-dimensional hypersurface spanned by the outgoing geodesic direction γ_v and the spatial directions $\gamma_{\pm}$. Such a 3-dimensional hypersurface can be constructed by starting with any spatial 2-surface and projecting “outgoing” null geodesics not necessarily orthogonally from it. Choose comoving coordinates along the null hypersurface consisting of the affine parameter λ along with 2 spatial coordinates x^α that remain constant along the geodesics of the congruence. The coordinate 3-velocity within the null hypersurface is $u^\mu \equiv dx^\mu/d\lambda = \{1, 0, 0\}$. Then analogously to equation (18.18) the null expansion satisfies, from equation (18.10) with $\Gamma_{vv}^v = 0$ because the congruence is being taken to be affinely parameterized,

$$2\vartheta = K_v = D_m u^m = D_\mu u^\mu = \frac{1}{\sqrt{g}} \frac{\partial(\sqrt{g} u^\mu)}{\partial x^\mu} = u^\mu \frac{\partial \ln \sqrt{g}}{\partial x^\mu} = \frac{d \ln \sqrt{g}}{d\lambda} , \quad (18.29)$$

where g is the determinant of 2-dimensional spatial metric of the comoving line-element. Thus the null expansion ϑ equals one half the logarithmic rate of change of the cross-sectional area of the comoving area element.

In terms of the expansion, vorticity, and shear, the Raychaudhuri equations (18.24) along the outgoing

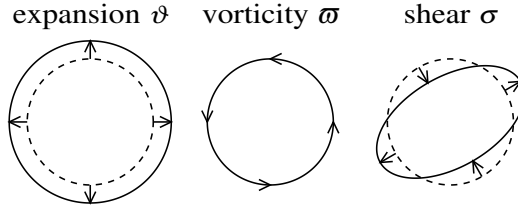


Figure 18.1 Illustrating how the Sachs optical coefficients, the expansion ϑ , the vorticity ϖ , and the shear σ , characterize the rate at which a congruence of light rays changes shape as it propagates. The congruence of light rays is coming vertically upward out of the paper.

null geodesic direction v are

$$(\hat{D}_v + \vartheta)\vartheta - \varpi^2 + \sigma_{++}\sigma_{++}^* = -4\pi T_{vv} , \quad (18.30a)$$

$$(\hat{D}_v + 2\vartheta)\varpi = 0 , \quad (18.30b)$$

$$(\hat{D}_v + 2\vartheta)\sigma_{++} = -C_{v+v+} . \quad (18.30c)$$

The restricted covariant derivatives in equations (18.30) are

$$\hat{D}_v\vartheta = (\partial_v + \Gamma_{uvv})\vartheta , \quad (18.31a)$$

$$\hat{D}_v\varpi = (\partial_v + \Gamma_{uvv})\varpi , \quad (18.31b)$$

$$\hat{D}_v\sigma_{++} = (\partial_v + \Gamma_{uvv} + 2\Gamma_{+-v})\sigma_{++} . \quad (18.31c)$$

18.5 Sachs optical coefficients

If the null axis γ_v and the two spatial axes $\gamma_{\pm}$ are taken to be parallel-transported along the null geodesic directions γ_v of the congruence, then the tetrad connections Γ_{uvv} and Γ_{-+v} in equations (18.30) vanish. In this case the expansion ϑ , vorticity ϖ , and the complex shear $\sigma \equiv \sigma_{++}$ are commonly called the **Sachs optical coefficients** (Sachs, 1961), often referred to as Sachs scalars. The Raychaudhuri equations (18.30) simplify to

$$(\partial_v + \vartheta)\vartheta - \varpi^2 + \sigma\sigma^* = -4\pi T_{vv} , \quad (18.32a)$$

$$(\partial_v + 2\vartheta)\varpi = 0 , \quad (18.32b)$$

$$(\partial_v + 2\vartheta)\sigma = -C_{v+v+} . \quad (18.32c)$$

The directed derivative ∂_v equals a derivative $d/d\lambda$ with respect to an affine parameter along the geodesic directions, equation (18.26).

The Sachs coefficients characterize how the shape of the congruence of light rays evolves as it propagates, as illustrated in Figure 18.1. The expansion represents how fast the congruence expands, the vorticity how

fast it rotates, and the shear how fast its ellipticity is changing. The amplitude and phase of the complex shear represent the amplitude and phase of the major axis of the shear ellipse.

Concept question 18.2 Can vorticity be non-zero while shear vanishes? Answer. Yes. The principal null congruences of the Λ -Kerr-Newman geometry provide an example of congruences that have non-zero vorticity but are shear-free, Exercise 23.12.

18.6 Interior mass

The outgoing and ingoing expansions θ_v and θ_u are multiplied by e^θ and $e^{-\theta}$ under a Lorentz boost in the vu -plane. Their product $\theta_v\theta_u$ is invariant under a Lorentz boost in the vu -plane.

$$2\theta^z\theta_z = 4\gamma^{vu}\theta_v\theta_u = 4\theta_v\theta_u = K_vK_u = K_{va}^aK_{ub}^b. \quad (18.33)$$

I think $\gamma_{vu} = 1$. equation (17.226g)

$$R_{abcd} = \hat{R}_{abcd} + K_{cvb}K_{dua} + K_{cub}K_{dva} - K_{cva}K_{dub} - K_{cua}K_{dvb} \quad (18.34)$$

$$R_{ac} = \hat{R}_{ac} + K_{cvb}K_{dua} + K_{cub}K_{dva} - K_{cva}K_{dub} - K_{cua}K_{dvb} \quad (18.35)$$

18.7 Hypersurface-orthogonality for a timelike congruence

Singularity theorems consider special congruences that are both geodesic and vorticity-free. The Raychaudhuri equation (18.19b) guarantees that if the vorticity ϖ_{ab} vanishes on the initial 3-dimensional hypersurface of a timelike geodesic congruence, then the vorticity will vanish identically everywhere along the congruence. This section shows that a timelike congruence is geodesic and vorticity-free if and only if it is **hypersurface-orthogonal**, that is, the 4-velocity $\mathbf{u}$ along the paths is normal to some hypersurface, equations (18.37), which proves to be a hypersurface of constant proper time, or equivalently of constant action. The next subsection, §18.7.1, shows how to construct a timelike hypersurface-orthogonal congruence.

The covariant curl of the 4-velocity $\mathbf{u} \equiv \gamma_0$ of a congruence of timelike paths is

$$\mathbf{D} \wedge \mathbf{u} = \gamma^m \wedge \gamma^n (\partial_m u_n - \Gamma_{nm}^k u_k) = \gamma^m \wedge \gamma^n \Gamma_{n0m} = \gamma^0 \wedge \gamma^a K_a - \gamma^a \wedge \gamma^b \varpi_{ab}. \quad (18.36)$$

The covariant curl is a 6-component bivector whose 3 time-space parts are the acceleration K_a , and whose 3 space-space parts are the vorticity ϖ_{ab} .

Equation (18.36) shows that the covariant curl $\mathbf{D} \wedge \mathbf{u}$ vanishes if and only if both the acceleration K_a and the vorticity ϖ_{ab} vanish. If the curl vanishes, and if torsion vanishes, then by Poincaré's lemma the 4-velocity $\mathbf{u}$ is, at least locally, the gradient of a scalar τ ,

$$\mathbf{D} \wedge \mathbf{u} = 0 \quad \Leftrightarrow \quad \mathbf{u} = -\partial\tau. \quad (18.37)$$

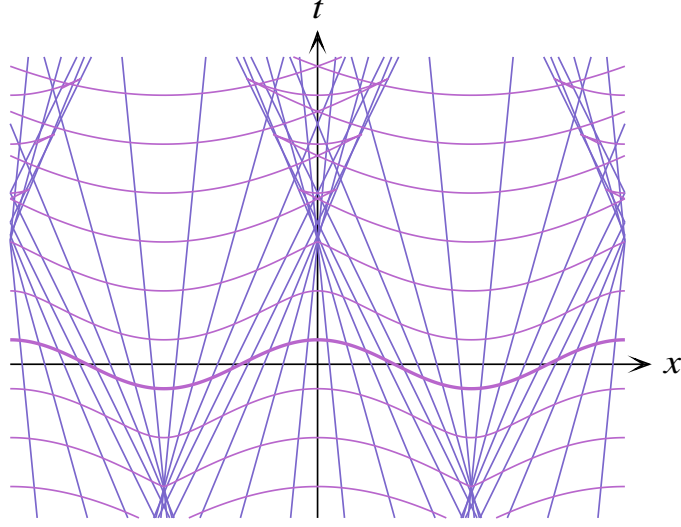


Figure 18.2 Spacetime diagram of Minkowski space illustrating a hypersurface-orthogonal congruence of timelike geodesics. The congruence is constructed by starting with an initial 3-dimensional spacelike hypersurface (thick line), here a cosine perturbation from the $t = 0$ hypersurface, and projecting geodesics (blue lines) along its timelike normal direction. Hypersurfaces of constant proper time τ (purple lines) to the past or future of the initial hypersurface remain orthogonal to the geodesics. Generically, as here, the geodesics cross, and the spatial hypersurfaces of constant proper time correspondingly develop caustics where the hypersurfaces fold and crease.

The scalar τ is just the proper time along the geodesics, as follows from

$$\mathbf{u} \cdot \mathbf{u} = -\mathbf{u} \cdot \partial\tau = -\frac{dx^\mu}{d\tau} \frac{\partial\tau}{\partial x^\mu} = -1 . \quad (18.38)$$

Thus the 4-velocity $\mathbf{u}$ is normal to 3-dimensional hypersurfaces of constant proper time τ .

The action S of a freely-falling particle of non-zero mass m is related to the proper time along the particle's worldline by, equation (4.7),

$$S = -m\tau . \quad (18.39)$$

Thus the hypersurfaces of a hypersurface-orthogonal timelike congruence are also hypersurfaces of constant action for massive, freely-falling particles. The covariant momentum $p_\mu = mu_\mu$ of the particle is the gradient of the action, equation (4.105),

$$p_\mu = \frac{\partial S}{\partial x^\mu} , \quad (18.40)$$

which reproduces the result $\mathbf{u} = -\partial\tau$.

A weaker condition than the vanishing of $\mathbf{D} \wedge \mathbf{u}$ is that the curl $\mathbf{D} \wedge (\mathbf{u}/\alpha)$ of the 4-velocity scaled by some



Figure 18.3 This deep image of the elliptical galaxy NGC 474 shows shells caused by caustics in collisionless streams of stars originating from small galaxies accreted by NGC 474 over the last billion years. Astronomy Picture of the Day, 2011 July 26. Image credit: P.-A. Duc (CEA, CFHT), Atlas 3D Collaboration.

arbitrary factor α vanishes. The covariant curl of the scaled 4-velocity $\mathbf{u}/\alpha$ is

$$\alpha \mathbf{D} \wedge (\mathbf{u}/\alpha) = \mathbf{D} \wedge \mathbf{u} + \mathbf{u} \wedge \partial \ln \alpha = \boldsymbol{\gamma}^0 \wedge \boldsymbol{\gamma}^a (K_a - \partial_a \ln \alpha) - \boldsymbol{\gamma}^a \wedge \boldsymbol{\gamma}^b \varpi_{ab} . \quad (18.41)$$

This curl of the scaled 4-velocity vanishes if and only if the acceleration K_a is the gradient of a scalar, and the vorticity ϖ_{ab} vanishes,

$$K_a = \partial_a \ln \alpha , \quad \varpi_{ab} = 0 . \quad (18.42)$$

The conditions (18.42) are precisely those established in the ADM formalism, with α being the lapse. If conditions (18.42) hold, then $\mathbf{D} \wedge (\mathbf{u}/\alpha)$ vanishes, and if torsion also vanishes, then by Poincaré's lemma $\mathbf{u}/\alpha$ is, at least locally, the gradient of a scalar t ,

$$\mathbf{D} \wedge (\mathbf{u}/\alpha) = 0 \quad \Leftrightarrow \quad \mathbf{u} = -\alpha \partial t . \quad (18.43)$$

The scalar coordinate t is just the ADM time coordinate, as follows from

$$\mathbf{u} = \boldsymbol{\gamma}_0 = -\boldsymbol{\gamma}^0 = -e^0_{\mu} \mathbf{e}^{\mu} = -\alpha \mathbf{e}^t = -\alpha \mathbf{e}^{\mu} \frac{\partial t}{\partial x^{\mu}} = -\alpha \partial t . \quad (18.44)$$

18.7.1 Construction of timelike, geodesic, hypersurface-orthogonal congruences

It is straightforward to construct a timelike, geodesic, hypersurface-orthogonal congruence by starting with any 3-dimensional spacelike hypersurface and projecting geodesics into the past and future along the normal to the spacelike hypersurface, as illustrated in Figure 18.2. The geodesics are orthogonal to hypersurfaces of constant proper time τ , or equivalently of constant action $S = -m\tau$, starting at $\tau = 0$ (or $S = 0$) on the initial spacelike hypersurface. Generically, the resulting geodesics will cross at some point in the past or future or both, and the hypersurface correspondingly develops caustics, as in Figure 18.2. Geodesics remain orthogonal to hypersurfaces of constant proper time τ even after they cross, but the proper time τ is multiply-valued at spacetime points crossed by multiple geodesics.

Caustics in collisionless streams of stars are often observed in deep images of elliptical galaxies, as illustrated in Figure 18.3. When galaxies collide, the gravitational potentials of the galaxies merge, but because galaxies are mostly empty space, the stars in the galaxies do not collide. When a small galaxy with a small velocity dispersion in its stars falls into a larger galaxy, the smaller galaxy is tidally disrupted by the larger galaxy, but the stars from the smaller galaxy continue to orbit the larger galaxy in coherent collisionless streams, forming caustics where the star streams turn around in the merged gravitational potential.

18.8 Hypersurface-orthogonality for a null congruence

For massive particles, the proper time τ , or equivalently the action $S = -m\tau = -m^2\lambda$, where λ is the affine parameter, progresses along geodesics, and momenta along geodesics are orthogonal to hypersurfaces of constant action, equation (18.40). For massless particles on the other hand, the action does not progress along null geodesics. For a null congruence, it is not possible to start from an initial 3-dimensional hypersurface over which the action vanishes, and to project null geodesics into the past and future from this initial hypersurface, because the failure of the action to progress along null geodesics would then imply that the action would vanish everywhere, and the spacetime would cease to be foliated into hypersurfaces of constant action to which geodesics were putatively orthogonal.

Rather, the action must be allowed to vary along the initial 3-dimensional hypersurface of a null congruence. The action on the initial 3-dimensional hypersurface foliates it into 2-dimensional spatial surfaces of constant action. At each point on each 2-dimensional surface there are exactly 2 null directions orthogonal to the spatial 2-surface, one “outgoing,” the other “ingoing.” Projecting null geodesics along these null directions defines a pair of 3-dimensional null hypersurfaces along which the action is constant. The result is a spacetime that is foliated into pairs of outgoing and ingoing 3-dimensional null hypersurfaces of constant outgoing (+) and ingoing (−) action $S_{\pm}$. The values of the actions are determined by their values on the initial non-null 3-dimensional hypersurface.

Null congruences constructed in this way are said to be hypersurface-orthogonal. This definition of hypersurface-orthogonality for null congruences does *not* require that equation (18.40) holds across all of the 4-dimensional spacetime. Rather, hypersurface-orthogonality for null congruences imposes that equa-

tion (18.40) holds in the massless limit along each 3-dimensional null hypersurface of constant action,

$$p_\mu = \lim_{m \rightarrow 0} m^2 \frac{\partial \lambda}{\partial x^\mu} . \quad (18.45)$$

To see why the definition of hypersurface-orthogonality for null congruences does not impose that the condition (18.40) hold over the entire 4-dimensional spacetime, suppose contrarily that it did. The 4-momentum along an outgoing null geodesic of the congruence satisfies $\mathbf{p} = p^v \boldsymbol{\gamma}_v = p_u \boldsymbol{\gamma}^u$ (no sum over v or u). Poincaré's lemma implies that equation (18.40) holds, at least locally, if and only if the covariant curl of the 4-momentum vanishes, $\mathbf{D} \wedge \mathbf{p} = 0$. The covariant curl of the 4-momentum is, similarly to equation (18.36),

$$\mathbf{D} \wedge \mathbf{p} = \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n (\partial_m p_n - \Gamma_{nm}^k p_k) = -p^v \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n \Gamma_{vnm} , \quad (18.46)$$

which vanishes if and only if $\Gamma_{v[nm]} = 0$. This is a set of 6 conditions on the tetrad connections, requiring not only that the 2 spatial components of the acceleration K_{avv} and the 1 component of vorticity $K_{v[-+]} \equiv \varpi_{+-} \equiv \varepsilon_{+-} \varpi$ vanish, but also that the 1 component of acceleration Γ_{uvv} along the null direction $\boldsymbol{\gamma}_v$ and the 2 components $\Gamma_{v[au]}$ vanish. While the 6 Lorentz gauge freedoms allow these 6 tetrad-frame connections to be chosen to vanish along the outgoing congruence, the corresponding 6 connections along the ingoing congruence cannot be made to vanish at the same time. Moreover the Raychaudhuri equations (18.30) have no dependence on the 2 components $\Gamma_{v[au]}$, and the vorticity equation (18.30b) allows vorticity to vanish without requiring that Γ_{uvv} vanishes.

Thus hypersurface-orthogonality for null congruences is conventionally defined by the weaker condition that the limiting equation (18.45) hold along each 3-dimensional null hypersurface. This requires that only the components $\mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p})$ of the covariant curl tangent to each 3-dimensional null hypersurface vanish, not that the covariant curl vanish identically throughout spacetime. The components of the covariant curl restricted to the null hypersurface are

$$\mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) = -(p^v)^2 \boldsymbol{\gamma}^u \wedge (\boldsymbol{\gamma}^v \wedge \boldsymbol{\gamma}^a K_{avv} - \boldsymbol{\gamma}^a \wedge \boldsymbol{\gamma}^b \varpi_{ab}) . \quad (18.47)$$

The covariant curl in parentheses in equation (18.47) is a 3-component bivector whose time-space part is proportional to the spatial acceleration K_{avv} , and whose space-space part is proportional to the vorticity $\varpi_{ab} \equiv \varepsilon_{ab} \varpi$. Unlike the timelike case, equation (18.41), the hypersurface-orthogonality condition (18.47) for null congruences is unchanged by scaling the momentum $\mathbf{p}$ by some arbitrary factor α , since

$$\alpha \mathbf{p} \wedge (\mathbf{D} \wedge (\mathbf{p}/\alpha)) = \mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) + \mathbf{p} \wedge \mathbf{p} \wedge \partial \ln \alpha = \mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) , \quad (18.48)$$

because $\mathbf{p} \wedge \mathbf{p} = 0$.

The Raychaudhuri equation (18.30b) for the vorticity ϖ along an outgoing geodesic of a null congruence implies that if the vorticity vanishes on the initial 2-dimensional spatial hypersurface spanned by $\boldsymbol{\gamma}_\pm$, then it is guaranteed to vanish thereafter. Thus a null geodesic congruence that is initially hypersurface-orthogonal will remain hypersurface-orthogonal thereafter. Note that equation (18.30b) allows the vorticity to vanish identically without imposing that the geodesic be affinely parameterized, that is, without imposing that Γ_{uvv} vanishes.

Hypersurface-orthogonality along the outgoing null congruence imposes only 3 conditions on the tetrad,

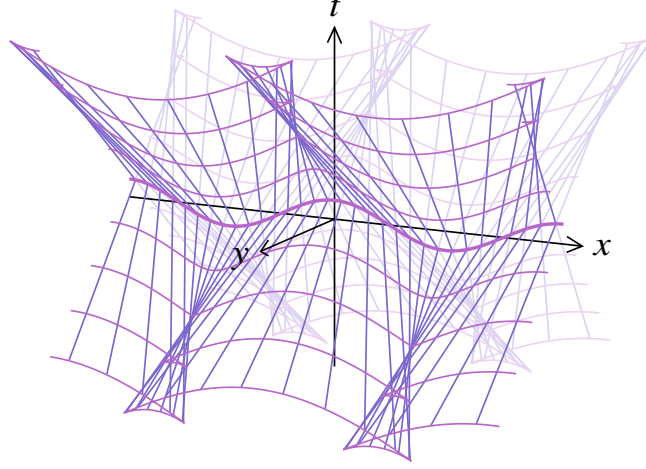


Figure 18.4 3D spacetime diagram of Minkowski space illustrating a pair of hypersurface-orthogonal congruences of null geodesics (blue lines) emerging from a 2-dimensional spacelike surface (thick line). The spacelike curves (purple lines) on the two null hypersurfaces are lines of constant affine parameter λ . These lines of constant affine parameter trace the intersections of null hypersurfaces in the remaining spacetime provided that the congruences are constructed to have translation symmetry in the y -direction (and in the suppressed z -direction), in which case other null hypersurfaces are parallel to the two shown, translated in the y -direction. This Figure would look the same as Figure 18.2 if projected on to the tx -plane.

namely that the outgoing spatial acceleration K_{avv} and the outgoing vorticity $\varpi_{+-} \equiv K_{v[-+]}$ vanish. The 6 Lorentz gauge freedoms allow hypersurface-orthogonality to be imposed simultaneously along both outgoing and ingoing null congruences, by demanding that the spatial accelerations K_{azz} and the vorticities $K_{z[+-]}$ along both congruences vanish.

18.8.1 Construction of double-null, geodesic, hypersurface-orthogonal congruences

To construct hypersurface-orthogonal congruences of outgoing and ingoing null geodesics, start with any non-null (timelike or spacelike) 3-dimensional hypersurface. Foliate the hypersurface into 2-dimensional spatial surfaces labelled by a time coordinate or a spatial coordinate according to whether the parent 3-hypersurface is timelike or spacelike. Project null geodesics along the two null directions normal to each spatial 2-surface. The null geodesics projecting from each 2-surface form a pair of 3-dimensional null hypersurfaces, as illustrated by Figure 18.4. Each null hypersurface is labelled by a constant null coordinate whose value is set by the value of the time or spatial coordinate on the 2-surface. The two geodesic null directions at each point define the null directions γ_v and γ_u of a Newman-Penrose tetrad. The spatial directions orthogonal to the two null directions define a plane whose tangent directions form the spatial directions γ_+ and γ_- of the Newman-Penrose tetrad.

Again it should be emphasized that hypersurface-orthogonality for null congruences is defined not by

condition (18.40) imposed over all spacetime, but rather by the limiting condition (18.45) imposed over each of the 3-dimensional null hypersurfaces of the congruence.

18.9 Focusing theorems

Focusing theorems exist for both timelike and null congruences. The focusing theorem follows from the Raychaudhuri equation for the expansion ϑ , coupled with assumptions about the sources in that equation.

The assumptions are:

1. the congruence is hypersurface-orthogonal;
2. the expansion is negative at some point, $\vartheta < 0$;
3. the energy-momentum tensor satisfies a positivity condition.

As shown in §§18.7 and 18.8, a hypersurface-orthogonal timelike or null congruence can be constructed by starting from some arbitrary (spacelike, for a timelike congruence, or non-null, for a null congruence) initial 3-dimensional hypersurface and projecting geodesics orthogonally from it. The requirement that the expansion be negative at some point is the reason that singularity theorems posit that a trapped surface has formed. A trapped surface is defined to be a closed 2-dimensional surface from which the expansions along both outgoing and ingoing orthogonal null directions are negative everywhere along the surface. Trapped surfaces exist inside the outer horizon of an ideal black hole, and it is plausible that the formation of a trapped surface is characteristic of the formation of a black hole. The final condition, a positivity condition on the energy-momentum tensor, ensures that the energy-momentum source in the Raychaudhuri equation is positive.

18.9.1 Focusing theorem for a null geodesic congruence

If the vorticity ϖ vanishes, then in the frame parallel-transported along the null congruence, the Raychaudhuri equation (18.32a) for the expansion ϑ along a null congruence simplifies to

$$\frac{d\vartheta}{d\lambda} + \vartheta^2 + \sigma\sigma^* + \frac{1}{2}G_{vv} = 0 . \quad (18.49)$$

The terms ϑ^2 and $\sigma\sigma^*$ are necessarily positive. The Newman-Penrose component G_{vv} of the Einstein tensor is related to the components in the parent orthonormal tetrad by

$$G_{vv} = \frac{1}{2}G_{00} + G_{03} + \frac{1}{2}G_{33} . \quad (18.50)$$

The Einstein component G_{vv} has boost weight 2, and is therefore multiplied by $e^{2\theta}$ under a boost by rapidity θ in the 3-direction. Consequently positivity of G_{vv} in one frame implies positivity of G_{vv} in any frame boosted in the 3-direction. Boosted along the 3-direction into the centre-of-mass frame, where $G_{03} = 0$, equation (18.50) reduces to

$$G_{vv} = \frac{1}{2}(G_{00} + G_{33}) = 4\pi(\rho + p_3) , \quad (18.51)$$

where ρ is the energy density and p_3 the pressure along the 3-direction. The Einstein component G_{vv} is therefore positive provided that

$$\boxed{\rho + p_3 \geq 0} , \quad (18.52)$$

which is called the **null energy condition**. If the null energy condition (18.52) holds, then the vorticity-free Raychaudhuri equation (18.49) shows that the expansion ϑ must always decrease.

The Raychaudhuri equation (18.49) can be arranged as

$$\frac{d(1/\vartheta)}{d\lambda} = 1 + \frac{\sigma\sigma^* + \frac{1}{2}G_{vv}}{\vartheta^2} , \quad (18.53)$$

whose right hand side is greater than or equal to 1, given the null energy condition (18.52). If the expansion ϑ is negative (meaning that light rays are converging), then equation (18.53) shows that $1/\vartheta$ will reach 0 at a finite value of the affine parameter λ . In other words, ϑ must become negative infinite at some finite value of λ . A negative infinite value of the expansion means, from equation (18.29), that the cross-sectional area of the null congruence has shrunk to zero. A zero cross-sectional area of a null congruence means that the congruence has shrunk to a line, not necessarily to a point.

Shrinking to zero area does not mean that a singularity has formed; it means simply that geodesics have reached a crossing point. For example, Figure 18.4 shows crossing geodesics of a null congruence in Minkowski space. It is only when *all* geodesics from a hypersurface-orthogonal congruence reach a crossing point that the spacetime encounters difficulties. In Figure 18.4, while the expansion is negative along some null geodesics, it is positive along others.

18.9.2 Focusing theorem for timelike geodesic congruence

The proof of the focusing theorem for timelike geodesics is similar to that for null geodesics. For vanishing vorticity, the Raychaudhuri equation (18.19a) along a timelike geodesic congruence is

$$\frac{d\vartheta}{d\tau} + \vartheta^2 + \frac{1}{3}\sigma^{ab}\sigma_{ab} + \frac{1}{3}R_{00} = 0 \quad (18.54)$$

in the orthonormal tetrad frame freely-falling along the geodesic. The component R_{00} of the Ricci tensor in the orthonormal tetrad is

$$R_{00} = 4\pi(\rho + 3p) , \quad (18.55)$$

where ρ is the energy density and $p \equiv \frac{1}{3}p_a^a$ is the isotropic pressure. The Ricci component R_{00} is positive provided that

$$\boxed{\rho + 3p \geq 0} , \quad (18.56)$$

which is called the **strong energy condition**. Note that a cosmological constant violates the strong energy condition (18.56), but not the null energy condition (18.52).

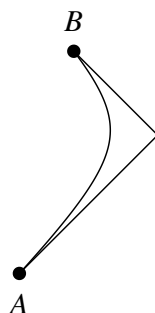


Figure 18.5 Spacetime diagram illustrating the dog-leg proposition. The dog-leg proposition asserts that any dog-leg path that joins 2 events A and B by a consecutive pair of null or timelike geodesics can be deformed into a strictly timelike path of longer proper time between A and B . The proposition is an assertion about the global causal structure of spacetime.

The timelike Raychaudhuri equation (18.54) can be written, similarly to the null Raychaudhuri equation (18.53),

$$\frac{d(1/\vartheta)}{d\lambda} = 1 + \frac{\frac{1}{3}\sigma^{ab}\sigma_{ab} + \frac{1}{3}R_{00}}{\vartheta^2}, \quad (18.57)$$

whose right hand side is greater than or equal to 1, given the strong energy condition (18.56). If the expansion ϑ is negative (meaning that volume is decreasing), then equation (18.57) shows that $1/\vartheta$ will reach 0 at a finite value of the affine parameter λ . In other words, ϑ must become negative infinite at some finite value of λ . A negative infinite value of the expansion means, from equation (18.18), that the volume of the congruence has shrunk to zero. A zero volume means that the congruence has flattened to a pancake, not necessarily to a point.

18.10 Singularity theorems

This section gives an account of one version of the singularity theorems, the original null version proved by Penrose (1965). See Senovilla (1998) for a review of singularity theorems.

18.10.1 Dog-leg proposition

A building block of singularity theorems is the dog-leg proposition. The dog-leg proposition asserts that any dog-leg path between two events A and B that consists of two different timelike or null geodesics joined together can be deformed into a strictly timelike path of longer proper time between A and B , as illustrated in Figure 18.5. The dog-leg proposition is a statement about the global causal structure of spacetime. The dog-leg proposition does not hold inside the inner horizon of a Kerr-Newman black hole, Concept question 18.3.

The dog-leg proposition can be replaced by other plausible hypotheses. Much of the content of the book

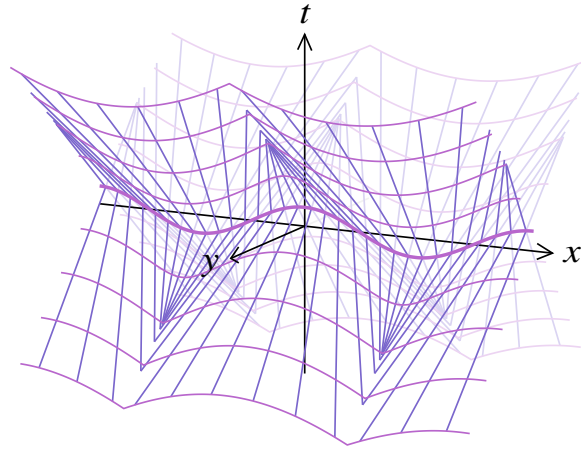


Figure 18.6 Null boundary of the future of a 2-dimensional spacelike surface. The null boundary is a pair of 3-dimensional null surfaces projecting orthogonally from the 2-surface (thick line), with the parts of the hypersurfaces excised after geodesic crossing, since the latter parts are connected by timelike geodesics to the 2-surface and are therefore not part of the null boundary. This is the same as Figure 18.4, but with geodesics terminated where they cross.

by Hawking and Ellis (1973) is concerned with exploring different plausible causality conditions. However, that will not be done here.

18.10.2 Null singularity theorem

Start with any 2-dimensional spatial surface. The future of this 2-surface is the 4-dimensional region of spacetime comprising all events that can be reached by some non-spacelike future-pointing path that starts at some point on the 2-surface. In a local neighbourhood of the 2-surface, the boundary of the future of the 2-surface comprises the pair of 3-dimensional null hypersurfaces projected orthogonally from the 2-surface, as illustrated by Figure 18.4. The dog-leg proposition implies that the future boundary is formed only from orthogonally-projected null geodesics. However, orthogonally-projected geodesics can intersect, as in Figure 18.4. After two orthogonally-projected geodesics intersect, a point to the future of the intersection can be reached from the 2-surface by starting on one geodesic and switching to the other at the crossing point. This dog-leg null path can be deformed into a timelike curve, and is therefore also not part of the future null boundary. Therefore the boundary of the future of the 2-surface comprises the pair of 3-dimensional null hypersurfaces projected orthogonally from the 2-surface, truncated where the null geodesics cross, as illustrated by Figure 18.6.

Now assume that the 2-dimensional surface is a trapped surface, meaning that the expansion along both the outgoing and ingoing null geodesic directions projected orthogonally from the 2-surface is negative at every point of the 2-surface. The focusing theorem implies that the expansion along every such null geodesic

reaches negative infinity at a finite value of the affine parameter λ , indicating that neighbouring null geodesics are crossing. Points on a null geodesic to the future of a crossing are no longer on the boundary of the future. Therefore the 3-dimensional boundary of the future of the trapped surface terminates after a finite affine parameter at a 2-dimensional caustic boundary. This is a contradiction, since the boundary of a boundary of a manifold is empty. Therefore the future must terminate, as it does for example inside the horizon of the Schwarzschild geometry.

Concept question 18.3 How do singularity theorems apply to the Kerr geometry? **Answer.**

The Kerr geometry violates the dog-leg proposition, so for this geometry the future does not terminate, but rather continues beyond the region where any trapped surface reaches a caustic boundary (see §23.25.1). As found in Exercise 23.2, the only geodesics that reach the ring singularity (Singularity or Parallel Singularity) of a Kerr black hole with $a \neq 0$ are null geodesics that lie in the equatorial plane. Therefore, to reach the singularity from a non-equatorial point, it is necessary to follow a geodesic down to the equatorial plane and then dog-leg to the singularity. Such a path cannot be deformed to a timelike geodesic. Similarly, a geodesic that starts at the singularity is confined to the equatorial plane, and a dog-leg is required to get out of the plane. The region that can be reached from the singularity by a dog-legged geodesic is the region inside the inner horizon. The ingoing and outgoing inner horizons of a Kerr black hole form the boundary of predictability, also known as the Cauchy horizon. A similar argument applies to the Kerr-Newman geometry, except that geodesics that hit the singularity must not only be null and equatorial, but also on one of the ingoing or outgoing principal null congruences, Exercise 23.2.

Concept question 18.4 How do singularity theorems apply to the Reissner-Nordström geometry? In Reissner-Nordström, the only geodesics that hit the singularity are radial null geodesics, Exercise 23.2. The Reissner-Nordström violates the dog-leg proposition because a dog-leg path that connects to the singularity cannot be deformed into a strictly timelike path: any path that connects to the singularity must be null asymptotically near the singularity.

Concept Questions

1. Explain how the equation for the Gullstrand-Painlevé metric (19.22) encodes not merely a metric but a full vierbein.
2. In what sense does the Gullstrand-Painlevé metric (19.22) depict a flow of space? [Are the coordinates moving? If not, then what is moving?]
3. If space has no substance, what does it mean that space falls into a black hole?
4. Would there be any gravitational field in a spacetime where space fell at constant velocity instead of accelerating?
5. In spherically symmetric spacetimes, what is the most important Einstein equation, the one that causes Reissner-Nordström black holes to be repulsive in their interiors, and causes mass inflation in non-empty (non Reissner-Nordström) charged black holes?

What's important?

1. The tetrad formalism provides a firm mathematical foundation for the concept that space falls faster than light inside a black hole.
2. Whereas the Kerr-Newman geometry of an ideal rotating black hole contains inside its horizon wormhole and white hole connections to other universes, real black holes are subject to the mass inflation stability discovered by Eric Poisson & Werner Israel (Poisson and Israel, [1990](#)).

Black hole waterfalls

19.1 Tetrads move through coordinates

As already discussed in §11.3, the way in which metrics are commonly written, as a (weighted) sum of squares of differentials,

$$ds^2 = \gamma_{mn} e^m{}_\mu e^n{}_\nu dx^\mu dx^\nu, \quad (19.1)$$

encodes not only a metric $g_{\mu\nu} = \gamma_{mn} e^m{}_\mu e^n{}_\nu$, but also a vierbein $e^m{}_\mu$, and consequently an inverse vierbein $e_m{}^\mu$, and associated tetrad γ_m . Most commonly the tetrad metric is orthonormal (Minkowski), $\gamma_{mn} = \eta_{mn}$, but other tetrad metrics, such as Newman-Penrose, occur. Usually it is self-evident from the form of the line-element what the tetrad metric γ_{mn} is in any particular case.

If the tetrad is orthonormal, $\gamma_{mn} = \eta_{mn}$, then the 4-velocity u^m of an object at rest in the tetrad, or equivalently the 4-velocity of the tetrad rest frame itself, is

$$u^m = \{1, 0, 0, 0\}. \quad (19.2)$$

The tetrad-frame 4-velocity (19.2) of the tetrad rest frame is transformed to a coordinate-frame 4-velocity u^μ in the usual way, by applying the inverse vierbein,

$$\frac{dx^\mu}{d\tau} \equiv u^\mu = e_m{}^\mu u^m = e_0{}^\mu. \quad (19.3)$$

Equation (19.3) says that the tetrad rest frame moves through the coordinates at coordinate 4-velocity given by the zeroth row of the inverse vierbein, $dx^\mu/d\tau = e_0{}^\mu$. The coordinate 4-velocity u^μ is related to the lapse α and shift β^α in the ADM formalism by $u^\mu = \{1, \beta^\alpha\}/\alpha$, equation (17.11).

The idea that locally inertial frames move through the coordinates provides the simplest way to conceptualize black holes. The motion of locally inertial frames through coordinates is what is meant by the “dragging of inertial frames” around rotating masses.

Exercise 19.1 Tetrad frame of a rotating wheel. Derive the line-element of Minkowski space adapted to the tetrad frame of a wheel uniformly rotating at angular velocity ω . Show that a clock attached to the wheel ticks slow by the Lorentz factor γ compared to a clock in the nonrotating frame, and that rulers

attached to the wheel measure the rim to be Lorentz-contracted by a factor γ compared to the nonrotating frame.

Solution. Start with the line-element of Minkowski space in cylindrical coordinates $x^\mu \equiv \{t, r, \phi, z\}$,

$$ds^2 = -dt^2 + dr^2 + r^2 d\phi^2 + dz^2. \quad (19.4)$$

The vierbein for the line-element (19.4) is $e^m_\mu = \text{diag}(1, 1, r, 1)$, and the corresponding inverse vierbein is $e_m^\mu = \text{diag}(1, 1, 1/r, 1)$. Lorentz boost the inverse vierbein into the tetrad frame of the wheel rotating at velocity $v = r\omega$ in the azimuthal ϕ direction,

$$e_m^\mu = \begin{pmatrix} \gamma & 0 & \gamma r\omega & 0 \\ 0 & 1 & 0 & 0 \\ \gamma r\omega & 0 & \gamma & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \gamma & 0 & \gamma\omega & 0 \\ 0 & 1 & 0 & 0 \\ \gamma r\omega & 0 & \gamma/r & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (19.5)$$

The coordinate-frame 4-velocity of the wheel's tetrad frame through the coordinates is

$$\frac{dx^\mu}{d\tau} \equiv u^\mu = e_0^\mu = \{\gamma, 0, \gamma\omega, 0\}, \quad (19.6)$$

confirming that indeed the wheel is moving at $d\phi/dt = \omega$. The line-element is

$$ds^2 = -\gamma^2(dt - r^2\omega d\phi)^2 + dr^2 + \gamma^2 r^2(d\phi - \omega dt)^2 + dz^2. \quad (19.7)$$

A point on the wheel follows $dr = d\phi - \omega dt = dz = 0$, so its proper time satisfies

$$d\tau = \gamma(dt - r^2\omega d\phi) = \gamma(1 - r^2\omega^2)dt = \frac{dt}{\gamma}, \quad (19.8)$$

demonstrating that a clock on the wheel runs slow by γ as claimed. Rulers attached to the rim of the wheel measure distances that are simultaneous in the frame of the wheel, corresponding to $dt - r^2\omega d\phi = 0$. Thus corotating rulers measure azimuthal distances along the rim of

$$dl = \gamma r(d\phi - \omega dt) = \gamma r(1 - r^2\omega^2)d\phi = \frac{r d\phi}{\gamma}, \quad (19.9)$$

demonstrating that the rim is Lorentz-contracted by γ as claimed.

19.2 Gullstrand-Painlevé waterfall

The Gullstrand-Painlevé metric is a version of the metric for a spherical (Schwarzschild or Reissner-Nordström) black hole discovered in 1921 independently by Allvar Gullstrand (Gullstrand, 1922) and Paul Painlevé (Painlevé, 1921). Although Gullstrand's paper was published in 1922, after Painlevé's, it appears that Gullstrand's work has priority. Gullstrand's paper was dated 25 May 1921, whereas Painlevé's is a write up of a presentation to the Académie des Sciences in Paris on 24 October 1921. Moreover, Gullstrand seems to have had a better grasp of what he had discovered than Painlevé, for Gullstrand recognized that observables such

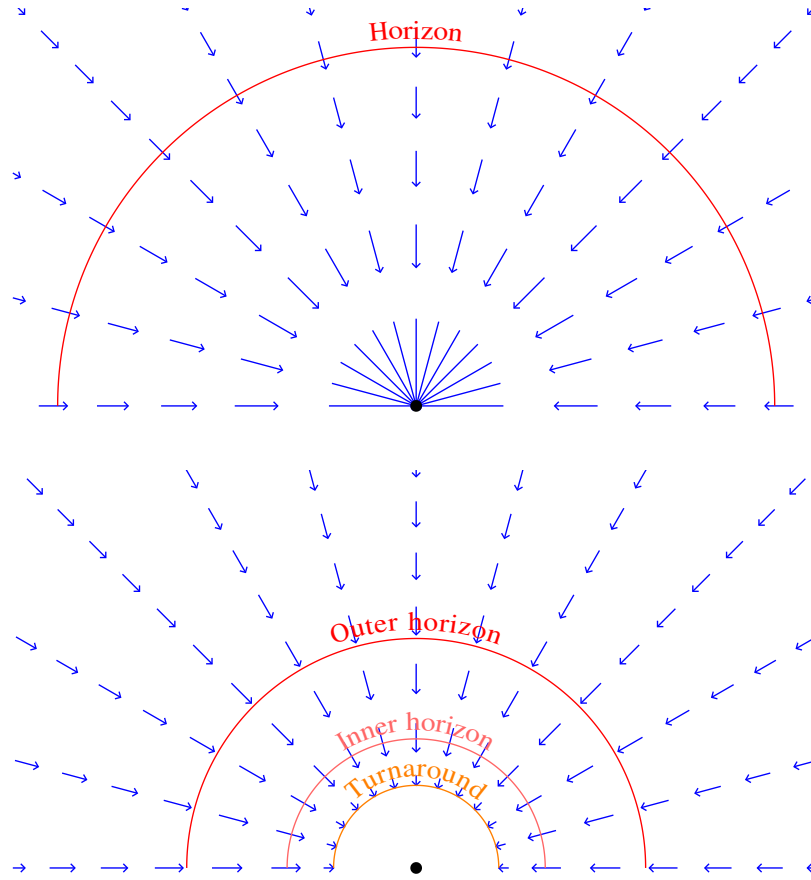


Figure 19.1 Radial velocity β in (upper panel) a Schwarzschild black hole, and (lower panel) a Reissner-Nordström black hole with electric charge $Q = 0.96$.

as the redshift of light from the Sun are unaffected by the choice of coordinates in the Schwarzschild geometry, whereas Painlevé, noting that the spatial metric was flat at constant free-fall time, $dt_{\text{ff}} = 0$, concluded in his final sentence that, as regards the redshift of light and such, “c’est pure imagination de prétendre tirer du ds^2 des conséquences de cette nature.”

Although neither Gullstrand nor Painlevé understood it, their metric paints a picture of space falling like a river, or waterfall, into a spherical black hole, Figure 6.1. The river has two key features: first, the river flows in Galilean fashion through a flat Galilean background, equation (19.25); and second, as a freely falling fishy swims through the river, its 4-velocity, or more generally any 4-vector attached to it, evolves by a series of infinitesimal Lorentz boosts induced by the change in the velocity of the river from place to place,

equation (19.30). Because the river moves in Galilean fashion, it can, and inside the horizon does, move faster than light through the background coordinates. However, objects moving in the river move according to the rules of special relativity, and so cannot move faster than light through the river.

19.2.1 Gullstrand-Painlevé tetrad

The Gullstrand-Painlevé metric (7.27) is

$$ds^2 = -dt_{\text{ff}}^2 + (dr - \beta dt_{\text{ff}})^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) , \quad (19.10)$$

where β is defined to be the radial velocity of a person who free-falls radially from rest at infinity,

$$\beta = \frac{dr}{d\tau} = \frac{dr}{dt_{\text{ff}}} , \quad (19.11)$$

and t_{ff} is the free-fall time, the proper time experienced by a person who free-falls from rest at infinity. The radial velocity β is the (apparently) Newtonian escape velocity

$$\beta = \mp \sqrt{\frac{2M(r)}{r}} , \quad (19.12)$$

where $M(r)$ is the interior mass within radius r , and the sign is $-$ (infalling) for a black hole, $+$ (outfalling) for a white hole. For the Schwarzschild or Reissner-Nordström geometry the interior mass $M(r)$ is the mass M at infinity minus the mass $Q^2/2r$ in the electric field outside r ,

$$M(r) = M - \frac{Q^2}{2r} . \quad (19.13)$$

Figure 19.1 illustrates the velocity fields in Schwarzschild and Reissner-Nordström black holes. Horizons occur where the radial velocity β equals the speed of light

$$\beta = \mp 1 , \quad (19.14)$$

with $-$ for black hole solutions, $+$ for white hole solutions. The phenomenology of Schwarzschild and Reissner-Nordström black holes has already been explored in Chapters 7 and 8.

Exercise 19.2 Coordinate transformation from Schwarzschild to Gullstrand-Painlevé. Show that the Schwarzschild metric transforms into the Gullstrand-Painlevé metric under the coordinate transformation of the time coordinate

$$dt_{\text{ff}} = dt - \frac{\beta}{1 - \beta^2} dr . \quad (19.15)$$

Exercise 19.3 Velocity of a person who free-falls radially from rest. Confirm that β given by equation (21.36) is indeed the velocity (19.11) of a person who free-falls radially from rest at infinity in the Reissner-Nordström geometry.

The Gullstrand-Painlevé line-element (19.10) encodes a vierbein with an orthonormal tetrad metric $\gamma_{mn} = \eta_{mn}$ through

$$e^0{}_{\mu} dx^{\mu} = dt_{\text{ff}} , \quad (19.16a)$$

$$e^1{}_{\mu} dx^{\mu} = dr - \beta dt_{\text{ff}} , \quad (19.16b)$$

$$e^2{}_{\mu} dx^{\mu} = r d\theta , \quad (19.16c)$$

$$e^3{}_{\mu} dx^{\mu} = r \sin \theta d\phi . \quad (19.16d)$$

Explicitly, the vierbein $e^m{}_{\mu}$ of the Gullstrand-Painlevé line-element (19.10), and the corresponding inverse vierbein $e_m{}^{\mu}$, are the matrices

$$e^m{}_{\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\beta & 1 & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & 0 & 0 & r \sin \theta \end{pmatrix} , \quad e_m{}^{\mu} = \begin{pmatrix} 1 & \beta & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1/(r \sin \theta) \end{pmatrix} . \quad (19.17)$$

According to equation (19.3), the coordinate 4-velocity of the tetrad frame through the coordinates is

$$\left\{ \frac{dt_{\text{ff}}}{d\tau}, \frac{dr}{d\tau}, \frac{d\theta}{d\tau}, \frac{d\phi}{d\tau} \right\} = u^{\mu} = e_0{}^{\mu} = \{1, \beta, 0, 0\} , \quad (19.18)$$

consistent with the claim (19.11) that β represents a radial velocity, while t_{ff} coincides with the proper time in the tetrad frame.

The tetrad and coordinate axes γ_m and e_{μ} are related to each other by the vierbein in the usual way, $\gamma_m = e_m{}^{\mu} e_{\mu}$ and $e_{\mu} = e^m{}_{\mu} \gamma_m$. The Gullstrand-Painlevé orthonormal tetrad axes γ_m are thus related to the coordinate axes e_{μ} by

$$\gamma_0 = e_{t_{\text{ff}}} + \beta e_r , \quad \gamma_1 = e_r , \quad \gamma_2 = e_{\theta}/r , \quad \gamma_3 = e_{\phi}/(r \sin \theta) . \quad (19.19)$$

Physically, the Gullstrand-Painlevé-Cartesian tetrad (19.19) are the axes of locally inertial orthonormal frames (with spatial axes γ_a oriented in the polar directions r, θ, ϕ) attached to observers who free-fall radially, without rotating, starting from zero velocity and zero angular momentum at infinity. The fact that the tetrad axes γ_m are parallel-transported, without precessing, along the worldlines of the radially free-falling observers can be confirmed by checking that the tetrad connections Γ_{nm0} with final index 0 all vanish, which implies that

$$\frac{d\gamma_m}{d\tau} = \partial_0 \gamma_m \equiv \Gamma_{m0}^n \gamma_n = 0 . \quad (19.20)$$

That the proper time derivative $d/d\tau$ in equation (19.20) of a person at rest in the tetrad frame, with 4-velocity (19.2), is equal to the directed time derivative ∂_0 follows from

$$\frac{d}{d\tau} = u^{\mu} \frac{\partial}{\partial x^{\mu}} = u^m \partial_m = \partial_0 . \quad (19.21)$$

19.2.2 Gullstrand-Painlevé-Cartesian tetrad

The manner in which the Gullstrand-Painlevé line-element depicts a flow of space into a black hole is elucidated further if the line-element is written in Cartesian rather than spherical polar coordinates. Introduce a Cartesian coordinate system $x^\mu \equiv \{t_{\text{ff}}, x^\alpha\} \equiv \{t_{\text{ff}}, x, y, z\}$. The Gullstrand-Painlevé metric in these Cartesian coordinates is

$$ds^2 = -dt_{\text{ff}}^2 + \delta_{\alpha\beta}(dx^\alpha - \beta^\alpha dt_{\text{ff}})(dx^\beta - \beta^\beta dt_{\text{ff}}) , \quad (19.22)$$

with implicit summation over spatial indices $\alpha, \beta = x, y, z$. The β^α in the metric (19.22) are the components of the radial velocity expressed in Cartesian coordinates

$$\beta^\alpha = \beta \left\{ \frac{x}{r}, \frac{y}{r}, \frac{z}{r} \right\} . \quad (19.23)$$

The vierbein $e^m{}_\mu$ and inverse vierbein $e_m{}^\mu$ encoded in the Gullstrand-Painlevé-Cartesian line-element (19.22) are

$$e^m{}_\mu = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\beta^1 & 1 & 0 & 0 \\ -\beta^2 & 0 & 1 & 0 \\ -\beta^3 & 0 & 0 & 1 \end{pmatrix} , \quad e_m{}^\mu = \begin{pmatrix} 1 & \beta^x & \beta^y & \beta^z \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad (19.24)$$

The tetrad axes γ_m of the Gullstrand-Painlevé-Cartesian line-element (19.22) are related to the coordinate tangent axes e_μ by

$$\gamma_0 = e_{t_{\text{ff}}} + \beta^\alpha e_\alpha , \quad \gamma_a = \delta_a^\alpha e_\alpha , \quad (19.25)$$

and conversely the coordinate tangent axes e_μ are related to the tetrad axes γ_m by

$$e_{t_{\text{ff}}} = \gamma_0 - \beta^a \gamma_a , \quad e_\alpha = \delta_\alpha^a \gamma_a . \quad (19.26)$$

Note that the tetrad-frame contravariant components β^a of the radial velocity coincide with the coordinate-frame contravariant components β^α ; for clarification of this point see the more general equation (19.56) for a rotating black hole. The Gullstrand-Painlevé-Cartesian tetrad axes (19.25) are the same as the tetrad axes (19.19), but rotated to point in Cartesian directions x, y, z rather than in polar directions r, θ, ϕ . Like the polar tetrad, the Cartesian tetrad axes γ_m are parallel-transported, without precessing, along the worldlines of radially free-falling observers, as can be confirmed by checking once again that the tetrad connections Γ_{nm0} with final index 0 all vanish.

Remarkably, the transformation (19.25) from coordinate to tetrad axes is just a Galilean transformation of space and time, which shifts the time axis by velocity β along the direction of motion, but which leaves unchanged both the time component of the time axis and all the spatial axes. In other words, the black hole behaves as if it were a river of space that flows radially inward through Galilean space and time at the Newtonian escape velocity.

19.2.3 Gullstrand-Painlevé fishies

The Gullstrand-Painlevé line-element paints a picture of locally inertial frames falling like a river of space into a spherical black hole. What happens to fishies swimming in that river? Of course general relativity supplies a mathematical answer in the form of the geodesic equation of motion (19.27). Does that mathematical answer lead to further conceptual insight?

Consider a fishy swimming in the Gullstrand-Painlevé river, with some arbitrary tetrad-frame 4-velocity u^m , and consider a tetrad-frame 4-vector p^k attached to the fishy. If the fishy is in free-fall, then the geodesic equation of motion for p^k is as usual

$$\frac{dp^k}{d\tau} + \Gamma_{mn}^k u^n p^m = 0 . \quad (19.27)$$

As remarked in §11.11, for a constant (for example Minkowski) tetrad metric, as here, the tetrad connections Γ_{mn}^k constitute a set of four generators of Lorentz transformations, one in each of the directions n . In particular $\Gamma_{mn}^k u^n$ is the generator of a Lorentz transformation along the path of a fishy moving with 4-velocity u^n . In a small (infinitesimal) time $\delta\tau$, the fishy moves a proper distance $\delta\xi^n \equiv u^n \delta\tau$ relative to the infalling river. This proper distance $\delta\xi^n = e_{\nu}^n \delta x^{\nu} = \delta_{\nu}^n (\delta x^{\nu} - \beta^{\nu} \delta t_{\text{ff}}) = \delta x^n - \beta^n \delta\tau$ equals the distance δx^n moved relative to the background Gullstrand-Painlevé-Cartesian coordinates, minus the distance $\beta^n \delta\tau$ moved by the river. The geodesic equation (19.27) says that the change δp^k in the tetrad 4-vector p^k in the time $\delta\tau$ is

$$\delta p^k = -\Gamma_{mn}^k \delta\xi^n p^m . \quad (19.28)$$

Equation (19.28) describes an infinitesimal Lorentz transformation $-\Gamma_{mn}^k \delta\xi^n$ of the 4-vector p^k .

Equation (19.28) is quite general in general relativity: it says that as a 4-vector p^k free-falls through a system of locally inertial tetrads, it finds itself Lorentz-transformed relative those tetrads. What is special about the Gullstrand-Painlevé-Cartesian tetrad is that the tetrad-frame connections, computed by the usual formula (11.54), are given by the coordinate gradient of the radial velocity (the following equation is valid component-by-component despite the nonmatching up-down placement of indices)

$$\boxed{\Gamma_{ab}^0 = \Gamma_{0b}^a = \partial_b \beta^a = \delta_b^{\beta} \frac{\partial \beta^a}{\partial x^{\beta}} \quad (a, b = 1, 2, 3)} . \quad (19.29)$$

The same property, that the tetrad connections are a pure coordinate gradient, holds also for the Doran-Cartesian tetrad for a rotating black hole, equation (19.59). With the connections (19.29), the change δp^k (19.28) in the tetrad 4-vector is

$$\delta p^0 = -\delta\beta^a p^a , \quad \delta p^a = -\delta\beta^a p^0 , \quad (19.30)$$

where $\delta\beta^a$ is the change in the velocity of the river as seen in the tetrad frame,

$$\delta\beta^a = \delta\xi^{\beta} \frac{\partial \beta^a}{\partial x^{\beta}} . \quad (19.31)$$

But equation (19.30) is nothing more than an infinitesimal Lorentz boost by a velocity change $\delta\beta^a$. This

shows that a fishy swimming in the river follows the rules of special relativity, being Lorentz boosted by tidal changes $\delta\beta^a$ in the river velocity from place to place.

Is it correct to interpret equation (19.31) as giving the change $\delta\beta^a$ in the river velocity seen by a fishy? Of course general relativity demands that equation (19.31) be mathematically correct; the issue is merely one of interpretation. Shouldn't the change in the river velocity really be

$$\delta\beta^a \stackrel{?}{=} \delta x^\nu \frac{\partial\beta^a}{\partial x^\nu}, \quad (19.32)$$

where δx^ν is the full change in the coordinate position of the fishy? No. Part of the change (19.32) in the river velocity can be attributed to the change in the velocity of the river itself over the time $\delta\tau$, which is $\delta x_{\text{river}}^\nu \partial\beta^a / \partial x^\nu$ with $\delta x_{\text{river}}^\nu = \beta^\nu \delta\tau = \beta^\nu \delta t_{\text{ff}}$. The change in the velocity relative to the flowing river is

$$\delta\beta^a = (\delta x^\nu - \delta x_{\text{river}}^\nu) \frac{\partial\beta^a}{\partial x^\nu} = (\delta x^\nu - \beta^\nu \delta t_{\text{ff}}) \frac{\partial\beta^a}{\partial x^\nu}, \quad (19.33)$$

which reproduces the earlier expression (19.31). Indeed, in the picture of fishies being carried by the river, it is essential to subtract the change in velocity of the river itself, as in equation (19.33), because otherwise fishies at rest in the river (going with the flow) would not continue to remain at rest in the river.

19.3 Boyer-Lindquist tetrad

The Boyer-Lindquist metric for an ideal rotating black hole was explored already in Chapter 9. With the tetrad formalism in hand, the advantages of the Boyer-Lindquist tetrad for portraying the Kerr-Newman geometry become manifest. With respect to the orthonormal Boyer-Lindquist tetrad, the electromagnetic field is purely radial, and the energy-momentum and Weyl tensors are diagonal. The Boyer-Lindquist tetrad is aligned with the principal (outgoing and ingoing) null congruences.

The Boyer-Lindquist orthonormal tetrad is encoded in the Boyer-Lindquist metric, equation (9.1),

$$ds^2 = -\frac{R^2\Delta}{\rho^2} (dt - a \sin^2\theta d\phi)^2 + \frac{\rho^2}{R^2\Delta} dr^2 + \rho^2 d\theta^2 + \frac{R^4 \sin^2\theta}{\rho^2} \left(d\phi - \frac{a}{R^2} dt\right)^2, \quad (19.34)$$

where

$$R \equiv \sqrt{r^2 + a^2}, \quad \rho \equiv \sqrt{r^2 + a^2 \cos^2\theta}, \quad \Delta \equiv 1 - \frac{2Mr}{R^2} + \frac{Q^2}{R^2} = 1 - \beta^2. \quad (19.35)$$

Explicitly, the vierbein $e^m{}_\mu$ of the Boyer-Lindquist orthonormal tetrad is

$$e^m{}_\mu = \begin{pmatrix} R\sqrt{\Delta}/\rho & 0 & 0 & -a \sin^2\theta R\sqrt{\Delta}/\rho \\ 0 & \rho/(R\sqrt{\Delta}) & 0 & 0 \\ 0 & 0 & \rho & 0 \\ -a \sin\theta/\rho & 0 & 0 & R^2 \sin\theta/\rho \end{pmatrix}, \quad (19.36)$$

with inverse vierbein $e_m{}^\mu$

$$e_m{}^\mu = \frac{1}{\rho} \begin{pmatrix} R/\sqrt{\Delta} & 0 & 0 & a/(R\sqrt{\Delta}) \\ 0 & R\sqrt{\Delta} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ a \sin \theta & 0 & 0 & 1/\sin \theta \end{pmatrix}, \quad (19.37)$$

With respect to the Boyer-Lindquist tetrad, only the time component A^t of the electromagnetic potential A^m is nonvanishing,

$$A^m = \left\{ \frac{Qr}{\rho R\sqrt{\Delta}}, 0, 0, 0 \right\}. \quad (19.38)$$

Only the radial components E and B of the electric and magnetic fields are nonvanishing, and they are given by the complex combination

$$E + IB = \frac{Q}{(r - Ia \cos \theta)^2}, \quad (19.39)$$

or explicitly

$$E = \frac{Q(r^2 - a^2 \cos^2 \theta)}{\rho^4}, \quad B = \frac{2Qar \cos \theta}{\rho^4}. \quad (19.40)$$

The electromagnetic field (19.39) satisfies Maxwell's equations (22.56) with zero electric charge and current, $j^n = 0$, except at the singularity $\rho = 0$.

The nonvanishing components of the tetrad-frame Einstein tensor G_{mn} are

$$G_{mn} = \frac{Q^2}{\rho^4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (19.41)$$

which is the energy-momentum tensor of the electromagnetic field. The nonvanishing components of the tetrad-frame Weyl tensor C_{klmn} are

$$-\frac{1}{2} C_{0101} = \frac{1}{2} C_{2323} = C_{0202} = C_{0303} = -C_{1212} = -C_{1313} = \text{Re } C, \quad (19.42a)$$

$$\frac{1}{2} C_{0123} = C_{0213} = -C_{0312} = \text{Im } C, \quad (19.42b)$$

where C is the complex Weyl scalar

$$C = -\frac{1}{(r - Ia \cos \theta)^3} \left(M - \frac{Q^2}{r + Ia \cos \theta} \right). \quad (19.43)$$

In the Boyer-Lindquist tetrad, the photon 4-velocity $v^m \equiv e^m{}_\mu v^\mu = e^m{}_\mu dx^\mu/d\lambda$ on the principal null congruences is radial,

$$v^t = \pm \frac{\rho}{R\sqrt{\Delta}}, \quad v^r = \pm \frac{\rho}{R\sqrt{\Delta}}, \quad v^\theta = 0, \quad v^\phi = 0. \quad (19.44)$$

Exercise 19.4 Dragging of inertial frames around a Kerr-Newman black hole. What is the coordinate-frame 4-velocity u^μ of the Boyer-Lindquist tetrad through the Boyer-Lindquist coordinates?

19.4 Doran waterfall

The picture of space falling into a black hole like a river or waterfall works also for rotating black holes. For Kerr-Newman rotating black holes, the counterpart of the Gullstrand-Painlevé metric is the Doran (2000) metric, equation (9.32). The relation between the Doran and Boyer-Lindquist tetrads is

$$\gamma_{\text{Doran}}^m = L^m_n \gamma_{\text{BL}}^n, \quad (19.45)$$

where L^m_n is a product of (noncommuting) Lorentz boosts in the azimuthal and radial directions,

$$L^m_n = \begin{pmatrix} R/\rho & 0 & 0 & a \sin \theta / \rho \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ a \sin \theta / \rho & 0 & 0 & R/\rho \end{pmatrix} \begin{pmatrix} 1/\sqrt{1-\beta^2} & -\beta/\sqrt{1-\beta^2} & 0 & 0 \\ -\beta/\sqrt{1-\beta^2} & 1/\sqrt{1-\beta^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (19.46)$$

with $\Delta = 1 - \beta^2$.

19.4.1 Doran-Cartesian coordinates

The space river that falls into a rotating black hole has a twist. One might have expected that the rotation of the black hole would be manifested by a velocity that spirals inward, but that is not the case. Instead, the river is characterized not merely by a velocity but also by a twist. The velocity and the twist together comprise a 6-dimensional river bivector ω_{km} , equation (19.60) below, whose electric part is the velocity, and whose magnetic part is the twist. Recall that the 6-dimensional group of Lorentz transformations is generated by a combination of 3-dimensional Lorentz boosts and 3-dimensional spatial rotations. A fishy that swims through the river is Lorentz boosted by tidal changes in the velocity, and rotated by tidal changes in the twist, equation (19.69).

Thanks to the twist, unlike the Gullstrand-Painlevé metric, the Doran metric is not spatially flat at constant free-fall time t_{ff} . Rather, the spatial metric is sheared in the azimuthal direction. Just as the velocity produces a Lorentz boost that makes the metric nonflat with respect to the time components, so also the twist produces a rotation that makes the metric nonflat with respect to the spatial components.

In place of the polar coordinates $\{r, \theta, \phi_{\text{ff}}\}$ of the Doran metric, equations (9.33), introduce corresponding Doran-Cartesian coordinates $\{x, y, z\}$ with z taken along the rotation axis of the black hole (the black hole rotates right-handedly about z , for positive spin parameter a)

$$x \equiv R \sin \theta \cos \phi_{\text{ff}}, \quad y \equiv R \sin \theta \sin \phi_{\text{ff}}, \quad z \equiv r \cos \theta. \quad (19.47)$$

The metric in Doran-Cartesian coordinates $x^\mu \equiv \{t_{\text{ff}}, x^\alpha\} \equiv \{t_{\text{ff}}, x, y, z\}$, is

$$\boxed{ds^2 = -dt_{\text{ff}}^2 + \delta_{\alpha\beta} (dx^\alpha - \beta^\alpha \alpha_\kappa dx^\kappa) (dx^\beta - \beta^\beta \alpha_\lambda dx^\lambda)} \quad (19.48)$$

where α_μ is the rotational velocity vector

$$\alpha_\mu = \left\{ 1, \frac{ay}{R^2}, -\frac{ax}{R^2}, 0 \right\} , \quad (19.49)$$

and β^μ is the velocity vector

$$\beta^\mu = \frac{\beta R}{\rho} \left\{ 0, \frac{xr}{R\rho}, \frac{yr}{R\rho}, \frac{zR}{r\rho} \right\} . \quad (19.50)$$

The rotational velocity and radial velocity vectors are orthogonal

$$\alpha_\mu \beta^\mu = 0 . \quad (19.51)$$

For the Kerr-Newman metric, the radial velocity β is

$$\beta = \mp \frac{\sqrt{2Mr - Q^2}}{R} \quad (19.52)$$

with $-$ for black hole (infalling), $+$ for white hole (outfalling) solutions. Horizons occur where

$$\beta = \mp 1 , \quad (19.53)$$

with $\beta = -1$ for black hole horizons, and $\beta = 1$ for white hole horizons. Note that the squared magnitude $\beta_\mu \beta^\mu$ of the velocity vector is not β^2 , but rather differs from β^2 by a factor of R^2/ρ^2 :

$$\beta_\mu \beta^\mu = \beta_m \beta^m = \frac{\beta^2 R^2}{\rho^2} . \quad (19.54)$$

The point of the convention adopted here is that $\beta(r)$ is any and only a function of r , rather than depending also on θ through ρ . Moreover, with the convention here, β is ∓ 1 at horizons, equation (19.53). Finally, the 4-velocity β^μ is simply related to β by $\beta^\mu = (\beta/r) \partial r / \partial x^\mu$.

The Doran-Cartesian metric (19.48) encodes a vierbein e^m_μ and inverse vierbein e_m^μ

$$e^m_\mu = \delta_\mu^m - \alpha_\mu \beta^m , \quad e_m^\mu = \delta_m^\mu + \alpha_m \beta^\mu . \quad (19.55)$$

Here the tetrad-frame components α_m of the rotational velocity vector and β^m of the radial velocity vector are

$$\alpha_m = e_m^\mu \alpha_\mu = \delta_m^\mu \alpha_\mu , \quad \beta^m = e^m_\mu \beta^\mu = \delta_\mu^m \beta^\mu , \quad (19.56)$$

which works thanks to the orthogonality (19.51) of α_μ and β^μ . Equation (19.56) says that the covariant tetrad-frame components of the rotational velocity vector are the same as its covariant coordinate-frame components in the Doran-Cartesian coordinate system, $\alpha_m = \alpha_\mu$, and likewise the contravariant tetrad-frame components of the radial velocity vector are the same as its contravariant coordinate-frame components, $\beta^m = \beta^\mu$.

19.4.2 Doran-Cartesian tetrad

Like the Gullstrand-Painlevé tetrad, the Doran-Cartesian tetrad $\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$ is aligned with the Cartesian rest frame $e_\mu \equiv \{e_{t_{\text{ff}}}, e_x, e_y, e_z\}$ at infinity, and is parallel-transported, without precessing, by observers who free-fall from zero velocity and zero angular momentum at infinity, as can be confirmed by checking that the tetrad connections with final index 0 all vanish, $\Gamma_{nm0} = 0$, equation (19.20).

Let $\parallel$ and $\perp$ subscripts denote horizontal radial and azimuthal directions respectively, so that

$$\begin{aligned}\gamma_\parallel &\equiv \cos \phi_{\text{ff}} \gamma_1 + \sin \phi_{\text{ff}} \gamma_2, & \gamma_\perp &\equiv -\sin \phi_{\text{ff}} \gamma_1 + \cos \phi_{\text{ff}} \gamma_2, \\ e_\parallel &\equiv \cos \phi_{\text{ff}} e_x + \sin \phi_{\text{ff}} e_y, & e_\perp &\equiv -\sin \phi_{\text{ff}} e_x + \cos \phi_{\text{ff}} e_y.\end{aligned}\tag{19.57}$$

Then the relation between Doran-Cartesian tetrad axes γ_m and the tangent axes e_μ of the Doran-Cartesian metric (19.48) is

$$\gamma_0 = e_{t_{\text{ff}}} + \beta^\alpha e_\alpha, \tag{19.58a}$$

$$\gamma_\parallel = e_\parallel, \tag{19.58b}$$

$$\gamma_\perp = e_\perp - \frac{a \sin \theta}{R} \beta^\alpha e_\alpha, \tag{19.58c}$$

$$\gamma_3 = e_z. \tag{19.58d}$$

The relations (19.58) resemble those (19.25) of the Gullstrand-Painlevé tetrad, except that the azimuthal tetrad axis $\gamma_\perp$ is shifted radially relative to the azimuthal tangent axis $e_\perp$. This shift reflects the fact that, unlike the Gullstrand-Painlevé metric, the Doran metric is not spatially flat at constant free-fall time, but rather is sheared azimuthally.

19.4.3 Doran fishies

The tetrad-frame connections equal the ordinary coordinate partial derivatives in Doran-Cartesian coordinates of a bivector (antisymmetric tensor) ω_{km}

$$\boxed{\Gamma_{kmn} = -\delta_n^\nu \frac{\partial \omega_{km}}{\partial x^\nu}}, \tag{19.59}$$

which I call the river field because it encapsulates all the properties of the infalling river of space. The bivector river field ω_{km} is

$$\boxed{\omega_{km} = \alpha_k \beta_m - \alpha_m \beta_k - \varepsilon_{0kma} \zeta^a}, \tag{19.60}$$

where $\beta_m = \eta_{mn} \beta^n$, the totally antisymmetric tensor ε_{klmn} is normalized so that $\varepsilon_{0123} = -1$, and the vector ζ^a points vertically upward along the rotation axis of the black hole

$$\zeta^a \equiv \{0, 0, 0, \zeta\}, \quad \zeta \equiv a \int_\infty^r \frac{\beta dr}{R^2}. \tag{19.61}$$

The electric part of ω_{km} , where one of the indices is time 0, constitutes the velocity vector β^a

$$\omega_{0a} = \beta^a \quad (19.62)$$

while the magnetic part of ω_{km} , where both indices are spatial, constitutes the twist vector μ^a defined by

$$\mu^a \equiv \frac{1}{2} \varepsilon^{0akm} \omega_{km} = \varepsilon^{0akm} \alpha_k \beta_m + \zeta^a . \quad (19.63)$$

The sense of the twist is that induces a right-handed rotation about an axis equal to the direction of μ^a by an angle equal to the magnitude of μ^a . In 3-vector notation, with $\boldsymbol{\mu} \equiv \mu^a$, $\boldsymbol{\alpha} \equiv \alpha_a$, $\boldsymbol{\beta} \equiv \beta^a$, $\boldsymbol{\zeta} \equiv \zeta^a$,

$$\boldsymbol{\mu} \equiv \boldsymbol{\alpha} \times \boldsymbol{\beta} + \boldsymbol{\zeta} . \quad (19.64)$$

In terms of the velocity and twist vectors, the river field ω_{km} is

$$\omega_{km} = \begin{pmatrix} 0 & \beta^1 & \beta^2 & \beta^3 \\ -\beta^1 & 0 & \mu^3 & -\mu^2 \\ -\beta^2 & -\mu^3 & 0 & \mu^1 \\ -\beta^3 & \mu^2 & -\mu^1 & 0 \end{pmatrix} . \quad (19.65)$$

Note that the sign of the electric part $\boldsymbol{\beta}$ of ω_{km} is opposite to the sign of the analogous electric field $\boldsymbol{E}$ associated with an electromagnetic field F_{km} , equation (4.46); but the adopted signs are natural in that the river field induces boosts in the direction of the velocity β^a , and right-handed rotations about the twist μ^a . Like a static electric field, the velocity vector β^a is the gradient of a potential

$$\beta^a = \delta_\alpha^a \frac{\partial}{\partial x^\alpha} \int^r \beta \, dr , \quad (19.66)$$

but unlike a magnetic field the twist vector μ^a is not pure curl: rather, it is $\mu^a + \zeta^a$ that is pure curl. Figure 19.2 illustrates the velocity and twist fields in a Kerr black hole.

With the tetrad connection coefficients given by equation (19.59), the equation of motion (19.27) for a 4-vector p^k attached to a fishy following a geodesic in the Doran river translates to

$$\frac{dp^k}{d\tau} = \delta_n^\nu \frac{\partial \omega^k_m}{\partial x^\nu} u^n p^m . \quad (19.67)$$

In a proper time $\delta\tau$, the fishy moves a proper distance $\delta\xi^m \equiv u^m \delta\tau$ relative to the background Doran-Cartesian coordinates. As a result, the fishy sees a tidal change $\delta\omega^k_m$ in the river field

$$\delta\omega^k_m = \delta\xi^n \frac{\partial \omega^k_m}{\partial x^n} . \quad (19.68)$$

Consequently the 4-vector p^k is changed by

$$p^k \rightarrow p^k + \delta\omega^k_m p^m . \quad (19.69)$$

But equation (19.69) corresponds to an infinitesimal Lorentz transformation by $\delta\omega^k_m$, equivalent to a Lorentz boost by $\delta\beta^a$ and a rotation by $\delta\mu^a$.

As discussed previously with regard to the Gullstrand-Painlevé river, §19.2.3, the tidal change $\delta\omega^k_m$,

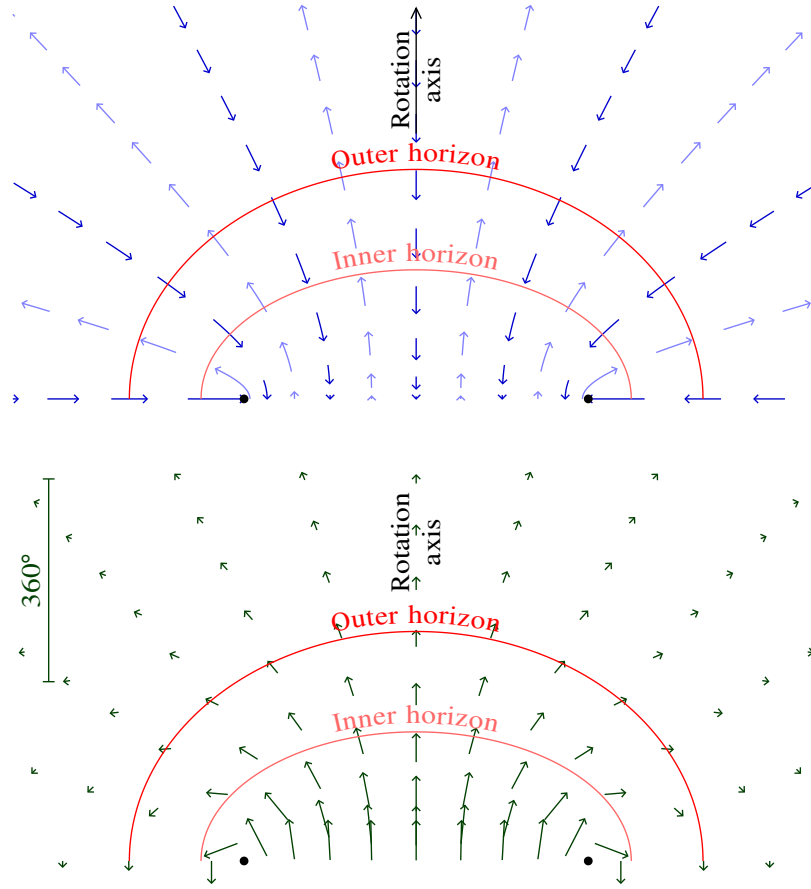


Figure 19.2 (Upper panel) velocity β^a and (lower panel) twist μ^a vector fields for a Kerr black hole with spin parameter $a = 0.96$. Both vectors lie, as shown, in the plane of constant free-fall azimuthal angle ϕ_{ff} . The vertical bar in the lower panel shows the length of a twist vector corresponding to a full rotation of 360° .

equation (19.68), in the river field seen by a fishy is not the full change $\delta x^\nu \partial \omega^k_m / \partial x^\nu$ relative to the background coordinates, but rather the change relative to the river

$$\delta \omega^k_m = (\delta x^\nu - \delta x^\nu_{\text{river}}) \frac{\partial \omega^k_m}{\partial x^\nu} = [\delta x^\nu - \beta^\nu (\delta t_{\text{ff}} - a \sin^2 \theta \delta \phi_{\text{ff}})] \frac{\partial \omega^k_m}{\partial x^\nu}, \quad (19.70)$$

with the change in the velocity and twist of the river itself subtracted off.

That there exists a tetrad (the Doran-Cartesian tetrad) where the tetrad-frame connections are a coordinate gradient of a bivector, equation (19.59), is a peculiar feature of ideal black holes. It is an intriguing

thought that perhaps the 6 physical degrees of freedom of a general spacetime might always be encoded in the 6 degrees of freedom of a bivector, but that is not true.

Exercise 19.5 River model of the Friedmann-Lemaître-Robertson-Walker metric. Show that the flat FLRW line-element

$$ds^2 = -dt^2 + a^2(dx^2 + x^2 do^2) \quad (19.71)$$

can be re-expressed as

$$ds^2 = -dt^2 + (dr - Hr dt)^2 + r^2 do^2, \quad (19.72)$$

where $r \equiv ax$ is the proper radial distance, and $H \equiv \dot{a}/a$ is the Hubble parameter. Interpret the line-element (19.72). Is there a generalization to a nonflat FLRW universe?

General spherically symmetric spacetimes

20.1 Spherical spacetime

In 3+1 spacetime dimensions, spherical spacetimes have 2 physical degrees of freedom. Spherical symmetry eliminates any angular degrees of freedom, leaving 4 adjustable metric coefficients g_{tt} , g_{tr} , g_{rr} , and $g_{\theta\theta}$. But coordinate transformations of the time t and radial r coordinates remove 2 degrees of freedom, leaving a spherical spacetime with a net 2 physical degrees of freedom. Spherical spacetimes have 4 distinct Einstein equations (20.39). But 2 of the Einstein equations serve to enforce energy-momentum conservation, so the evolution of the spacetime is governed by 2 Einstein equations, in agreement with the number of physical degrees of freedom of spherical spacetime.

The 2 degrees of freedom mean that spherical spacetimes in general relativity have a richer structure than in Newtonian gravity, which has only one degree of freedom, the Newtonian potential Φ . The richer structure is most striking in the case of the mass inflation instability, Chapter 21, which is an intrinsically general relativistic instability, with no Newtonian analogue.

The generalization to arbitrary spacetime dimensions is addressed in Exercise 20.3.

20.2 Spherical line-element

The spherical line-element adopted in this Chapter is, in spherical polar coordinates $x^\mu \equiv \{t, r, \theta, \phi\}$,

$$ds^2 = -\alpha^2 dt^2 + \frac{1}{\beta_1^2} (dr - \alpha\beta_0 dt)^2 + r^2 d\theta^2. \quad (20.1)$$

Here r is the circumferential radius, defined such that the circumference around any great circle is $2\pi r$. The line-element (20.1) is in ADM form (17.8) with lapse α and shift $\alpha\beta_0$. The notation β_m is motivated by fact that $\{\beta_0, \beta_1, 0, 0\}$ forms a tetrad-frame 4-vector, equation (20.9). Spherical symmetry imposes that α , β_0 , β_1 are functions of time t and radius r but independent of the angles θ and ϕ . As expounded in §11.3, through $ds^2 = \eta_{mn} e^m{}_\mu e^n{}_\nu dx^\mu dx^\nu$ the line-element (20.1) encodes not only a metric, but also a locally inertial tetrad $\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$. The off-diagonal character of the line-element allows the tetrad to flow through the

coordinates. This flexibility is especially useful for black holes, since no locally inertial frame can remain at rest inside the horizon of a black hole.

The vierbein $e^m{}_\mu$ can be read off from the line-element (20.1):

$$e^0{}_\mu dx^\mu = \alpha dt , \quad (20.2a)$$

$$e^1{}_\mu dx^\mu = \frac{1}{\beta_1} (dr - \alpha\beta_0 dt) , \quad (20.2b)$$

$$e^2{}_\mu dx^\mu = r d\theta , \quad (20.2c)$$

$$e^3{}_\mu dx^\mu = r \sin \theta d\phi . \quad (20.2d)$$

The vierbein $e^m{}_\mu$ and inverse vierbein $e_m{}^\mu$ corresponding to the spherical line-element (20.1) are

$$e^m{}_\mu = \begin{pmatrix} \alpha & 0 & 0 & 0 \\ -\alpha\beta_0/\beta_1 & 1/\beta_1 & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & 0 & 0 & r \sin \theta \end{pmatrix} , \quad e_m{}^\mu = \begin{pmatrix} 1/\alpha & \beta_0 & 0 & 0 \\ 0 & \beta_1 & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1/(r \sin \theta) \end{pmatrix} . \quad (20.3)$$

As in the ADM formalism, §17.1, the tetrad time axis γ_0 is chosen to be orthogonal to hypersurfaces of constant time t . The directed derivatives ∂_0 and ∂_1 along the time and radial tetrad axes γ_0 and γ_1 are

$$\partial_0 = e_0{}^\mu \frac{\partial}{\partial x^\mu} = \frac{1}{\alpha} \frac{\partial}{\partial t} + \beta_0 \frac{\partial}{\partial r} , \quad \partial_1 = e_1{}^\mu \frac{\partial}{\partial x^\mu} = \beta_1 \frac{\partial}{\partial r} . \quad (20.4)$$

The tetrad-frame 4-velocity u^m of a person at rest in the tetrad frame is by definition $u^m = \{1, 0, 0, 0\}$. It follows that the coordinate 4-velocity u^μ of such a person is

$$u^\mu = e_m{}^\mu u^m = e_0{}^\mu = \{1/\alpha, \beta_0, 0, 0\} . \quad (20.5)$$

A person instantaneously at rest in the tetrad frame satisfies $dr/dt = \alpha\beta_0$ according to equation (20.5), so it follows from the line-element (20.1) that the proper time τ of a person at rest in the tetrad frame is related to the coordinate time t by

$$d\tau = \alpha dt \quad \text{in tetrad rest frame} . \quad (20.6)$$

The directed time derivative ∂_0 is just the proper time derivative along the worldline of a person continuously at rest in the tetrad frame (and who is therefore not in free-fall, but accelerating with the tetrad frame), which follows from

$$\frac{d}{d\tau} = \frac{dx^\mu}{d\tau} \frac{\partial}{\partial x^\mu} = u^\mu \frac{\partial}{\partial x^\mu} = u^m \partial_m = \partial_0 . \quad (20.7)$$

By contrast, the proper time derivative measured by a person who is instantaneously at rest in the tetrad frame, but is in free-fall, is the covariant time derivative

$$\frac{D}{D\tau} = \frac{dx^\mu}{d\tau} D_\mu = u^\mu D_\mu = u^m D_m = D_0 . \quad (20.8)$$

Since the coordinate radius r has been defined to be the circumferential radius, a gauge-invariant definition,

it follows that the tetrad-frame gradient ∂_m of the coordinate radius r is a tetrad-frame 4-vector (a coordinate gauge-invariant object),

$$\partial_m r = e_m^\mu \frac{\partial r}{\partial x^\mu} = e_m^r = \beta_m = \{\beta_0, \beta_1, 0, 0\} \quad \text{a tetrad 4-vector} . \quad (20.9)$$

This accounts for the notation β_0 and β_1 introduced above. The component β_0 can be interpreted as the radial velocity of the tetrad frame, equation (20.5),

$$\beta_0 = \frac{dr}{d\tau} . \quad (20.10)$$

The component β_1 can be interpreted as the energy per unit mass of an object at rest in the tetrad frame, equation (20.52).

Since β_m is a tetrad 4-vector, its scalar product with itself must be a scalar. This scalar defines the **interior mass** $M(t, r)$, also called the Misner-Sharp mass (Misner and Sharp, 1964), by

$$\boxed{1 - \frac{2M}{r} \equiv \beta_m \beta^m = -\beta_0^2 + \beta_1^2} \quad \text{a coordinate and tetrad scalar} . \quad (20.11)$$

The interpretation of M as the interior mass will become evident below, §20.9.

The horizon function Δ is defined by

$$\Delta \equiv \beta^m \beta_m = 1 - \frac{2M}{r} . \quad (20.12)$$

Apparent horizons occur where the horizon function is zero, $\Delta = 0$, that is, where the 4 vector β_m is null, a gauge-invariant condition. The condition for an apparent horizon is

$$r = 2M , \quad (20.13)$$

which holds in any spherically symmetric geometry, not just the Schwarzschild geometry. In general the interior mass M varies with radius r ; only in the Schwarzschild geometry is the interior mass M constant.

Inside horizons, where the horizon function Δ is negative, the velocity β_0 cannot be zero: the tetrad must move superluminally through the radial coordinate. Similarly, outside horizons, where the horizon function Δ is positive, the energy per unit mass β_1 cannot be zero. Inside horizons, the energy per unit mass β_1 can be either positive, in which case the tetrad frame is called ingoing, or negative, in which case the tetrad frame is called outgoing. The tetrad can switch between ingoing and outgoing only inside horizons.

Exercise 20.1 Apparent horizon. Show that radial null geodesics in a spherical geometry satisfy

$$\frac{dr}{dt} = \alpha(\beta_0 \pm \beta_1) . \quad (20.14)$$

An apparent horizon occurs where outgoing radial null geodesics are not moving radially, $dr/dt = 0$. Conclude that an apparent horizon occurs where (choosing α and β_1 positive without loss of generality)

$$\beta_0 = -\beta_1 . \quad (20.15)$$

20.3 Rest diagonal line-element

Although this is not the choice adopted here, the line-element (20.1) can always be brought to diagonal form by a coordinate transformation $t \rightarrow t_\times$ (subscripted $\times$ for diagonal) of the time coordinate. The t - r part of the metric is

$$g_{tt} dt^2 + 2 g_{tr} dt dr + g_{rr} dr^2 = \frac{1}{g_{tt}} [(g_{tt} dt + g_{tr} dr)^2 + (g_{tt} g_{rr} - g_{tr}^2) dr^2] . \quad (20.16)$$

This can be diagonalized by choosing the time coordinate $t_\times$ such that

$$f dt_\times = g_{tt} dt + g_{tr} dr \quad (20.17)$$

for some integrating factor $f(t, r)$. Equation (20.17) can be solved by choosing $t_\times$ to be constant along integral curves

$$\frac{dr}{dt} = - \frac{g_{tt}}{g_{tr}} . \quad (20.18)$$

The resulting diagonal **rest line-element** is

$$ds^2 = -\alpha_\times^2 dt_\times^2 + \frac{dr^2}{1 - 2M/r} + r^2 d\phi^2 . \quad (20.19)$$

The line-element (20.19) corresponds physically to the case where the tetrad frame is taken to be at rest in the spatial coordinates, $\beta_0 = 0$, as can be seen by comparing it to the earlier line-element (20.1). In changing the tetrad frame from one moving at $dr/dt = \alpha\beta_0$ to one that is at rest (at constant circumferential radius r), a tetrad transformation has in effect been done at the same time as the coordinate transformation (20.17), the tetrad transformation being precisely that needed to make the line-element (20.19) diagonal. The metric coefficient g_{rr} in the line-element (20.19) follows from the fact that $\beta_1^2 = 1 - 2M/r$ when $\beta_0 = 0$, equation (20.11). The transformed time coordinate $t_\times$ is unspecified up to a transformation $t_\times \rightarrow f(t_\times)$. If the spacetime is asymptotically flat at infinity, then a natural way to fix the transformation is to choose $t_\times$ to be the proper time at rest at infinity.

20.4 Comoving diagonal line-element

Although once again this is not the path followed here, the line-element (20.1) can also be brought to diagonal form by a coordinate transformation $r \rightarrow r_\times$, where, analogously to equation (20.17), $r_\times$ is chosen to satisfy

$$f dr_\times = g_{tr} dt + g_{rr} dr \equiv \frac{1}{\beta_1} (dr - \alpha\beta_0 dt) \quad (20.20)$$

for some integrating factor $f(t, r)$. The new coordinate $r_\times$ is constant along the worldline of an object at rest in the tetrad frame, with $dr/dt = \alpha\beta_0$, equation (20.5), so $r_\times$ can be regarded as a comoving radial coordinate. The comoving radial coordinate $r_\times$ could for example be chosen to equal the circumferential

radius r at some fixed instant of coordinate time t (say $t = 0$). The diagonal **comoving line-element** in this comoving coordinate system takes the form

$$ds^2 = -\alpha^2 dt^2 + \lambda^2 dr_\times^2 + r^2 d\theta^2, \quad (20.21)$$

where the circumferential radius $r(t, r_\times)$ is considered to be a function of time t and the comoving radial coordinate $r_\times$. Whereas in the rest line-element (20.19) the tetrad was changed from one that was moving at $dr/dt = \alpha\beta_0$ to one that was at rest, here the transformation keeps the tetrad unchanged. In both the rest and comoving diagonal line-elements (20.19) and (20.21) the tetrad is at rest relative to the respective radial coordinate r or $r_\times$; but whereas in the rest line-element (20.19) the radial coordinate was fixed to be the circumferential radius r , in the comoving line-element (20.21) the comoving radial coordinate $r_\times$ is a label that follows the tetrad. Because the tetrad is unchanged by the transformation to the comoving radial coordinate $r_\times$, the directed time and radial derivatives ∂_0 and ∂_1 are unchanged:

$$\partial_0 = \frac{1}{\alpha} \frac{\partial}{\partial t} \Big|_{r_\times} = \frac{1}{\alpha} \frac{\partial}{\partial t} \Big|_r + \beta_0 \frac{\partial}{\partial r} \Big|_t, \quad \partial_1 = \frac{1}{\lambda} \frac{\partial}{\partial r_\times} \Big|_t = \beta_1 \frac{\partial}{\partial r} \Big|_t. \quad (20.22)$$

20.5 Tetrad connections

Now turn the handle to proceed towards the Einstein equations. The nonvanishing tetrad connections coefficients Γ_{kmn} corresponding to the spherical line-element (20.1) are

$$\Gamma_{100} = h_0, \quad (20.23a)$$

$$\Gamma_{101} = h_1, \quad (20.23b)$$

$$\Gamma_{202} = \Gamma_{303} = \frac{\beta_0}{r}, \quad (20.23c)$$

$$\Gamma_{212} = \Gamma_{313} = \frac{\beta_1}{r}, \quad (20.23d)$$

$$\Gamma_{323} = \frac{\cot \theta}{r}, \quad (20.23e)$$

where h_0 is the proper radial acceleration (minus the gravitational force) experienced by a person at rest in the tetrad frame

$$h_0 \equiv \partial_1 \ln \alpha = \beta_1 \frac{\partial \ln \alpha}{\partial r}, \quad (20.24)$$

and h_1 is the “Hubble parameter” of the radial flow, as measured in the tetrad rest frame, defined by

$$h_1 \equiv \beta_0 \frac{\partial \ln \alpha \beta_0}{\partial r} - \partial_0 \ln \beta_1. \quad (20.25)$$

The interpretation of h_0 as a proper acceleration and h_1 as a radial Hubble parameter goes as follows. The tetrad-frame 4-velocity u^m of a person at rest in the tetrad frame is by definition $u^m = \{1, 0, 0, 0\}$. If the person at rest were in free fall, then the proper acceleration would be zero, but because this is a general

spherical spacetime, the tetrad frame is not necessarily in free fall. The proper acceleration experienced by a person continuously at rest in the tetrad frame is the proper time derivative $Du^m/D\tau$ of the 4-velocity, which is

$$\frac{Du^m}{D\tau} = D_0 u^m = \partial_0 u^m + \Gamma_{00}^m u^0 = \Gamma_{00}^m = \{0, \Gamma_{00}^1, 0, 0\} = \{0, h_0, 0, 0\} , \quad (20.26)$$

the first step of which follows from equation (20.8). Similarly, a person at rest in the tetrad frame will measure the 4-velocity of an adjacent person at rest in the tetrad frame a small proper radial distance $\delta\xi^1$ away to differ by $\delta\xi^1 D_1 u^m$. The Hubble parameter of the radial flow is thus the covariant radial derivative $D_1 u^m$, which is

$$D_1 u^m = \partial_1 u^m + \Gamma_{01}^m u^0 = \Gamma_{01}^m = \{0, \Gamma_{01}^1, 0, 0\} = \{0, h_1, 0, 0\} . \quad (20.27)$$

Confined to the $(\gamma_0 \gamma_1)$ -plane (that is, considering only Lorentz transformations in the (tr) -plane, which is to say radial Lorentz boosts), the acceleration h_0 and Hubble parameter h_1 constitute the components of a tetrad-frame 2-vector $h_n = \{h_0, h_1\}$:

$$h_n = \Gamma_{10n} . \quad (20.28)$$

The Riemann tensor, equations (20.30) below, involves covariant derivatives $D_m h_n$ of h_n . These should be interpreted either as 4D covariant derivatives of the 4-vector $h_n \equiv \{h_0, h_1, 0, 0\}$ with zero angular parts, or equivalently as 2D covariant derivatives $D_m^{(2)} h_n$ confined to the $(\gamma_0 \gamma_1)$ -plane. The contraction $h^n h_n = -h_0^2 + h_1^2$ is a scalar with respect to radial Lorentz boosts.

Since h_1 is a kind of radial Hubble parameter, it can be useful to define a corresponding radial scale factor λ by

$$h_1 \equiv \partial_0 \ln \lambda . \quad (20.29)$$

The scale factor λ is the same as the λ in the comoving line-element of equation (20.21). This is true because h_1 is a tetrad connection and therefore coordinate gauge-invariant, and the line-element (20.21) is related to the line-element (20.1) being considered by a coordinate transformation $r \rightarrow r_\times$ that leaves the tetrad unchanged.

20.6 Riemann, Einstein, and Weyl tensors

The nonvanishing components of the tetrad-frame Riemann tensor R_{klmn} corresponding to the spherical line-element (20.1) are

$$R_{0101} = D_1 h_0 - D_0 h_1 , \quad (20.30a)$$

$$R_{0202} = R_{0303} = -\frac{1}{r} D_0 \beta_0 , \quad (20.30b)$$

$$R_{1212} = R_{1313} = -\frac{1}{r} D_1 \beta_1 , \quad (20.30c)$$

$$R_{0212} = R_{0313} = -\frac{1}{r} D_0 \beta_1 = -\frac{1}{r} D_1 \beta_0 , \quad (20.30d)$$

$$R_{2323} = \frac{2M}{r^3} , \quad (20.30e)$$

where D_m denotes the covariant derivative as usual. The nonvanishing components of the tetrad-frame Ricci tensor R_{km} are

$$R_{00} = R_{0101} + 2R_{0202} , \quad (20.31a)$$

$$R_{11} = -R_{0101} + 2R_{1212} , \quad (20.31b)$$

$$R_{01} = 2R_{0212} , \quad (20.31c)$$

$$R_{22} = R_{33} = -R_{0202} + R_{1212} + R_{2323} , \quad (20.31d)$$

whence

$$R_{00} = D_1 h_0 - D_0 h_1 - \frac{2}{r} D_0 \beta_0 , \quad (20.32a)$$

$$R_{11} = -D_1 h_0 + D_0 h_1 - \frac{2}{r} D_1 \beta_1 , \quad (20.32b)$$

$$R_{01} = -\frac{2}{r} D_0 \beta_1 = -\frac{2}{r} D_1 \beta_0 , \quad (20.32c)$$

$$R_{22} = R_{33} = \frac{1}{r} D_0 \beta_0 - \frac{1}{r} D_1 \beta_1 + \frac{2M}{r^3} . \quad (20.32d)$$

The Ricci scalar is

$$R = -2D_1 h_0 + 2D_0 h_1 + \frac{4}{r} D_0 \beta_0 - \frac{4}{r} D_1 \beta_1 + \frac{4M}{r^3} . \quad (20.33)$$

The nonvanishing components of the tetrad-frame Einstein tensor G^{km} are

$$G^{00} = 2R_{1212} + R_{2323} , \quad (20.34a)$$

$$G^{11} = 2R_{0202} - R_{2323} , \quad (20.34b)$$

$$G^{01} = -2R_{0212} , \quad (20.34c)$$

$$G^{22} = G^{33} = R_{0101} + R_{0202} - R_{1212} , \quad (20.34d)$$

whence

$$G^{00} = \frac{2}{r} \left(-D_1\beta_1 + \frac{M}{r^2} \right) , \quad (20.35a)$$

$$G^{11} = \frac{2}{r} \left(-D_0\beta_0 - \frac{M}{r^2} \right) , \quad (20.35b)$$

$$G^{01} = \frac{2}{r} D_0\beta_1 = \frac{2}{r} D_1\beta_0 , \quad (20.35c)$$

$$G^{22} = G^{33} = D_1h_0 - D_0h_1 + \frac{1}{r} (D_1\beta_1 - D_0\beta_0) . \quad (20.35d)$$

The nonvanishing components of the tetrad-frame Weyl tensor C_{klmn} are

$$\frac{1}{2} C_{0101} = -C_{0202} = -C_{0303} = C_{1212} = C_{1313} = -\frac{1}{2} C_{2323} = C , \quad (20.36)$$

where C is the Weyl scalar (the spin 0 component of the Weyl tensor),

$$C \equiv \frac{1}{6} (R_{0101} - R_{0202} + R_{1212} - R_{2323}) = \frac{1}{6} (G^{00} - G^{11} + G^{22}) - \frac{M}{r^3} . \quad (20.37)$$

20.7 Einstein equations

The tetrad-frame Einstein equations

$$G^{km} = 8\pi T^{km} \quad (20.38)$$

imply that

$$\begin{pmatrix} G^{00} & G^{01} & 0 & 0 \\ G^{01} & G^{11} & 0 & 0 \\ 0 & 0 & G^{22} & 0 \\ 0 & 0 & 0 & G^{33} \end{pmatrix} = 8\pi T^{km} = 8\pi \begin{pmatrix} \rho & f & 0 & 0 \\ f & p & 0 & 0 \\ 0 & 0 & p_{\perp} & 0 \\ 0 & 0 & 0 & p_{\perp} \end{pmatrix} \quad (20.39)$$

where $\rho \equiv T^{00}$ is the proper energy density, $f \equiv T^{01}$ is the proper radial energy flux, $p \equiv T^{11}$ is the proper radial pressure, and $p_{\perp} \equiv T^{22} = T^{33}$ is the proper transverse pressure. Proper here means as measured by a person at rest in the tetrad frame.

20.8 Choose your frame

So far the radial motion of the tetrad frame has been left unspecified. Any arbitrary choice can be made. For example, the tetrad frame could be chosen to be at rest,

$$\beta_0 = 0 , \quad (20.40)$$

as in the Schwarzschild or Reissner-Nordström line-elements. Alternatively, the tetrad frame could be chosen to be in free-fall,

$$h_0 = 0 , \quad (20.41)$$

as in the Gullstrand-Painlevé line-element. For situations where the spacetime contains matter, one natural choice is the **centre-of-mass** frame, defined to be the frame in which the energy flux f is zero

$$G^{01} = 8\pi f = 0 . \quad (20.42)$$

Whatever the choice of radial tetrad frame, tetrad-frame quantities in different radial tetrad frames are related to each other by a radial Lorentz boost.

20.9 Interior mass

Equations (20.35b) with the middle expression of (20.35c), and (20.35a) with the final expression of (20.35c), respectively, along with the definition (20.11) of the interior mass M , and the Einstein equations (20.39), imply (note that $D_m M = \partial_m M$ since M is a scalar)

$$p = \frac{1}{\beta_0} \left(-\frac{1}{4\pi r^2} \partial_0 M - \beta_1 f \right) , \quad (20.43a)$$

$$\rho = \frac{1}{\beta_1} \left(\frac{1}{4\pi r^2} \partial_1 M - \beta_0 f \right) . \quad (20.43b)$$

In the centre-of-mass frame, $f = 0$, these equations reduce to

$$\partial_0 M = -4\pi r^2 \beta_0 p , \quad (20.44a)$$

$$\partial_1 M = 4\pi r^2 \beta_1 \rho . \quad (20.44b)$$

Equations (20.44) amply justify the interpretation of M as the interior mass. The first equation (20.44a) can be written

$$\frac{dM}{dr} = -4\pi r^2 p , \quad (20.45)$$

where $dM/dr = \partial_0 M / \partial_0 r$ is the total derivative of the mass M with respect to radius r along the path of the matter, in the centre-of-mass frame. Equation (20.45) can be recognized as an expression of the first law of thermodynamics,

$$dE + p dV = 0 , \quad (20.46)$$

with mass-energy E equal to M and volume V equal to $\frac{4}{3}\pi r^3$. The second equation (20.44b) can be written, since $\partial_1 = \beta_1 \partial / \partial r$, equation (20.4),

$$\frac{\partial M}{\partial r} = 4\pi r^2 \rho , \quad (20.47)$$

which looks exactly like the Newtonian relation between interior mass M and density ρ . Equation (20.47) is the Hamiltonian constraint for spherically symmetric spacetimes.

Actually, the apparently Newtonian equation (20.47) is deceiving. The total mass-energy dM in a radial shell should be distinguished from the proper mass-energy dm of the shell in its own frame. The proper 3-volume element d^3r in the centre-of-mass tetrad frame is given by¹, equation (15.86),

$$d^3r = e d^3x^{r\theta\phi} = \frac{r^2 \sin \theta dr d\theta d\phi}{\beta_1} , \quad (20.48)$$

where $e = |e^a{}_\alpha|$ is the determinant of the the 3×3 spatial vierbein matrix. Thus the proper 3-volume element $dV \equiv d^3r$ of a radial shell of width dr is

$$dV = \frac{4\pi r^2 dr}{\beta_1} . \quad (20.49)$$

Consequently the proper mass-energy dm associated with the proper density ρ in a proper radial volume element dV is

$$dm = \rho dV = \frac{4\pi r^2 \rho dr}{\beta_1} , \quad (20.50)$$

whereas the total mass-energy dM from equation (20.47) is

$$dM = \rho 4\pi r^2 dr = \beta_1 \rho dV . \quad (20.51)$$

The factor β_1 can be interpreted as the energy per unit mass of the matter,

$$\beta_1 = \frac{dM}{dm} . \quad (20.52)$$

The difference between the total and proper mass-energy

$$dM - dm = (\beta_1 - 1)\rho dV \quad (20.53)$$

can be interpreted as a combination of the kinetic and gravitational energy of the matter.

20.10 Energy-momentum conservation

Covariant conservation of the Einstein tensor $D_m G^{mn} = 0$ implies conservation of energy-momentum $D_m T^{mn} = 0$. The transverse component, $n = 2, 3$, of the conservation equations vanish identically. The remaining two nontrivial equations represent conservation of energy and of radial momentum, and are

$$D_m T^{m0} = \partial_0 \rho + \frac{2\beta_0}{r}(\rho + p_\perp) + h_1(\rho + p) + \left(\partial_1 + \frac{2\beta_1}{r} + 2h_0\right)f = 0 , \quad (20.54a)$$

$$D_m T^{m1} = \partial_1 p + \frac{2\beta_1}{r}(p - p_\perp) + h_0(\rho + p) + \left(\partial_0 + \frac{2\beta_0}{r} + 2h_1\right)f = 0 . \quad (20.54b)$$

¹ The same conclusion follows from considering the spherical line element (20.1). In the tetrad frame, by construction $dr - \alpha\beta_0 dt = 0$, and the proper time satisfies $d\tau = \alpha dt$. At constant proper time, the proper radial distance is dr/β_1 , from the line element (20.1).

In the centre-of-mass frame, $f = 0$, these energy-momentum conservation equations reduce to

$$\partial_0 \rho + \frac{2\beta_0}{r}(\rho + p_\perp) + h_1(\rho + p) = 0, \quad (20.55a)$$

$$\partial_1 p + \frac{2\beta_1}{r}(p - p_\perp) + h_0(\rho + p) = 0. \quad (20.55b)$$

In a general situation where the mass-energy is a sum over several individual components x ,

$$T^{mn} = \sum_{\text{species } x} T_x^{mn}, \quad (20.56)$$

the individual mass-energy components x of the system each satisfy an energy-momentum conservation equation of the form

$$D_m T_x^{mn} = F_x^n, \quad (20.57)$$

where F_x^n is the flux of energy into component x . Einstein's equations enforce energy-momentum conservation of the system as a whole, so the sum of the energy fluxes must be zero

$$\sum_{\text{species } x} F_x^n = 0. \quad (20.58)$$

20.10.1 First law of thermodynamics

For an individual species x , the energy conservation equation (20.54a) in the centre-of-mass frame of the species, $f_x = 0$, can be written

$$D_m T_x^{m0} = \partial_0 \rho_x + (\rho_x + p_{\perp x}) \partial_0 \ln r^2 + (\rho_x + p_x) \partial_0 \ln \lambda_x = F_x^0, \quad (20.59)$$

where λ_x is the radial “scale factor,” equation (20.29), in the centre-of-mass frame of the species (the scale factor is different in different frames). Equation (20.59) can be recognized as an expression of the first law of thermodynamics for a volume element V of species x , in the form

$$V^{-1} \left[\partial_0 (\rho_x V) + p_{\perp x} V_r \partial_0 V_\perp + p_x V_\perp \partial_0 V_r \right] = F_x^0, \quad (20.60)$$

with transverse volume (area) $V_\perp \propto r^2$, radial volume (width) $V_r \propto \lambda_x$, and total volume $V \propto V_\perp V_r$. The flux F_x^0 on the right hand side is the heat per unit volume per unit time going into species x . If the pressure of species x is isotropic, $p_{\perp x} = p_x$, then equation (20.60) simplifies to

$$V^{-1} \left[\partial_0 (\rho_x V) + p_x \partial_0 V \right] = F_x^0, \quad (20.61)$$

with volume $V \propto r^2 \lambda_x$.

20.11 Structure of the Einstein equations

The spherically symmetric spacetime under consideration is described by 3 vierbein coefficients, α , β_0 , and β_1 . However, some combination of the 3 coefficients represents a gauge freedom, since the spherically symmetric spacetime has only two physical degrees of freedom. As commented in §20.8, various gauge-fixing choices can be made, such as choosing to work in the centre-of-mass frame, $f = 0$.

Equations (20.35) give 4 equations for the 4 nonvanishing components of the Einstein tensor. The two expressions for G^{01} are identical when expressed in terms of the vierbein and vierbein derivatives, so are not distinct equations. Conservation of energy-momentum of the system as a whole is built in to the Einstein equations, a consequence of the Bianchi identities, so 2 of the Einstein equations are effectively equivalent to the energy-momentum conservation equations (20.54). If the matter equations are arranged to satisfy energy-momentum conservation, as they should, then 2 of the Einstein equations are redundant, and can be dropped.

This leaves 2 independent Einstein equations to describe the 2 physical degrees of freedom of the spacetime. The 2 equations may be taken to be the evolution equations (20.35c) and (20.35b) for the velocity β_0 and energy per unit mass β_1 ,

$$\boxed{D_0\beta_0 = -\frac{M}{r^2} - 4\pi r p}, \quad (20.62a)$$

$$\boxed{D_0\beta_1 = 4\pi r f}, \quad (20.62b)$$

which are valid for any choice of tetrad frame, not just the centre-of-mass frame. The covariant derivatives on the left hand side of equations (20.62) are more explicitly

$$D_0\beta_0 = \partial_0\beta_0 - h_0\beta_1, \quad D_0\beta_1 = \partial_0\beta_1 - h_0\beta_0, \quad (20.63)$$

where h_0 is the proper radial acceleration, equation (20.24).

Equations (20.62) can be taken to be the fundamental equations governing the gravitational field in spherically symmetric spacetimes. It is these equations that are responsible (to the extent that equations may be considered responsible) for the strange internal structure of Reissner-Nordström black holes, and for mass inflation. The coefficient β_0 equals the coordinate radial 4-velocity $dr/d\tau = \partial_0 r = \beta_0$ of the tetrad frame, equation (20.5), and thus equation (20.62a) can be regarded as giving the proper radial acceleration $D^2 r / D\tau^2 = D\beta_0 / D\tau = D_0\beta_0$ of the tetrad frame as measured by a person who is in free-fall and instantaneously at rest in the tetrad frame. If the acceleration is measured by an observer who is continuously at rest in the tetrad frame (as opposed to being in free-fall), then the proper acceleration is $\partial_0\beta_0 = D_0\beta_0 + h_0\beta_1$. The presence of the extra term $h_0\beta_1$, proportional to the proper acceleration h_0 actually experienced by the observer continuously at rest in the tetrad frame, reflects the principle of equivalence of gravity and acceleration.

The right hand side of equation (20.62a) can be interpreted as the radial gravitational force, which consists of two terms. The first term, $-M/r^2$, looks like the familiar Newtonian gravitational force, which is attractive (negative, inward) in the usual case of positive mass M . The second term, $-4\pi r p$, proportional to the radial pressure p , is what makes spherical spacetimes in general relativity interesting. In a Reissner-Nordström black

hole, the negative radial pressure produced by the radial electric field produces a radial gravitational repulsion (positive, outward), according to equation (20.62a), and this repulsion dominates the gravitational force at small radii, producing an inner horizon. In mass inflation, the (positive) radial pressure of relativistically counter-streaming outgoing and ingoing streams just above the inner horizon dominates the gravitational force (inward), and it is this that drives mass inflation.

Like the second half of a vaudeville act, the second Einstein equation (20.62b) also plays an indispensable role. The energy per unit mass $\beta_1 \equiv \partial_1 r$ on the left hand side is the proper radial gradient of the circumferential radius r measured by a person at rest in the tetrad frame. The sign of β_1 determines which way an observer at rest in the tetrad frame thinks is “outwards,” the direction of larger circumferential radius r . A positive β_1 means that the observer thinks the outward direction points away from the black hole, while a negative β_1 means that the observer thinks the outward direction points towards from the black hole. Outside the outer horizon β_1 is necessarily positive, because β_m must be spacelike there. But inside the horizon β_1 may be either positive or negative. A tetrad frame can be defined as “ingoing” if the proper radial gradient β_1 is positive, and “outgoing” if β_1 is negative. In the Reissner-Nordström geometry, ingoing geodesics have positive energy, and outgoing geodesics have negative energy. However, the definition of outgoing or ingoing based on the sign of β_1 is general; there is no need for a timelike Killing vector such as would be necessary to define the (conserved) energy of a geodesic.

Equation (20.62b) shows that the proper rate of change $D_0\beta_1$ in the radial gradient β_1 measured by an observer who is in free-fall and instantaneously at rest in the tetrad frame is proportional to the radial energy flux f in that frame. But ingoing observers (β_1 positive) tend to see energy flux pointing away from the black hole (f positive), while outgoing observers (β_1 negative) tend to see energy flux pointing towards the black hole (f negative). Thus the change in β_1 tends to be in the same direction as β_1 , amplifying β_1 whatever its sign.

Exercise 20.2 Birkhoff’s theorem. Prove Birkhoff’s theorem from equations (20.62). Birkhoff’s theorem states that any spherically symmetric spacetime that is devoid of energy-momentum between some inner and outer radii is Schwarzschild between those radii.

Exercise 20.3 Spherically symmetric spacetimes in arbitrary dimensions. Generalize the solution for spherically symmetric spacetimes to arbitrary spacetime dimensions $N \geq 3$. The case of $N = 2$ spacetime dimensions was addressed in §11.19.

Solution. The line-element (20.1) is the same, with do^2 interpreted as the line-element of a unit sphere in $N-2$ dimensions. The interior mass $M(t, r)$ is defined by, generalizing equation (20.11),

$$1 - \frac{2M}{r^{N-3}} \equiv \beta_m \beta^m = -\beta_0^2 + \beta_1^2. \quad (20.64)$$

The tetrad-frame connections (20.23a)–(20.23d) with at least one time (0) or radial (1) index remain the same, but the angular connections Γ_{aba} with $a > b \geq 2$ generalize to those of a unit $(N-2)$ -sphere. The Riemann components (20.30a)–(20.30d) with at least one time (0) or radial (1) index remain the same, but

the angular components generalize to

$$R_{abab} = \frac{2M}{r^{N-3}} \quad (2 \leq a < b) . \quad (20.65)$$

The nonvanishing components of the Einstein tensor G^{km} are, generalizing equations (20.35),

$$G^{00} = \frac{N-2}{r} \left(-D_1\beta_1 + \frac{(N-3)M}{r^{N-2}} \right) , \quad (20.66a)$$

$$G^{11} = \frac{N-2}{r} \left(-D_0\beta_0 - \frac{(N-3)M}{r^{N-2}} \right) , \quad (20.66b)$$

$$G^{01} = \frac{N-2}{r} D_0\beta_1 = \frac{N-2}{r} D_1\beta_0 , \quad (20.66c)$$

$$G^{aa} = D_1h_0 - D_0h_1 + \frac{N-3}{r} (D_1\beta_1 - D_0\beta_0) - \frac{2(N-4)M}{r^{N-1}} \quad (a \geq 2) . \quad (20.66d)$$

The nonvanishing components of the Weyl tensor are, generalizing equations (20.36),

$$\frac{1}{N-2} C_{0101} = -C_{0a0a} = C_{1a1a} = -\frac{N-3}{2} C_{abab} = C \quad (2 \leq a < b) , \quad (20.67)$$

where the Weyl scalar C (the spin 0 component of the Weyl tensor) is, generalizing equations (20.37),

$$C \equiv (N-3) \left[\frac{2}{(N-1)(N-2)} (G^{00} - G^{11} + G^{22}) - \frac{M}{r^{N-1}} \right] . \quad (20.68)$$

The nonvanishing Einstein equations are, generalizing equations (20.39),

$$\{G^{00}, G^{01}, G^{11}, G^{aa}\} = \kappa_N \{\rho, f, p, p_\perp\} , \quad (20.69)$$

where κ_N is the generalization of the gravitational coupling parameter $\kappa_4 \equiv 8\pi G$ to N spacetime dimensions (see Exercise 16.14). The Hamiltonian constraint is, generalizing equation (20.47),

$$\frac{\partial M}{\partial r} = \frac{\kappa_N}{N-2} r^{N-2} \rho . \quad (20.70)$$

The equations governing the evolution of spherical spacetimes can be taken to be, generalizing equations (20.62),

$$D_0\beta_0 = -\frac{(N-3)M}{r^{N-2}} - \frac{\kappa_N}{N-2} rp , \quad (20.71a)$$

$$D_0\beta_1 = \frac{\kappa_N}{N-2} rf , \quad (20.71b)$$

the covariant derivatives on the left hand side of which are given more explicitly by equations (20.63).

Concept question 20.4 Naked singularities in spherical spacetimes? A singularity forms at zero radius, $r = 0$, when an apparent horizon develops there, that is, when space starts falling into $r = 0$ at the speed of light. Can geodesics emerge from such a singularity? A singularity from which geodesics can emerge is called a **naked singularity**. **Answer.** The surprising answer is yes, naked singularities can occur

in spherical spacetimes. To see that this conclusion is surprising, consider the following “proof” that naked singularities do not exist. The proof relies on the assumption that the interior mass M and radial pressure p are both positive, or more precisely, that $M/r^3 + 4\pi p$ is positive; this is certainly a reasonable physical assumption for real black holes. As seen in Exercise 20.1, outgoing and ingoing radial null geodesics in a spherical spacetime follow $dr/dt = \alpha(\beta_0 \pm \beta_1)$, equation (20.14). An apparent horizon forms when the outgoing null geodesic ceases to move outward, $\beta_0 + \beta_1 = 0$. The outgoing and ingoing null geodesics bound the future lightcone emerging from the apparent horizon: all radial geodesics, timelike or lightlike, must lie inside or on the lightcone, so that $dr/dt \leq 0$ for all radial geodesics at an apparent horizon, with $dr/dt = 0$ for the outgoing radial null geodesic. But the Einstein equation (20.62a), which is valid in any frame arbitrarily Lorentz-boosted in the radial direction, shows that β_0 , which equals $dr/d\tau$ in that Lorentz-boosted frame, must decrease along any geodesic, as long as $M/r^3 + 4\pi p$ is positive. Thus once $dr/d\tau$ is zero or negative along a geodesic, it cannot become positive. In particular, this holds true at zero radius, $r = 0$: as long as $M/r^3 + 4\pi p$ is positive, once $dr/d\tau$ is negative at $r = 0$, indicating the appearance of a singularity, then $dr/d\tau$ cannot become positive, and therefore no light ray can emerge from the singularity.

The foregoing “proof” that naked singularities cannot exist in spherical spacetimes is flawed because the infall velocity β_0 can be multi-valued at the point at zero radius where a singularity first forms. Section 20.16 gives an explicit example for the case of spherically symmetric collapse of pressureless dust.

20.11.1 Comment on the lapse α

Whereas the Einstein equations (20.62) give evolution equations for the vierbein coefficients β_0 and β_1 , there is no evolution equation for the vierbein coefficient α , the lapse. Indeed, the Einstein equations involve the lapse α only through the connections h_m , equations (20.23a) and (20.23b), and thus only as the radial derivative $\partial \ln \alpha / \partial r$, equations (20.24) and (20.25). This reflects the fact that, even after the tetrad frame is fixed, there is still a coordinate freedom $t \rightarrow t'(t)$ in the choice of coordinate time t . Under such a gauge transformation, α transforms as $\alpha \rightarrow \alpha' = f(t) \alpha$ where $f(t) = \partial t / \partial t'$ is an arbitrary function of coordinate time t . Only the radial derivative $\partial \ln \alpha / \partial r$ is independent of this coordinate gauge freedom, and thus the tetrad-frame Einstein equations depend, through h_m , only on this radial derivative, not on α itself.

These results are consistent with the arguments in §16.15.1 and §17.2.3 that the lapse α can be treated as a gauge variable, arbitrarily adjustable by a coordinate transformation of the time coordinate.

A possible gauge choice is to set $\alpha = 1$ everywhere. According to equation (20.24), this choice requires that the proper acceleration in the tetrad-frame vanish, $h_0 = 0$, that is, the tetrad-frame is everywhere in free fall, as for example in the Gullstrand-Painlevé line-element. I like to think of a free-fall frame as being realised physically by tracer “dark matter” particles that free-fall radially (from zero velocity, typically) at infinity, and stream freely, without interacting, through any actual matter that may be present.

20.12 Comparison to ADM (3+1) formulation

The line-element (20.1) is in ADM form with lapse α , shift $\alpha\beta_0$, and spatial metric

$$g_{\alpha\beta} = \text{diag}(1/\beta_1^2, r^2, r^2 \sin^2\theta) . \quad (20.72)$$

The nonvanishing components of the acceleration $K_a \equiv \Gamma_{a00}$ and of the extrinsic curvature $K_{ab} \equiv \Gamma_{a0b}$ are

$$K_1 = h_0 , \quad (20.73a)$$

$$K_{11} = h_1 , \quad K_{22} = K_{33} = \frac{\beta_0}{r} . \quad (20.73b)$$

20.13 Spherical electromagnetic field

The internal structure of a charged black hole resembles that of a rotating black hole because the negative pressure (tension) of the radial electric field produces a gravitational repulsion analogous to the centrifugal repulsion in a rotating black hole. Since it is much easier to deal with spherical than rotating black holes, it is common to use charge as a surrogate for rotation in exploring black holes.

20.13.1 Electromagnetic field

The assumption of spherical symmetry means that any electromagnetic field can consist only of a radial electric field (in the absence of magnetic monopoles). The only nonvanishing components of the electromagnetic field F_{mn} are then

$$-F_{01} = F_{10} = E = \frac{Q}{r^2} , \quad (20.74)$$

where E is the radial electric field, and $Q(t, r)$ is the interior electric charge. Equation (20.74) can be regarded as defining what is meant by the electric charge Q interior to radius r at time t .

20.13.2 Maxwell's equations

A radial electric field automatically satisfies the two source-free Maxwell equations. For the radial electric field (20.74), the other two Maxwell's equations, the sourced ones (16.34), are

$$\partial_1 Q = 4\pi r^2 q , \quad (20.75a)$$

$$\partial_0 Q = -4\pi r^2 j , \quad (20.75b)$$

where $q \equiv j^0$ is the proper electric charge density and $j \equiv j^1$ is the proper radial electric current density in the tetrad frame.

20.13.3 Electromagnetic energy-momentum tensor

For the radial electric field (20.74), the electromagnetic energy-momentum tensor (16.151) in the tetrad frame is the diagonal tensor

$$T_e^{mn} = \frac{Q^2}{8\pi r^4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (20.76)$$

The radial electric energy-momentum tensor is independent of the radial motion of the tetrad frame, which reflects the fact that the electric field is invariant under a radial Lorentz boost. The energy density ρ_e and radial and transverse pressures p_e and $p_{\perp e}$ of the electromagnetic field are the same as those from a spherical charge distribution with interior electric charge Q in flat space

$$\rho_e = -p_e = p_{\perp e} = \frac{Q^2}{8\pi r^4} = \frac{E^2}{8\pi}. \quad (20.77)$$

The nonvanishing components of the covariant derivative $D_m T_e^{mn}$ of the electromagnetic energy-momentum (20.76) are

$$D_m T_e^{m0} = \partial_0 \rho_e + \frac{4\beta_0}{r} \rho_e = -\frac{Q}{4\pi r^4} \partial_0 Q = -\frac{jQ}{r^2} = -jE, \quad (20.78a)$$

$$D_m T_e^{m1} = \partial_1 p_e + \frac{4\beta_1}{r} p_e = -\frac{Q}{4\pi r^4} \partial_1 Q = -\frac{qQ}{r^2} = -qE. \quad (20.78b)$$

The first expression (20.78a), which gives the rate of energy transfer out of the electromagnetic field as the current density j times the electric field E , is the same as in flat space. The second expression (20.78b), which gives the rate of transfer of radial momentum out of the electromagnetic field as the charge density q times the electric field E , is the Lorentz force on a charge density q , and again is the same as in flat space.

20.14 General relativistic stellar structure

Even with the assumption of spherical symmetry, it is by no means easy to solve the system of partial differential equations that comprise the Einstein equations coupled to mass-energy of various kinds. However, the system simplifies in some cases.

One simple case is that of a system that is not only spherically symmetric but also static, such as a star. In this case all time derivatives can be taken to vanish, $\partial/\partial t = 0$, and, since the centre-of-mass frame coincides with the rest frame, it is natural to choose the tetrad frame to be at rest, $\beta_0 = 0$. The Einstein equation (20.62b) then vanishes identically, while the Einstein equation (20.62a) becomes

$$h_0 \beta_1 = \frac{M}{r^2} + 4\pi r p, \quad (20.79)$$

which expresses the proper acceleration h_0 in the rest frame in terms of the familiar Newtonian gravitational

force M/r^2 plus a term $4\pi r p$ proportional to the radial pressure. The radial pressure p , if positive as is the usual case for a star, enhances the inward gravitational force, helping to destabilize the star. Because β_0 is zero, the interior mass M given by equation (20.11) reduces to

$$1 - 2M/r = \beta_1^2 . \quad (20.80)$$

When equations (20.79) and (20.80) are substituted into the momentum equation (20.55b), and if the pressure is taken to be isotropic, so $p_\perp = p$, the result is the **Oppenheimer-Volkov equation** for general relativistic hydrostatic equilibrium

$$\boxed{\frac{\partial p}{\partial r} = -\frac{(\rho + p)(M + 4\pi r^3 p)}{r^2(1 - 2M/r)}} . \quad (20.81)$$

In the Newtonian limit $p \ll \rho$ and $M \ll r$ this goes over to (with units restored)

$$\frac{\partial p}{\partial r} = -\rho \frac{GM}{r^2} , \quad (20.82)$$

which is the usual Newtonian equation of spherically symmetric hydrostatic equilibrium.

Exercise 20.5 Constant density star. Shortly after communicating to Einstein his celebrated solution, Schwarzschild (1916) sent Einstein a second letter describing the solution for a constant density star. By adjoining the interior solution to his exterior solution, Schwarzschild had a consistent solution with no troubling “singularity” at its horizon.

In a spherically symmetric static spacetime, Einstein’s equations reduce to an equation for the mass M interior to r

$$\frac{dM}{dr} = 4\pi r^2 \rho , \quad (20.83)$$

and to the Volkov-Oppenheimer equation of hydrostatic equilibrium (20.81).

1. **Interior mass.** Suppose that the density ρ is constant. From equation (20.83) obtain an expression for the interior mass M as a function of radius r and the density ρ . [Hint: This is easy.]
2. **Hydrostatic equilibrium.** Given your expression for M , show that the Volkov-Oppenheimer equation (20.81) rearranges to

$$\int_{p_c} \frac{dp}{(\rho + p)(\rho + 3p)} = - \int_0^r \frac{4\pi r dr}{3 - 8\pi r^2 \rho} \quad (20.84)$$

where p_c is the central pressure, where the radius is zero, $r = 0$.

3. **Solve.** Integrate equation (20.84). From the integral evaluated at the edge of the star, where the pressure is zero, $p = 0$, and the radius is the stellar radius, $r = R_\star$, argue that

$$\frac{\rho + 3p_c}{\rho + p_c} = \sqrt{\frac{1}{1 - 2M_\star/R_\star}} \quad (20.85)$$

where $M_\star \equiv \frac{4}{3}\pi\rho R_\star^3$ is the total mass of the star.

4. **Limits.** From the condition that the central pressure be positive and finite, $0 < p_c < \infty$, deduce that

$$0 < \frac{2M_\star}{R_\star} < \frac{8}{9} . \quad (20.86)$$

5. **Comment.** Comment on what equation (20.86) implies physically. [Hint: What is the Schwarzschild radius?]
-

20.15 Freely-falling dust without shell-crossing

Another case where the spherically symmetric equations simplify is that of neutral, radially freely falling, pressureless matter, at least as long as shells of matter do not cross each other. Pressureless matter is commonly referred to as “dust” in the literature. The collapse of a uniform sphere of dust was first solved by Oppenheimer and Snyder (1939). The formalism of freely falling dust is applied in §20.16 to illustrate the formation of a naked singularity.

It is natural to choose the tetrad frame to be the rest frame of the freely falling dust. In the dust rest frame, the energy flux and pressure vanish, $f = p = p_\perp = 0$. The geodesic equation for the freely falling dust implies that the proper acceleration vanishes, $h_0 = 0$, equation (20.26).

The equations admit two integrals of motion. The first integral of motion is the interior mass M , which equation (20.44a) shows is constant, $\partial_0 M = 0$, along the path of the freely falling dust.

The second integral of motion is β_1 , as follows from the second of the 2 Einstein equations (20.62). Since the acceleration vanishes, the covariant time derivative coincides with the directed time derivative, $D_0 = \partial_0$. The 2 Einstein equations (20.62) are then

$$\partial_0 \beta_0 = -\frac{M}{r^2} , \quad (20.87a)$$

$$\partial_0 \beta_1 = 0 . \quad (20.87b)$$

The second equation (20.87b) shows that β_1 is constant as claimed, an integral of motion along the path of the freely falling dust. The first equation (20.87a), in combination with the definition (20.11) of the interior mass M and the constancy of β_1 , recovers the constancy of M . The definition (20.11) of the interior mass M implies that the radial velocity $\beta_0 \equiv dr/d\tau$ of the freely falling dust is (the minus sign assumes infalling dust)

$$\beta_0 = -\sqrt{\beta_1^2 - 1 + 2M/r} . \quad (20.88)$$

Comparing this to the solution $u^r \equiv dr/d\tau$ of radially free-falling particles in a Schwarzschild geometry of mass M , equation (7.36), shows that β_1 may be interpreted as the energy E per unit mass that the freely falling dust would have if there were no further matter (i.e. the geometry were Schwarzschild) outside the radius of the dust. This interpretation of β_1 is consistent its earlier interpretation as energy per mass, equation (20.52).

As discussed in §20.11.1, in a free-fall tetrad the lapse α can be set equal to unity everywhere, $\alpha = 1$. This

corresponds to setting the time coordinate t equal to, up to a shell-dependent constant, the proper time τ attached to the freely falling dust. The relation between time t and radius r along the path of the dust is obtained by integrating the equation for $\beta_0 \equiv dr/d\tau$,

$$t - t_M = \tau = \int_0^r \frac{dr}{\beta_0} = \int_0^r \frac{dr}{-\sqrt{\beta_1^2 - 1 + 2M/r}} , \quad (20.89)$$

where the proper time τ is fixed to zero at the time t_M when the shell collapses to zero radius. The condition that shells of positive density collapse to zero radius without crossing requires that the collapse time t_M be an increasing function of interior mass M . A parametric solution for the radius r of the freely falling dust is, with $\kappa \equiv 1 - \beta_1^2$,

$$r = 2M \begin{cases} \kappa^{-1} \sin^2(\kappa^{1/2}\eta/2) & |\beta_1| < 1 \\ \eta^2/4 & |\beta_1| = 1 \\ |\kappa|^{-1} \sinh^2(|\kappa|^{1/2}\eta/2) & |\beta_1| > 1 \end{cases} , \quad \tau = M \begin{cases} \kappa^{-3/2} [\kappa^{1/2}\eta - \sin(\kappa^{1/2}\eta)] & |\beta_1| < 1 \\ \eta^3/6 & |\beta_1| = 1 \\ |\kappa|^{-3/2} [\sinh(|\kappa|^{1/2}\eta) - |\kappa|^{1/2}\eta] & |\beta_1| > 1 \end{cases} , \quad (20.90)$$

where η is negative, going to zero as the dust radius r collapses to 0. Bound dust, $|\beta_1| < 1$, reaches a maximum radius at $|\kappa|^{1/2}\eta = -\pi$.

It is possible to consider the situation of outgoing dust inside the horizon, for which β_1 is negative. However, there is a coordinate singularity in the line-element (20.1) at $\beta_1 = 0$, and care needs to be taken interpreting solutions where β_1 passes through zero. The coordinate singularity may be removed by transforming to a time coordinate different from the free-fall time coordinate. The conclusion is that trajectories with different signs of β_1 belong to distinct pieces of spacetime that abut along the $\beta_1 = 0$ trajectory.

The relation between energy density ρ and the interior mass M is determined by equation (20.47). The initial conditions must be set up to satisfy this equation, but the evolution equations guarantee that equation (20.47) holds thereafter. The equation is a constraint equation: it is the Hamiltonian constraint. An explicit expression for the proper (centre-of-mass) density ρ at time t and radius r is

$$\rho = -\frac{1}{4\pi r^2 \beta_0 \partial t / \partial M|_r} , \quad (20.91)$$

where the time t is given as a function of M (and $\beta_1(M)$) and r by equation (20.89), $t(M, r) = t_M + \int dr/\beta_0$. The proper pressure vanishes, as it must for freely falling dust.

Exercise 20.6 Oppenheimer-Snyder collapse. Solve the Oppenheimer and Snyder (1939) problem of the spherical collapse of a uniform density sphere of pressureless matter that starts from zero velocity at infinity.

20.16 Naked singularities in dust collapse

Christodoulou (1984) initiated the study of the formation of naked singularities in spherically symmetric collapse of dust. Christodoulou showed that if the collapsing dust were sufficiently centrally concentrated,

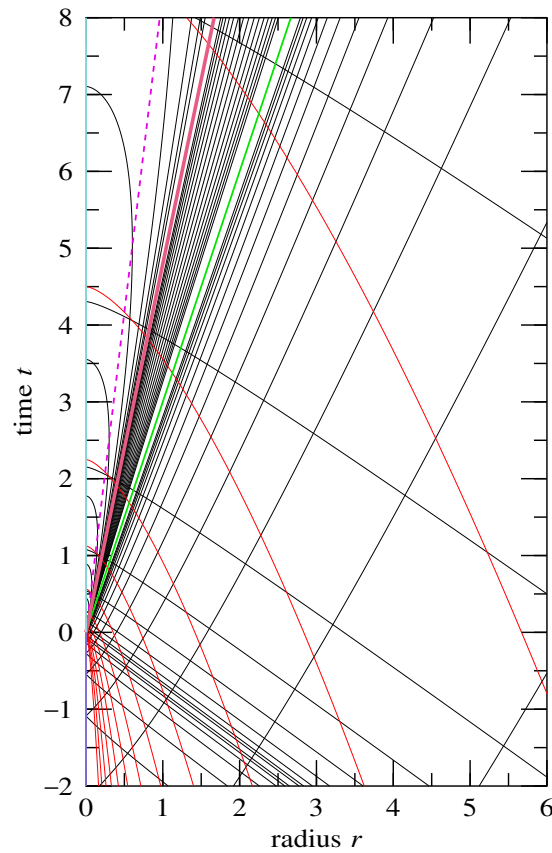


Figure 20.1 Spacetime diagram illustrating the formation of a naked singularity in self-similar collapse of dust, for $a = 18$, equation (20.94). Infalling (red) lines show trajectories of infalling dust, which are also contours of constant interior mass M . Approximately diagonal (black) lines show outgoing and ingoing radial null geodesics. Contours are drawn at intervals of factors of 2. A singularity (cyan) forms where the first shell of mass collapses to zero radius. The naked singularity is the point at the origin $\{t, r\} = \{0, 0\}$ where the singularity first forms. The apparent horizon (dashed pink line) is the locus of points where outgoing null rays turn around. The true horizon (thick pink line) divides outgoing null rays that do not and do reach infinity. In the region of spacetime between the true horizon (thick pink line) and the Cauchy horizon (thick green line), outgoing null rays emanate from the naked singularity and extend to infinity. The apparent, true, and Cauchy horizons are all straight lines emanating from the naked singularity at the origin.

then the point at which the singularity first formed would be visible to the outside world, a “naked” singularity. The appearance of naked singularities in spherical collapse of dust is generic, requiring only that the collapsing dust be sufficiently centrally concentrated.

Since the appearance of naked singularities is generic, it suffices to illustrate the situation in a simple case. One simplifying assumption is that the dust falls from zero velocity at infinity, so that $\beta_1 = 1$, in which case

the infall velocity β_0 of dust shells is (with the index on β_0 dropped for brevity)

$$\beta \equiv \beta_0 = -\sqrt{\frac{2M}{r}} . \quad (20.92)$$

Integrating equation (20.89) with $\beta_1 = 1$ gives the relation between the radius r and time t along the trajectory of a shell with interior mass M ,

$$\frac{2}{3}r^{3/2} = \sqrt{2M}(t_M - t) , \quad (20.93)$$

where t_M , a function of mass M , is the time at which the shell collapses to zero radius, $r = 0$.

A second simplifying assumption is self-similarity (see §20.18). In the present case, self-similar solutions occur when the collapse time t_M is proportional to the interior mass M ,

$$t_M = aM , \quad (20.94)$$

with a some positive dimensionless constant. Given the self-similar assumption (20.94), the relation (20.93) between the radius r and time t reduces to a cubic in the infall velocity β ,

$$a\beta^3 - \frac{2t}{r}\beta + \frac{4}{3} = 0 . \quad (20.95)$$

Equation (20.95) shows that the infall velocity β is constant along lines $t/r = \text{constant}$, that is, along straight lines emanating from the origin at $\{t, r\} = \{0, 0\}$. The infall velocity β varies from 0 at $t/r = -\infty$, to $-\infty$ at $t/r = +\infty$. The line-element is Gullstrand-Painlevé, equation (7.27), with β the real negative solution of the cubic (20.95). The proper pressure in the tetrad frame is zero (as it should be for dust), and the proper density ρ is

$$\rho = \frac{1}{4\pi r^2} \frac{\beta^2}{2 - 3a\beta^3} , \quad (20.96)$$

which is positive everywhere.

Radial outgoing (+) and ingoing (−) null rays passing through the infalling dust follow

$$\frac{dr}{dt} = \beta \pm 1 . \quad (20.97)$$

Equation (20.97) for null geodesics can be recast as a differential equation between r and β , which integrates to

$$r \propto \exp \left[\int \frac{2(\beta \pm 1)(2 - 3a\beta^3) d\beta}{\beta(\pm 4 - 2\beta \pm 3a\beta^3 + 3a\beta^4)} \right] . \quad (20.98)$$

The integrand in equation (20.98) is a rational function of β , so is integrable in terms of elementary functions. Special sets of null geodesics occur where the integrand has poles. At poles, null geodesics follow $\beta = \text{constant}$, corresponding to straight lines emanating from the origin. For outgoing null geodesics (+ in equation (20.98)), the quartic denominator $4 - 2\beta + 3a\beta^3 + 3a\beta^4$ has two real roots at $\beta < 0$ provided that the positive constant a exceeds the threshold value

$$a \geq \frac{26}{3} + 5\sqrt{3} \approx 17.3 . \quad (20.99)$$

For values of a exceeding the threshold (20.99), there is a naked singularity at the origin. The more negative of the two real roots (smaller radius) marks the location of the true horizon, while the less negative (larger radius) marks the so-called Cauchy horizon. Radial outgoing null rays inside the true horizon turn around and fall to the spacelike singularity, never reaching infinity. Radial outgoing null rays between the true and Cauchy horizons propagate from the naked singularity to infinity.

In mathematics, a Cauchy horizon is defined to be the boundary of predictability. In the present case, the naked singularity at the origin is considered to be a source of unpredictability, since the direction in which geodesics emerge from the naked singularity is ambiguous, not determined uniquely by the direction of geodesics impinging on it.

Figure 20.1 is a spacetime diagram that illustrates the formation of a naked singularity in self-similar collapse of dust for the case $a = 18$, which slightly exceeds the threshold (20.99). The roots of the quartic in this case are

$$\begin{aligned} \beta &= -0.791475, & t/r &= 4.79558 & \text{true horizon,} \\ \beta &= -\frac{2}{3}, & t/r &= 3 & \text{Cauchy horizon.} \end{aligned} \quad (20.100)$$

All outgoing null geodesics inside the true horizon, $\beta < -0.791475$, in due course turn around and fall to zero radius, $r = 0$. Outgoing null geodesics between the true and Cauchy horizons, $-0.791475 < \beta < -\frac{2}{3}$, start at the naked singularity at the origin and reach infinity. Outgoing null geodesics outside the Cauchy horizon, $\beta > -\frac{2}{3}$, start at zero radius before the singularity has formed, and propagate to infinity. The apparent horizon, where outgoing null rays turn around, $\beta = -1$, occurs at $t/r = \frac{25}{3}$.

The naked singularity in spherical dust collapse has the property that future-directed geodesics can emerge from it in some directions but not in others. This is a generic feature of naked singularities in general relativity.

20.16.1 Are naked singularities important?

As might be imagined, there is a diversity of opinion regarding the importance of naked singularities in general relativity. One school of thought holds that singularities that are hidden behind horizons (clothed singularities) have no effect on outside observers, and in that sense do not matter, at least to the outside observer. From this perspective naked singularities are important precisely because they can affect an outside observer. This seems to me a somewhat anthropocentric point of view. It may be that no human ever falls into a black hole; but in the cosmos objects fall into black holes all the time. Singularity theorems, Chapter 18, indicate that general relativity fails inside black holes (more generally, wherever a trapped surface has formed). The question of what physics replaces general relativity where it fails is profound, regardless of whether humans can see it.

The possible appearance of naked singularities in gravitational collapse offers a potential window to physics beyond general relativity. However, the collapse of a real black hole is one of the most violent events in observational astronomy, attended by supernovae and gamma-ray bursts. It is moot whether the signal from a naked singularity, whatever it might be, would be discernible against the cacophony of astrophysical processes.

20.17 Thin spherical shells

Sections 20.15 and 20.16 addressed pressureless matter that falls freely without shell crossing. Another problem that can be solved is that of a thin spherical shell. The shell may have internal pressure, and the spherical spacetime in which it falls need not be empty. The thin shell formalism is used in §20.17.1 to explore the evolution of a bubble of vacuum energy in empty space, a problem considered by Blau, Guendelman, and Guth (1987).

As remarked around equation (20.49), the proper radial volume element in the tetrad frame is not $4\pi r^2 dr$, but rather $4\pi r^2 dr/\beta_1$. The surface density $\hat{\rho}$, energy flux $\hat{f}$, radial pressure $\hat{p}$, and transverse pressure $\hat{p}_\perp$ of a thin shell are defined to be integrals over the proper radial element dr/β_1 ,

$$\hat{\rho} \equiv \int_-^+ \rho \frac{dr}{\beta_1}, \quad \hat{f} \equiv \int_-^+ f \frac{dr}{\beta_1}, \quad \hat{p} \equiv \int_-^+ p \frac{dr}{\beta_1}, \quad \hat{p}_\perp \equiv \int_-^+ p_\perp \frac{dr}{\beta_1}. \quad (20.101)$$

Minus the surface transverse pressure $-\hat{p}_\perp$ is called the surface tension. The Einstein equations governing the shell are obtained by equating 8π (in units, $8\pi G$) times the surface energy-momenta (20.101) to integrals of the Einstein tensor (20.35) over the proper volume of the shell. The integrals can be done by inspection: any term involving a covariant radial derivative D_1 integrates to its argument. The Einstein equations for the spherical shell in its own frame are then

$$-\frac{2[\beta_1]_-^+}{r} = 8\pi\hat{\rho}, \quad (20.102a)$$

$$0 = \frac{2[\beta_0]_-^+}{r} = 8\pi\hat{f}, \quad (20.102b)$$

$$0 = 8\pi\hat{p}, \quad (20.102c)$$

$$[h_0]_-^+ + \frac{[\beta_1]_-^+}{r} = 8\pi\hat{p}_\perp. \quad (20.102d)$$

Equation (20.102b) says that the velocity β_0 is constant across the shell, and equations (20.102b) and (20.102c) say that the radial energy flux $\hat{f}$ and radial pressure $\hat{p}$ vanish in the shell's own frame, which makes physical sense.

The Riemann, Ricci, and Einstein tensors are defined in terms of derivatives of the tetrad connections Γ_{klm} . Unsurprisingly, integrals of these tensors over the shell are expressible in terms of $[\Gamma_{klm}]_-^+$. The set of tetrad connections that are tensors under Lorentz transformations within the shell constitute the extrinsic curvature $\hat{K}_{km}$ of the shell, defined to be the set of tetrad connections $[\Gamma_{k1m}]_-^+$ with middle index the radial index 1,

$$\hat{K}_{km} \equiv [\Gamma_{k1m}]_-^+. \quad (20.103)$$

Recall that in the ADM formalism the extrinsic curvature $K_{km} \equiv \Gamma_{k0m}$ is defined to be the set of tetrad connections with middle index the time index 0, equations (17.21). In the ADM case the time axis γ_0 is a spatial scalar, and the extrinsic curvature K_{km} is therefore a tetrad tensor with respect to spatial transformations. In the present case the radial axis γ_1 is a scalar with respect to Lorentz transformations

within the shell, and the connections $[\Gamma_{k1m}]_{-}^{+}$ with middle index the radial index 1 form a tetrad tensor with respect to Lorentz transformations within the shell. The Ricci tensor $\hat{R}_{km} \equiv \int R_{km} dr / \beta_1$ integrated over the shell, with indices k, m running over 0, 2, 3, equals minus the extrinsic curvature of the shell,

$$\hat{R}_{km} = -\hat{K}_{km} . \quad (20.104)$$

The Einstein tensor $\hat{G}_{km}$ integrated over the shell, again with indices k, m running over 0, 2, 3, is then

$$\hat{G}_{km} = \hat{R}_{km} - \frac{1}{2} \eta_{km} \hat{R} . \quad (20.105)$$

One can confirm that the Einstein tensor (20.105) recovers the left hand sides of the Einstein equations (20.102a) and (20.102d) for the surface density and transverse pressure $\hat{\rho}$ and $\hat{p}_{\perp}$.

The proper mass-energy $\hat{m}$ of the shell is, equation (20.102a),

$$\hat{m} \equiv 4\pi r^2 \hat{\rho} = \int_{-}^{+} \rho \frac{4\pi r^2 dr}{\beta_1} = -r[\beta_1]_{-}^{+} . \quad (20.106)$$

The proper mass-energy $\hat{m}$ of the shell is to be distinguished from the total mass-energy $\hat{M}$ in the shell,

$$\hat{M} \equiv [M]_{-}^{+} = \int_{-}^{+} \rho 4\pi r^2 dr = -\frac{r[\beta_1^2]_{-}^{+}}{2} , \quad (20.107)$$

the final expression of which follows from the definition (20.11) of interior mass M and the fact that the velocity β_0 is constant across the shell, equation (20.102b). The proper mass $\hat{m}$ of the shell is related to the interior mass M by

$$\hat{m} = \int_{-}^{+} \frac{dM}{\beta_1} = -r[\beta_1]_{-}^{+} , \quad (20.108)$$

in agreement with equation (20.106). The ratio $\hat{M}/\hat{m}$ of total to proper mass-energy in the shell is

$$\frac{\hat{M}}{\hat{m}} = \frac{[\beta_1^2]_{-}^{+}}{2[\beta_1]_{-}^{+}} = \frac{\beta_1^{-} + \beta_1^{+}}{2} = \bar{\beta}_1 , \quad (20.109)$$

the average $\bar{\beta}_1$ of the energies per unit mass $\beta_1^{\pm}$ either side of the shell. The energies per unit mass $\beta_1^{\pm}$ either side of the shell are

$$\beta_1^{\pm} = \bar{\beta}_1 \pm \frac{1}{2}[\beta_1]_{-}^{+} = \frac{\hat{M}}{\hat{m}} \mp \frac{\hat{m}}{2r} . \quad (20.110)$$

The definition (20.11) of interior mass, along with the expressions (20.110) for $\beta_1^{\pm}$, implies that average interior mass $\bar{M}$ of the shell is

$$\bar{M} \equiv \frac{M^{-} + M^{+}}{2} = \frac{r}{2} \left(1 + \beta_0^2 - \frac{(\beta_1^{-})^2 + (\beta_1^{+})^2}{2} \right) = \frac{r(1 + \beta_0^2 - \bar{\beta}_1^2)}{2} - \frac{\hat{m}^2}{8r} . \quad (20.111)$$

The Einstein equation (20.102b) implies that the shell velocity β_0 is constant across the shell. The definition (20.11) of interior mass implies expressions for the velocity β_0 in terms of the interior masses $M^{\pm}$ and

energies per unit mass $\beta_1^\pm$ either side of the shell, and equation (20.111) supplies a second expression for β_0 in terms of the mean interior mass $\bar{M}$ and the mean energy per unit mass $\bar{\beta}_1$,

$$\beta_0 = \sqrt{(\beta_1^\pm)^2 - 1 + \frac{2M^\pm}{r}} \quad (20.112a)$$

$$= \sqrt{\bar{\beta}_1^2 - 1 + \frac{2\bar{M}}{r} + \frac{\hat{m}^2}{4r^2}} . \quad (20.112b)$$

The sign of the velocity β_0 is + for outfalling, – for infalling.

Evolution equations for the various energies per unit mass β_1 and for the velocity β_0 follow from evolution equations for the proper mass $\hat{m}$ of the shell and for the various interior masses. The Einstein equation (20.62b) in the centre-of-mass frame, $f = 0$, is $\partial_0\beta_1 - h_0\beta_0 = 0$, which with the Einstein equations (20.102) of the shell implies

$$0 = \partial_0[\beta_1]_-^\pm - \beta_0[h_0]_-^\pm = -4\pi \left(\partial_0(r\hat{\rho}) + \beta_0(\hat{\rho} + 2\hat{p}_\perp) \right) = -4\pi r \left(\partial_0\hat{\rho} + \frac{2\beta_0}{r}(\hat{\rho} + \hat{p}_\perp) \right) . \quad (20.113)$$

Equation (20.113) implies that the proper mass-energy $\hat{m}$ of the shell evolves as

$$\partial_0\hat{m} + 8\pi r\hat{p}_\perp\beta_0 = 0 , \quad (20.114)$$

which looks like the first law of thermodynamics in the form $\partial_0\hat{m} + \hat{p}_\perp\partial_0A = 0$ where $A \equiv 4\pi r^2$ is the proper area of the shell. If the equation of state of the shell is $\hat{p}_\perp/\hat{\rho} = \hat{w}$, then $\hat{\rho} \propto r^{-2(1+\hat{w})}$ and $\hat{m} \propto r^{-2\hat{w}}$. In particular, if the equation of state of the shell is vacuum, $\hat{w} = -1$, then the density $\hat{\rho}$ of the shell is constant. The interior masses $M^\pm$ evolve according to the Einstein equation (20.44a),

$$\partial_0M^\pm + 4\pi r^2p^\pm\beta_0 = 0 . \quad (20.115)$$

The two equations (20.115) may be recast as evolution equations for the total mass $\hat{M} \equiv [M]_-^\pm$ of the shell and for the average interior mass $\bar{M}$,

$$\partial_0\hat{M} + 4\pi r^2[p]_-^\pm\beta_0 = 0 , \quad (20.116a)$$

$$\partial_0\bar{M} + 4\pi r^2\bar{p}\beta_0 = 0 , \quad (20.116b)$$

where $[p]_-^\pm$ and $\bar{p} \equiv \frac{1}{2}(p^- + p^+)$ are respectively the difference and average of the external radial pressures $p^\pm$ on the shell. The evolution (20.114) of the proper mass-energy $\hat{m}$ of the shell depends on its equation of state $\hat{p}_\perp/\hat{\rho}$, while the evolution (20.116) of the total mass-energies $\hat{M}$ and $\bar{M}$ depends on the external pressures $p^\pm$.

Usually it is most straightforward to solve the evolution equations (20.114) and (20.116) for the various masses $\hat{m}$, $\hat{M}$, and $\bar{M}$, and then to infer the energies per unit mass $\beta_1^\pm$ and their average $\bar{\beta}_1$ from equation (20.110), and the velocity β_0 either any of the two equivalent equations (20.112). However, evolution equations for β_0 , $\beta_1^\pm$, and $\bar{\beta}_1$ can be deduced directly, either from the evolution equations for the masses, or

from the Einstein equations (20.62),

$$\partial_0 \beta_0 = \beta_1^\pm h_0^\pm - \frac{M^\pm}{r^2} - 4\pi r p^\pm \quad (20.117a)$$

$$= \bar{\beta}_1 \bar{h}_0 - \frac{\bar{M}}{r^2} - \frac{\hat{m}^2}{4r^3} - 4\pi r \bar{p} - \frac{2\pi \hat{m} \hat{p}_\perp}{r}, \quad (20.117b)$$

$$\partial_0 \beta_1^\pm = \beta_0 h_0^\pm, \quad (20.117c)$$

$$\partial_0 \bar{\beta}_1 = \partial_0 \frac{\hat{M}}{\hat{m}} = \beta_0 \bar{h}_0, \quad (20.117d)$$

where $h_0^\pm$ are proper accelerations experienced by observers in the tetrad frame on each side of the shell, and $\bar{h}_0$ is their average,

$$h_0^\pm = \bar{h}_0 \pm 2\pi(\hat{\rho} + 2\hat{p}_\perp), \quad \bar{h}_0 = -\frac{[p]_\pm^+}{\hat{\rho}} + \frac{2\bar{\beta}_1 \hat{p}_\perp}{r\hat{\rho}}. \quad (20.118)$$

Exercise 20.7 Free fall of a thin, pressureless, spherical shell in vacuo. Solve for the evolution of a thin, pressureless, spherical shell that free falls in vacuo from rest at infinity. This exercise provides the mathematics behind the calculations reported in §7.28.

Solution. In the particular case of a pressureless shell, $\hat{p}_\perp = 0$, freely falling in vacuo, $p^- = p^+ = 0$, the evolution equations (20.114) and (20.114) imply that the proper, total, and mean interior masses $\hat{m}$, $\hat{M}$, and $\bar{M}$ are all constants. The constancy of $\hat{m}$ and $\hat{M}$ implies the constancy of $\bar{\beta}_1$, equation (20.109) (but not of $\beta_1^\pm$, equation (20.110)). The infall velocity β_0 is given by equation (20.112b). If the shell free falls from rest at infinity, then $\bar{\beta}_1 = 1$, and the total mass-energy of the shell equals its proper mass-energy, $\hat{M} = \hat{m}$, equation (20.109). The infall velocity is

$$\beta_0 = -\sqrt{\frac{2\bar{M}}{r} + \frac{\hat{M}^2}{4r^2}}. \quad (20.119)$$

If the spherical shell is falling towards an object of mass $M_\bullet$ (a black hole, say), then the interior masses inside and outside the shell are $M^- = M_\bullet$ and $M^+ = M_\bullet + \hat{M}$, and the mean interior mass is $\bar{M} = M_\bullet + \frac{1}{2}\hat{M}$.

20.17.1 A bubble of vacuum

Blau, Guendelman, and Guth (1987) explored the scenario of a spherically symmetric bubble of positive vacuum energy density ρ_Λ that evolves in otherwise empty space. As chronicled by Merali (2017), Blau et al. were motivated at least in part by the question of what might happen to a mote of vacuum energy that was somehow created in empty space. Could such a mote develop into an inflating universe? If so, would the new universe expand out and destroy the surrounding space? Or would the new universe create its own spacetime?

The geometry is de Sitter inside the bubble, Schwarzschild outside. The interface between the de Sitter and empty spaces cannot itself be empty, because the finite pressure of the vacuum and the zero pressure of empty space do not balance. For simplicity, Blau et al. modelled the interface as a thin spherical shell, which they assumed itself had a vacuum equation of state, $\hat{p}_\perp = -\hat{\rho}$, a so-called domain wall. The interior masses $M^\pm$ inside (–) and outside (+) the shell, and the proper mass $\hat{m}$ of the shell, are then

$$M^- = \frac{4}{3}\pi r^3 \rho_\Lambda, \quad \hat{m} = 4\pi r^2 \hat{\rho}, \quad M^+ = M, \quad (20.120)$$

where the vacuum density ρ_Λ , shell density $\hat{\rho}$, and the mass M are all constants. The mass M is the mass of the bubble perceived by an observer in the empty space outside the bubble. Equations (20.110) for the energy per unit mass $\beta_1^\pm$ inside (–) and outside (+) the shell, and their average $\bar{\beta}_1$, become

$$\beta_1^\pm = \frac{M - \frac{4}{3}\pi r^3(\rho_\Lambda \pm 6\pi\hat{\rho}^2)}{4\pi r^2 \hat{\rho}}, \quad \bar{\beta}_1 = \frac{M - \frac{4}{3}\pi r^3 \rho_\Lambda}{4\pi r^2 \hat{\rho}}. \quad (20.121)$$

The velocity β_0 , equation (20.112), is

$$\beta_0 = \sqrt{(\beta_1^\pm)^2 - \Delta^\pm}, \quad (20.122)$$

where $\Delta^\pm \equiv 1 - 2M^\pm/r$ is the horizon function either side of the shell. The energies per unit mass $\beta_1^\pm$ and $\bar{\beta}_1$ are respectively zero at radii $r_1^\pm$ and r_1 given by

$$r_1^\pm = \left(\frac{M}{\frac{4}{3}\pi(\rho_\Lambda \pm 6\pi\hat{\rho}^2)} \right)^{1/3}, \quad r_1 = \left(\frac{M}{\frac{4}{3}\pi\rho_\Lambda} \right)^{1/3}. \quad (20.123)$$

For positive mass M and vacuum density ρ_Λ , the radii r_1^+ and r_1 are always positive. The radius r_1^- is positive or negative as ρ_Λ is larger or smaller than $6\pi\hat{\rho}^2$. Blau et al. argue that if the vacuum is GUT scale, then it might be expected that $\rho_\Lambda \sim m_{\text{GUT}}^4$ and $\hat{\rho} \sim m_{\text{GUT}}^3$ in Planck units $c = G = \hbar = 1$, in which case $\hat{\rho}^2/\rho_\Lambda \sim m_{\text{GUT}}^2$, which is small compared to 1 if the GUT scale is significantly smaller than the Planck scale, $m_{\text{GUT}} \ll 1$. In that case all of $r_1^\pm$ and r_1 are positive, and they are ordered

$$0 < r_1^+ \lesssim r_1 \lesssim r_1^-. \quad (20.124)$$

It is convenient to work in radial units $r_1^+ = 1$ (Blau et al. introduce the dimensionless variable $z \equiv r/r_1^+$). In these units, the energy per unit mass β_1^+ , equation (20.121), is

$$\beta_1^+ = \frac{1}{\sqrt{-E}} \left(\frac{1 - r^3}{r^2} \right), \quad (20.125)$$

and the velocity β_0 , equation (20.122), satisfies

$$-E\beta_0^2 + V = E, \quad (20.126)$$

where $V(r)$ is a dimensionless effective potential and the (negative) constant E is an effective dimensionless

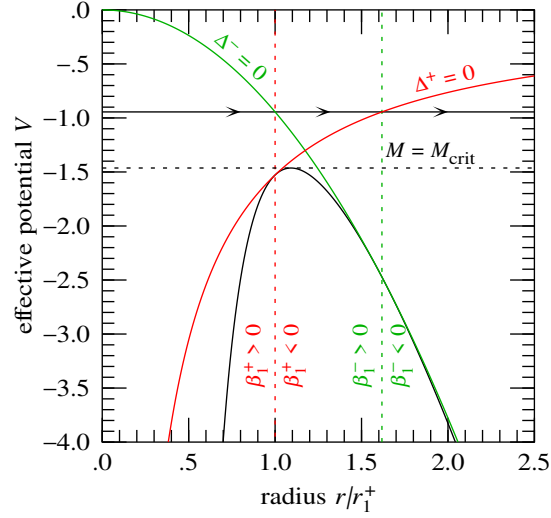


Figure 20.2 Effective potential V , equation (20.127a), of a spherical shell sandwiched between a bubble of vacuum and empty space, as a function of the radius r in units $r_1^+ = 1$, for the case $r_{\max} = 1.092$. A more realistic case would have $r_{\max}$ closer to 1, but the larger choice of $r_{\max}$ brings out the behaviour more clearly. The arrowed horizontal line is an illustrative unbound trajectory of a shell that expands from zero radius to infinity. Unbound trajectories occur for $M > M_{\text{crit}}$. The radii where the trajectory passes through the de Sitter horizon $\Delta^- = 0$, the Schwarzschild horizon $\Delta^+ = 0$, and the places where the energies per unit mass $\beta_1^\pm$ pass through zero, are marked. The choice $r_{\max} = (1 - \sqrt{5}/2 + \sqrt{17/4 - \sqrt{5}})^{1/3} = 1.092$ is a special value for which there happens to be a special trajectory, the one shown, where the locations $\Delta^- = 0$ and $\beta_1^+ = 0$ coincide, and also the locations $\Delta^+ = 0$ and $\beta_1^- = 0$ coincide. This is similar to Figure 6 of Blau, Guendelman, and Guth (1987).

energy,

$$V \equiv - \left(\frac{1 - r^3}{r^2} \right)^2 - \frac{\mu}{r}, \quad (20.127a)$$

$$E \equiv - \left(\frac{\mu^2}{16\pi\hat{\rho}M} \right)^{2/3}, \quad (20.127b)$$

with the constant μ given by

$$\mu \equiv \frac{24\pi\hat{\rho}^2}{\rho_\Lambda + 6\pi\hat{\rho}^2}. \quad (20.128)$$

If M , ρ_Λ , and $\hat{\rho}$ are all positive, then the constant μ is positive, while V and E are negative. Equation (20.126) agrees with equation (5.9) of Blau et al. with the translations (there $\rightarrow$ here)

$$z \rightarrow r/r_1^+, \quad \beta_D \rightarrow \beta_1^-, \quad \beta_S \rightarrow \beta_1^+, \quad \rho_0 \rightarrow \rho_\Lambda, \quad \chi^2 \rightarrow \frac{8}{3}\pi\rho_\Lambda, \quad \gamma^2 \rightarrow \mu, \quad \sigma \rightarrow \hat{\rho}. \quad (20.129)$$

The effective potential V defined by equation (20.127a) is a hill that goes through a maximum at a value

$r = r_{\max}$ that depends on μ . Equivalently, μ depends on $r_{\max}$. Figure 20.2 illustrates the effective potential V for the case $r_{\max} = 1.092$. The value of the constant μ , and of the potential $V_{\max} \equiv V(r_{\max})$ at its maximum, are

$$\mu = \frac{2(r_{\max}^3 - 1)(r_{\max}^3 + 2)}{r_{\max}^3}, \quad V_{\max} = -\frac{3(r_{\max}^6 - 1)}{r_{\max}^4}. \quad (20.130)$$

As $6\pi\hat{\rho}^2/\rho_{\Lambda}$ varies from 0 to 1 to ∞ , the constant μ varies from 0 to 2 to 4, and the apex $r_{\max}$ of the potential varies from 1 to $2^{1/6}$ to $2^{1/3}$. The motion of the shell is bounded if $E < V_{\max}$, unbounded if $E > V_{\max}$. The critical case $E = V_{\max}$ occurs at a mass $M = M_{\text{crit}}$,

$$M_{\text{crit}} = \sqrt{\frac{(2 - r_{\max}^3)(r_{\max}^3 + 2)^3}{72\pi\rho_{\Lambda}(r_{\max}^3 + 1)^2}}. \quad (20.131)$$

If $M < M_{\text{crit}}$, then the motion is bounded, while for $M > M_{\text{crit}}$ the motion is unbounded. For vacuum densities sufficiently below the Planck scale, where $\hat{\rho}^2/\rho_{\Lambda} \ll 1$ and hence $r_{\max} \approx 1$, the critical mass is $M_{\text{crit}} \approx \sqrt{3/(32\pi\rho_{\Lambda})}$, or about 6 grams for $M_{\text{GUT}} \approx 10^{16}$ GeV.

Blau et al. were interested in the fate of a mote of vacuum that materializes at small radius and initially expands. If the mass of the mote is less than the critical mass M_{crit} , then the mote momentarily expands, but then turns around and collapses. No new universe.

If on the other hand the mass of the mote exceeds the critical mass M_{crit} , then the mote expands from zero radius to infinity. A new universe is created.

From the perspective of an observer in the pre-existing empty space, the shell materializes at zero radius, $r = 0$, with zero proper mass, $\hat{m} = 0$, but with finite total mass $\hat{M} = M$, hence infinite energy per unit mass. The outside observer sees a white hole of mass M and horizon size $2M$ suddenly come into being. The shell is inside the White Hole part of the Schwarzschild geometry, Figure 7.17. The outside observer, in the Universe part of the Schwarzschild geometry, sees the shell born at the white hole singularity, possibly with attending fireworks, and watches the shell expand into the empty space inside the white hole (the contents of a white hole are, unlike a black hole, visible to an outside observer). The shell switches from ingoing ($\beta_1^+ > 0$) to outgoing ($\beta_1^+ < 0$) inside the white hole, and, now having negative energy per unit mass β_1^+ , exits into the Parallel Universe part of the Schwarzschild geometry. The observer in the Universe sees the exiting shell redshift and dim to obscurity.

The more interesting perspective is that of an observer who rides with the shell. The shell does not expand into a pre-existing spacetime, but rather creates its own new spacetime, with both empty and de Sitter components. The shell can be conceptualized in one lower dimension as the leading circular edge of an expanding two-sided disk, on the one side of which is empty space, and on the other is de Sitter space. Looking backwards, the shell observer sees empty space at smaller radii, going back to the white hole. Looking forwards, the shell observer sees de Sitter space also at smaller radii. The forward looking observer is looking in the direction where the radius should be larger, but because the spherical shell is expanding faster than light ($\Delta^- < 0$) away from the origin of de Sitter space at $r = 0$, any light that the shell observer sees necessarily comes from behind them, at smaller radius.

An observer at the origin $r = 0$ of de Sitter space sees the shell expand away from them. Either before or

shortly after passing through the White Hole horizon into the Parallel Universe, the shell expands beyond the de Sitter horizon of the observer at the origin $r = 0$. The origin observer truly finds themselves in an inflating universe.

Key to this remarkable behaviour is the transition of the shell's total mass $\hat{M} = M - \frac{4}{3}\pi r^3 \rho_\Lambda$ from positive to negative, which happens between the times that the shell passes through the White Hole and de Sitter horizons. Does a large negative total shell mass $\hat{M}$ make sense? Recall that the total mass $\hat{M}$ includes not only rest mass but also kinetic and gravitational contributions, and the gravitational contribution can be negative. The proper mass $\hat{m} = 4\pi r^2 \hat{\rho}$ of the shell is always positive (and increasing). The mass $\frac{4}{3}\pi r^3 \rho_\Lambda$ of vacuum energy grows huge as the bubble expands, a mass balanced by the negative gravitational total mass of the shell.

Is the creation of a bubble of vacuum from a white hole singularity realistic? Nope.

20.17.2 A bubble of vacuum from a magnetic monopole

Sakai et al. (2006) argue that a more realistic origin for an inflating universe is a Grand Unified Theory (GUT) magnetic monopole ('t Hooft, 1974; Polyakov, 1974). Magnetic monopoles are predicted by GUTs, where the electromagnetic field gets knotted up in spacetime. GUT monopoles are predicted to have masses approximately $\alpha^{-1} = 137$ times the GUT mass, or about 10^{18} GeV, close to the Planck mass. No monopole has been observed in Nature, but that is not too surprising given their large mass.

Sakai et al. model the scenario using the thin shell formalism, with Reissner-Nordström geometry outside the shell, de Sitter inside. The parameters of the RN geometry are the mass M and magnetic charge $\mathcal{Q}$ of the magnetic monopole. The de Sitter geometry has positive vacuum density ρ_Λ . The shell carries all the magnetic charge of the monopole, so the monopole looks charged from the RN side, uncharged from the de Sitter side. Sakai et al. model the shell as having a constant mass $\hat{m}_\mathcal{Q}$ attributable to the rest mass of its magnetic charge, plus a constant vacuum shell density $\hat{\rho}_\lambda$. The interior masses $M^\pm$ inside (−) and outside (+) the shell, and the proper mass $\hat{m}$ of the shell, are then

$$M^- = \frac{4}{3}\pi r^3 \rho_\Lambda, \quad \hat{m} = \hat{m}_\mathcal{Q} + 4\pi r^2 \hat{\rho}_\lambda, \quad M^+ = M - \frac{\mathcal{Q}^2}{2r}, \quad (20.132)$$

where M , $\mathcal{Q}$, ρ_Λ , $\hat{m}_\mathcal{Q}$, and $\hat{\rho}_\lambda$ are all constants. The translation from Sakai et al.'s notation is (there $\rightarrow$ here)

$$\beta^\pm \rightarrow \beta_1^\pm, \quad \rho \rightarrow \rho_\Lambda, \quad \sigma_0 \rightarrow \hat{\rho}_\lambda, \quad \sigma_1 \rightarrow \hat{\rho}_\mathcal{Q} = \frac{\hat{m}_\mathcal{Q}}{4\pi r^2}. \quad (20.133)$$

An effective potential V for the shell may be defined by

$$V \equiv -\beta_0^2 = -(\beta_1^\pm)^2 + 1 - \frac{M^\pm}{r}, \quad (20.134)$$

with energies per unit mass $\beta_1^\pm$ given by equation (20.110).

The interesting parameter regime is where the RN geometry is near extremal, $\mathcal{Q} \approx M$. Parameters may be chosen such that the effective potential V , equation (20.134), has a stable or marginally stable point at

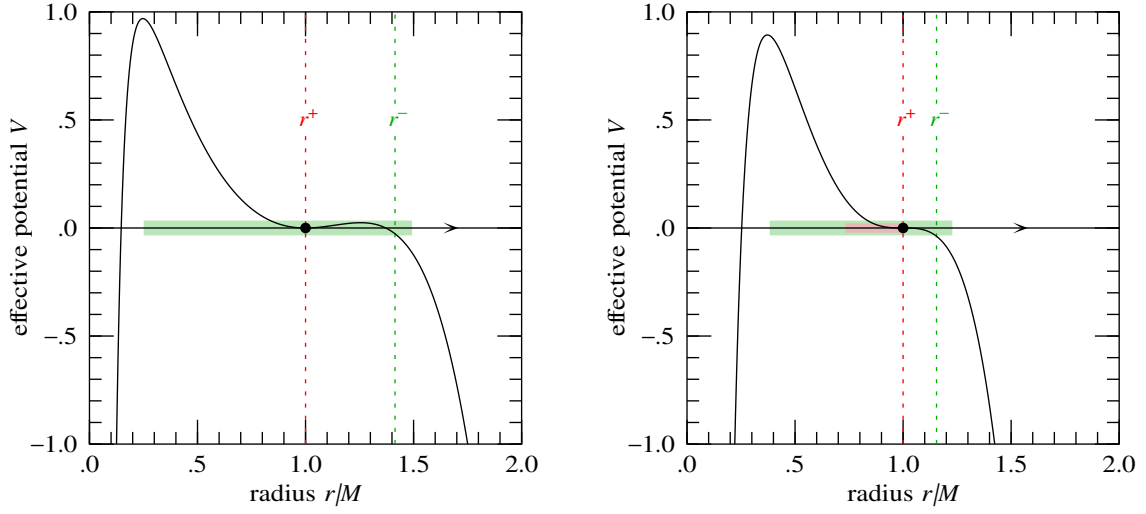


Figure 20.3 Effective potential V , equation (20.134), of a magnetically charged shell enclosing a bubble of positive energy vacuum. The parameters are given by equation (20.135). Both configurations have an extremal RN geometry $\mathcal{Q} = M$. On the left, the shell (dot) is in a classically stable configuration. On the right, the vacuum density is somewhat larger, and the configuration has marginal stability. Thick horizontal bands show the radial ranges where β_1^+ (pinkish) and β_1^- (greenish) are positive. Vertical dashed lines mark RN (red) and de Sitter (green) horizons r^+ and r^- .

zero velocity β_0 , as illustrated in Figure 20.3. The left panel of Figure 20.3 shows a stable point; the right panel a marginally stable point. The parameters of the two cases illustrated are

$$\mathcal{Q} = M, \quad H \equiv \sqrt{\frac{8}{3}\pi\rho_\Lambda} = \begin{cases} \sqrt{\frac{1}{2}} M^{-1}, \\ \sqrt{\frac{3}{4}} \end{cases}, \quad \hat{m}_\mathcal{Q} = \begin{cases} \sqrt{\frac{1}{2}} M, \\ \sqrt{\frac{1}{2}} \end{cases}, \quad \hat{\rho}_\Lambda = 0. \quad (20.135)$$

Both examples are for an extremal RN geometry, $\mathcal{Q} = M$, and in both cases the point of (marginal) stability is at the RN horizon. The first, stable, choice (left panel of Figure 20.3) is special in that not only β_0 but also β_1^+ vanishes at the point of stability. The second, marginally stable, choice (right panel of Figure 20.3) is special by virtue of its marginal stability.

The initial configuration illustrated in the left panel of Figure 20.3 is classically stable. An outside observer sees a magnetic monopole with magnetic charge $\mathcal{Q}$ equal to its mass M . The shell is located at the horizon of the extremal RN geometry. Inside the shell is vacuum with positive energy ρ_Λ .

The shell could potentially quantum tunnel out of the stable configuration, or alternatively the monopole could perhaps be perturbed out of its stable state by a collision of some kind. Or, the parameters might perhaps be tuned so that the configuration is close to or at marginal stability, as illustrated in the right panel of Figure 20.3.

Once out of the stable or marginally stable configuration, the shell starts expanding. In both cases shown

in Figure 20.3, the RN energy per unit mass β_1^+ starts at zero in the initial (marginally) stable configuration. More generally, β_1^+ can be initially positive or negative. But regardless of the initial sign, β_1^+ becomes negative as the shell expands, indicating that the shell has made its way to a Parallel Universe or Parallel Antiverse part of the RN geometry, Figure 8.6. As the shell expands, it exits the de Sitter horizon of an observer at $r = 0$.

As in the situation of a bubble of vacuum in empty space considered by Blau, Guendelman, and Guth (1987), the shell does not expand into a pre-existing spacetime, but rather creates its own new spacetime, with both RN and de Sitter components. Looking backward, an observer riding the shell sees RN spacetime at smaller radii. Looking forward, an observer riding the shell sees de Sitter spacetime, also at smaller radii. Even though the forward-looking observer is looking in the direction of larger radii, they see only smaller radii because the shell is moving superluminally outward outside the de Sitter horizon of the origin at $r = 0$.

How realistic is the scenario of the creation of an inflating universe from a GUT magnetic monopole? An object moving outwards in radius with negative RN energy per unit mass β_1^+ is necessarily in a Parallel part of the RN geometry, and must have negotiated an inner horizon where the outside universe appeared infinitely blueshifted. In the extremal cases illustrated in Figure 20.3, the inner and outer horizons coincide, and an object at rest at the horizon sees the outside universe infinitely blueshifted. In realistic situations, the diverging concentration of energy at the inner horizon drives an instability that is the principal topic of Chapter 21. Bottom line: the model is not realistic as it stands.

20.18 Self-similar spherically symmetric spacetime

A fourth way to simplify the system of spherically symmetric equations, transforming them into ordinary differential equations, is to consider self-similar solutions. The system is more complicated than that of a static system, or of freely falling dust, or of thin shells, but still straightforward.

Self-similar solutions are flexible enough to admit multiple components of energy-momentum, which may interact with each other. Self-similar solutions are especially useful for exploring the inflationary instability in the vicinity of the inner horizon of a charged spherical black hole, considered in the next Chapter 21. Charged spherical black holes are not realistic as models of real astronomical black holes, but they have inner horizons like realistic rotating black holes, so admit inflation.

20.18.1 Self-similarity

The assumption of **self-similarity** (also known as homothety, if you can pronounce it) is the assumption that the system possesses conformal time translation invariance. This implies that there exists a conformal time coordinate t such that the geometry at any one time is conformally related to the geometry at any other time, $g_{\mu\nu} = e^{2\psi t} \tilde{g}_{\mu\nu}$, where the conformal metric coefficients $\tilde{g}_{\mu\nu}(r)$ are functions only of conformal radius r , not of conformal time t . In terms of conformal coordinates $x^\mu = \{t, r, \theta, \phi\}$, the self-similar line-element is

$$ds^2 = e^{2\psi t} [\tilde{g}_{tt}(r) dt^2 + 2\tilde{g}_{tr}(r) dt dr + \tilde{g}_{rr}(r) dr^2 + e^{2r} d\theta^2] . \quad (20.136)$$

The choice e^{2r} of the coefficient of do^2 is a gauge choice of the conformal radius r , chosen here so as to bring the self-similar line-element into a form (20.140) below that resembles as far as possible the spherical line-element (20.1). The proper circumferential radius R is

$$R \equiv e^{vt+r} \quad (20.137)$$

which is to be considered as a function $R(t, r)$ of the conformal coordinates t and r . The circumferential radius R has a gauge-invariant meaning, whereas neither t nor r are independently gauge-invariant. The conformal factor R has the dimensions of length. In self-similar solutions, all quantities are proportional to some power of R , and that power can be determined by dimensional analysis. Quantities that depend only on the conformal radial coordinate r , independent of the circumferential radius R , are called dimensionless.

The fact that dimensionless quantities such as the conformal metric coefficients $\tilde{g}_{\mu\nu}(r)$ are independent of conformal time t implies that the tangent vector $\mathbf{e}_t$, which by definition satisfies

$$\frac{\partial}{\partial t} = \mathbf{e}_t \cdot \boldsymbol{\partial} , \quad (20.138)$$

is a conformal Killing vector, §7.32.4, also known as the homothetic vector. The tetrad-frame components of the conformal Killing vector $\mathbf{e}_t$ defines the tetrad-frame conformal Killing 4-vector ξ^m ,

$$\frac{\partial}{\partial t} \equiv R \xi^m \partial_m , \quad (20.139)$$

in which the factor R is introduced so as to make ξ^m dimensionless. The conformal Killing vector $\mathbf{e}_t$ is the generator of the conformal time translation symmetry, and as such it is gauge-invariant (up to a global rescaling of conformal time, $t \rightarrow at$ for some constant a). It follows that its dimensionless tetrad-frame components ξ^m constitute a tetrad 4-vector (again, up to global rescaling of conformal time).

20.18.2 Self-similar line-element

The self-similar line-element can be taken to have the same form as the spherical line-element (20.1), but with the dependence on the dimensionless conformal Killing vector ξ^m made manifest:

$$ds^2 = R^2 \left[-(\xi^0 dt)^2 + \frac{1}{\beta_1^2} (dr + \beta_1 \xi^1 dt)^2 + do^2 \right] . \quad (20.140)$$

The vierbein e^m_{μ} and inverse vierbein e_m^{μ} corresponding to the self-similar line-element (20.140) are

$$e^m_{\mu} = R \begin{pmatrix} \xi^0 & 0 & 0 & 0 \\ \xi^1 & 1/\beta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sin \theta \end{pmatrix} , \quad e_m^{\mu} = \frac{1}{R} \begin{pmatrix} 1/\xi^0 & -\beta_1 \xi^1/\xi^0 & 0 & 0 \\ 0 & \beta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/\sin \theta \end{pmatrix} . \quad (20.141)$$

It is straightforward to see that the coordinate time components of the vierbein must be $e^m_t = R \xi^m$, since $\partial/\partial t = e^m_t \partial_m$ equals $R \xi^m \partial_m$, equation (20.139).

20.18.3 Tetrad-frame scalars and vectors

Since the conformal factor R is gauge-invariant, the directed gradient $\partial_m R$ constitutes a tetrad-frame 4-vector β_m (which unlike ξ^m is independent of any global rescaling of conformal time),

$$\beta_m \equiv \partial_m R . \quad (20.142)$$

It is straightforward to check that β_1 defined by equation (20.142) is consistent with its appearance in the vierbein (20.141) provided that $R \propto e^r$ as earlier assumed, equation (20.137). The tetrad time component β_0 satisfies

$$\beta_0 = \frac{v - \beta_1 \xi^1}{\xi^0} . \quad (20.143)$$

With two distinct dimensionless tetrad 4-vectors in hand, β_m and the conformal Killing vector ξ^m , three gauge-invariant dimensionless scalars can be constructed, $\beta^m \beta_m$, $\xi^m \beta_m$, and $\xi^m \xi_m$,

$$1 - \frac{2M}{R} = \beta^m \beta_m = -\beta_0^2 + \beta_1^2 , \quad (20.144a)$$

$$v \equiv \xi^m \beta_m = \xi^0 \beta_0 + \xi^1 \beta_1 = \frac{1}{R} \frac{\partial R}{\partial t} , \quad (20.144b)$$

$$\Delta \equiv -\xi^m \xi_m = (\xi^0)^2 - (\xi^1)^2 . \quad (20.144c)$$

The M in equation (20.144a), which is essentially the same as equation (20.11), is the interior mass. Equation (20.144a) is dimensionless, which implies that the interior mass at fixed conformal radius r increases in proportion to the conformal factor, $M \propto R$. The dimensionless constant v in equation (20.144b) may be interpreted as a measure of the expansion velocity of the self-similar spacetime. Because of the freedom of a global rescaling of conformal time, it is possible to set $v = 1$ without loss of generality; but that scaling obscures the physical significance of v as an expansion rate. The choice adopted in the next Chapter, equation (21.7), is to set v equal to the rate $\dot{M}_\bullet$ of increase of the interior mass M evaluated at a specific conformal radius, taken to be the sonic point outside the horizon where the boundary conditions are established; the rate is with respect to the proper time τ_d of collisionless “dark matter” that free-falls radially from zero velocity far from the black hole,

$$v \equiv \dot{M}_\bullet \equiv \frac{dM_{\text{sonic}}}{d\tau_d} . \quad (20.145)$$

The proper time τ_d is essentially the free-fall time t_{ff} of the Gullstrand-Painlevé line-element (19.10), or equivalently T in the line-element (20.148) with $\alpha = 1$ and $\beta_1 = 1$. The dimensionless quantity Δ in equation (20.144c) is the dimensionless horizon function: horizons occur where the horizon function vanishes,

$$\Delta = 0 \quad \text{at horizons} . \quad (20.146)$$

Note that if v is rescaled, then $\Delta \propto v^2$.

Exercise 20.8 Self-similar line-element. Let T and R denote time and radius coordinates

$$T \equiv e^{vt} , \quad R \equiv e^{vt+r} . \quad (20.147)$$

Show that the self-similar line-element (20.140) in terms of T and R is

$$ds^2 = -\alpha^2 dT^2 + \frac{1}{\beta_1^2} (dR - \beta_0 dT)^2 + R^2 d\sigma^2 , \quad (20.148)$$

with lapse

$$\alpha = \frac{\xi^0 R}{vT} . \quad (20.149)$$

The line-element (20.148) is the same as the spherical line-element (20.1) with t and r in the latter relabelled T and R .

20.18.4 Self-similar diagonal line-element

The self-similar line-element (20.140) can be brought to diagonal form by a coordinate transformation to diagonal conformal coordinates $t_\times, r_\times$ (subscripted $\times$ for diagonal),

$$t \rightarrow t_\times = t + f(r) , \quad r \rightarrow r_\times = r - v f(r) , \quad (20.150)$$

which leaves unchanged the conformal factor R , equation (20.137). The resulting diagonal metric is (compare equation (20.19))

$$ds^2 = R^2 \left(-\Delta dt_\times^2 + \frac{dr_\times^2}{1 - 2M/R + v^2/\Delta} + d\sigma^2 \right) . \quad (20.151)$$

The diagonal line-element (20.151) corresponds physically to the case where the tetrad frame is at rest in the similarity frame, $\xi^1 = 0$, as can be seen by comparing it to the line-element (20.140). The frame can be called the **similarity frame**. The form of the metric coefficients in the line-element (20.151) follows from the line-element (20.140) and the gauge-invariant scalars (20.144).

The conformal Killing vector in the similarity frame is $\xi^m = \{\sqrt{\Delta}, 0, 0, 0\}$, and the 4-velocity of the similarity frame in its own frame is $u^m = \{1, 0, 0, 0\}$. Since both are tetrad 4-vectors, it follows that with respect to a general tetrad frame (20.140),

$$\xi^m = u^m \sqrt{\Delta} \quad (20.152)$$

where u^m is the 4-velocity of the similarity frame with respect to the general tetrad frame. This shows that the conformal Killing vector ξ^m in a general tetrad frame is proportional to the 4-velocity of the similarity frame through the tetrad frame. In particular, the proper 3-velocity of the similarity frame through the tetrad frame is

$$\text{proper 3-velocity of similarity frame through tetrad frame} = \frac{\xi^1}{\xi^0} . \quad (20.153)$$

In the models considered in Chapter 21, fluids generically fall inward into the black hole. The velocity of the tetrad rest frame of an infalling fluid is negative relative to the similarity frame, so the velocity ξ^1/ξ^0 of the similarity frame through the tetrad frame is positive.

In the rest frame of any fluid, the Killing vector ξ^m remains finite and continuous across horizons, where $\Delta = 0$, whereas the related 4-velocity u^m , equation (20.152), diverges at horizons. The infall velocity hits the speed of light at the outer horizon, $\xi^1/\xi^0 = 1$, both ξ^1 and ξ^0 remaining positive there (while u^m diverges). Inside the horizon, the conformal Killing vector ξ^m becomes timelike, with positive ξ^1 exceeding ξ^0 . In some models, the fluid later drops through an outgoing inner horizon, where $\xi^1/\xi^0 = -1$ with ξ^1 positive and ξ^0 negative. In general, ξ^m is lightlike at horizons,

$$\left| \frac{\xi^1}{\xi^0} \right| = 1 \quad \text{at a horizon} . \quad (20.154)$$

20.18.5 Ray-tracing line-element

It proves useful to introduce a “ray-tracing” conformal radial coordinate x related to the coordinate $r_\times$ of the diagonal line-element (20.151) by

$$dx \equiv \frac{\Delta dr_\times}{[(1 - 2M/R)\Delta + v^2]^{1/2}} . \quad (20.155)$$

In terms of the ray-tracing coordinate x , the diagonal metric (20.151) is

$$ds^2 = R^2 \left(-\Delta dt_\times^2 + \frac{dx^2}{\Delta} + do^2 \right) . \quad (20.156)$$

The line-element (20.156) defines the same similarity tetrad frame as (20.151).

20.18.6 Geodesics

Spherical symmetry and conformal time translation symmetry imply that geodesic motion in spherically symmetric self-similar spacetimes is described by a complete set of integrals of motion.

The integral of motion associated with conformal time translation symmetry can be obtained from Lagrange’s equations of motion,

$$\frac{d}{d\tau} \frac{\partial L}{\partial u^t} = \frac{\partial L}{\partial t} , \quad (20.157)$$

with effective Lagrangian $L = \frac{1}{2} g_{\mu\nu} u^\mu u^\nu$ for a particle with coordinate 4-velocity u^μ . The self-similar metric depends on the conformal time t only through the overall conformal factor $g_{\mu\nu} \propto R^2$. The derivative of the conformal factor is given by $\partial \ln R / \partial t = v$, equation (20.144b), so it follows that $\partial L / \partial t = 2vL$. For a massive particle, for which conservation of rest mass implies $g_{\mu\nu} u^\mu u^\nu = -1$, Lagrange’s equations (20.157) thus yield

$$\frac{du_t}{d\tau} = -v . \quad (20.158)$$

In the limit of zero accretion rate, $v \rightarrow 0$, equation (20.158) would integrate to give u_t as a constant, the energy per unit mass of the geodesic. But here there is conformal time translation symmetry in place of time translation symmetry, and equation (20.158) integrates to

$$u_t = -v\tau , \quad (20.159)$$

in which an arbitrary constant of integration has been absorbed into a shift in the zero point of the proper time τ . Although the above derivation was for a massive particle, it holds also for a massless particle, with the understanding that the proper time τ is constant along a null geodesic. The quantity u_t in equation (20.159) is the covariant time component of the coordinate-frame 4-velocity u^μ of the particle; it is related to the covariant components u_m of the tetrad-frame 4-velocity of the particle by

$$u_t = e^m{}_t u_m = R \xi^m u_m . \quad (20.160)$$

Without loss of generality, geodesic motion can be taken to lie in the equatorial plane $\theta = \pi/2$ of the spherical spacetime. The integrals of motion associated with conformal time translation symmetry, rotational symmetry about the polar axis, and conservation of rest mass, are, for a massive particle,

$$u_t = -v\tau , \quad u_\phi = L , \quad u_\mu u^\mu = -1 , \quad (20.161)$$

where L is the orbital angular momentum per unit rest mass of the particle. The coordinate 4-velocity $u^\mu \equiv dx^\mu/d\tau$ that follows from equations (20.161) takes its simplest form in the conformal coordinates $\{t_\times, x, \theta, \phi\}$ of the ray-tracing metric (20.156),

$$u^{t_\times} = \frac{v\tau}{R^2\Delta} , \quad u^x = \pm \frac{1}{R^2} [v^2\tau^2 - (R^2 + L^2)\Delta]^{1/2} , \quad u^\phi = \frac{L}{R^2} . \quad (20.162)$$

20.18.7 Null geodesics

The important case of a massless particle follows from taking the limit of a massive particle with infinite energy and angular momentum, $v\tau \rightarrow \infty$ and $L \rightarrow \infty$ (note that τ is constant along a null geodesic, and $v\tau$ can be treated as constant in the limit of a massive particle of infinite energy). To obtain finite results, define an affine parameter λ by $d\lambda \equiv v\tau d\tau$, and a 4-velocity in terms of it by $v^\mu \equiv dx^\mu/d\lambda$. The integrals of motion (20.161) then become, for a null geodesic,

$$v_{t_\times} = -1 , \quad v_\phi = J , \quad v_\mu v^\mu = 0 , \quad (20.163)$$

where $J \equiv L/(v\tau)$ is the (dimensionless) conformal angular momentum of the particle. The 4-velocity v^μ along the null geodesic is then, in terms of the coordinates of the ray-tracing metric (20.156),

$$v^{t_\times} = \frac{1}{R^2\Delta} , \quad v^x = \pm \frac{1}{R^2} (1 - J^2\Delta)^{1/2} , \quad v^\phi = \frac{J}{R^2} . \quad (20.164)$$

Equations (20.164) yield the shape of a null geodesic by quadrature,

$$\phi = \int \frac{J dx}{(1 - J^2\Delta)^{1/2}} . \quad (20.165)$$

Equation (20.165) shows that the shape of null geodesics in spherically symmetric self-similar spacetimes hinges on the behaviour of the dimensionless horizon function $\Delta(x)$ as a function of the dimensionless ray-tracing variable x . Null geodesics go through periapsis or apoapsis in the self-similar frame where the denominator of the integrand of (20.165) is zero, corresponding to $v^x = 0$.

In the Reissner-Nordström geometry there is a radius, the photon sphere, where photons can orbit in circles for ever. In nonstationary self-similar solutions there is no conformal radius where photons can orbit for ever (to remain at fixed conformal radius r , the photon angular momentum would have to increase in proportion to the conformal factor R). There is however a separatrix between null geodesics that do or do not fall into the black hole, and the conformal radius where this occurs can be called the photon sphere equivalent. The photon sphere equivalent occurs where the denominator of the integrand of equation (20.165) not only vanishes, $v^x = 0$, but is an extremum, which happens where the horizon function Δ is an extremum,

$$\frac{d\Delta}{dx} = 0 \quad \text{at photon sphere equivalent} . \quad (20.166)$$

20.18.8 Dimensional analysis

The spatial conformal coordinates $\{r, \theta, \phi\}$ are by definition dimensionless. The tetrad metric γ_{mn} is dimensionless, while the coordinate metric $g_{\mu\nu}$ scales as R^2 ,

$$\gamma_{mn} \propto R^0 , \quad g_{\mu\nu} \propto R^2 . \quad (20.167)$$

The vierbein $e^m{}_\mu$, and inverse vierbein $e_m{}^\mu$ equations (20.141), scale as

$$e^m{}_\mu \propto R , \quad e_m{}^\mu \propto R^{-1} . \quad (20.168)$$

The tetrad connections Γ_{kmn} and the tetrad-frame Riemann tensor R_{klmn} scale as

$$\Gamma_{kmn} \propto R^{-1} , \quad R_{klmn} \propto R^{-2} . \quad (20.169)$$

20.18.9 Variety of self-similar solutions

Self-similar solutions exist provided that the properties of the energy-momentum introduce no additional dimensional parameters. Dimensional analysis shows that the proper density ρ and radial and transverse pressure p and $p_\perp$ of any species must scale with conformal factor R as

$$\rho \propto p \propto p_\perp \propto R^{-2} . \quad (20.170)$$

The pressure-to-density ratio $w \equiv p/\rho$ of any species is dimensionless, and since the ratio can depend only on the nature of the species itself, not for example on where it happens to be located in the spacetime, it follows that the ratio w must be a constant. It is legitimate for the pressure-to-density ratio to be different in the radial and transverse directions (as it is for a radial electric field), but otherwise self-similarity requires that

$$w \equiv p/\rho , \quad w_\perp \equiv p_\perp/\rho , \quad (20.171)$$

be constants for each species. For example, $w = 1$ for an ultrahard fluid (which can mimic the behaviour of a massless scalar field (Babichev et al., 2008)), $w = 1/3$ for a relativistic fluid, $w = 0$ for pressureless cold dark matter, $w = -1$ for vacuum energy, and $w = -1$ with $w_\perp = 1$ for a radial electric field.

Self-similarity allows that the energy-momentum may consist of several distinct components, such as a relativistic fluid, plus dark matter, plus an electric field. The components may interact with each other provided that the properties of the interaction introduce no additional dimensional parameters. Dimensional analysis shows that the flux F^n of energy and momentum transferred between any two species, equation (20.57), must scale as

$$F^n \propto R^{-3} . \quad (20.172)$$

20.18.10 Electrical conductivity

The principal reason to consider charged black holes is that stationary charged black holes have inner horizons like rotating black holes, and it is easier to model spherical charged black holes than rotating black holes. The big question, explored using spherical charged black holes in the next Chapter 21, is what happens near their inner horizons? In exploring this question one should bear in mind that charge is really a surrogate for rotation.

In self-similar models, a charged black hole acquires its electrical charge from accretion of charged fluid. A charged fluid will experience a Lorentz force from the electric field, and will therefore exchange momentum with the electric field. If the fluid is nonconducting, then there is no dissipation, and the interaction between the charged fluid and electric field automatically introduces no additional dimensional parameters. However, if the charged fluid is electrically conducting, then the electrical conductivity of the fluid could potentially introduce an additional dimensional parameter, and this must not be allowed if self-similarity is to be maintained. Dimensional analysis shows that the electric charge density $q \equiv j^0$, the radial electric current $j \equiv j^1$, and the radial electric field $E \equiv Q/R^2$ scale as

$$q \propto j \propto R^{-2} , \quad E \propto R^{-1} , \quad (20.173)$$

consistent with the requirement that the flux of energy and momentum on the right hand sides of equations (20.78) scale as $F^n \propto R^{-3}$. In diffusive electrical conduction in a fluid of conductivity σ , an electric field E gives rise to a current in the fluid rest frame,

$$j = \sigma E , \quad (20.174)$$

which is just Ohm's law. Dimensional analysis then requires that the conductivity must scale as $\sigma \propto R^{-1}$. The conductivity can depend only on the intrinsic properties of the conducting fluid, and the only intrinsic property available is its density, which scales as $\rho \propto R^{-2}$. It follows that the conductivity must be proportional to the square root of the density ρ of the conducting fluid,

$$\sigma = \kappa \rho^{1/2} , \quad (20.175)$$

where κ is a dimensionless conductivity constant. The form (20.175) is required by self-similarity, and is

not necessarily realistic (although it is realistic that the conductivity increases with density). However, the conductivity (20.175) is adequate for the purpose of exploring the consequences of dissipation in simple models of black holes.

A realistic value of the electrical conductivity of a baryonic plasma at a relativistic temperature T is (Arnold, Moore, and Yaffe, 2000)

$$\sigma = \frac{C}{e^2 \ln e^{-1}} \frac{kT}{\hbar} \quad (20.176)$$

where e is the dimensionless charge of the electron, the square root of the fine-structure constant, and the factor $C \approx 15$ depends on the mix of particle species. This electrical conductivity is huge. A dimensionless measure of the conductivity (which has units 1/time) is the conductivity σ times the characteristic timescale $t_{\text{BH}} \equiv GM/c^3$ of the black hole, which is of order

$$\sigma t_{\text{BH}} \sim \frac{T}{T_{\text{BH}}} \quad (20.177)$$

where $kT_{\text{BH}} \equiv \hbar/t_{\text{BH}}$ is the characteristic temperature of the black hole (for a Schwarzschild black hole, this characteristic temperature T_{BH} is 8π times the Hawking temperature). In the astronomical situation considered here the temperature T of the plasma is huge compared to the characteristic temperature T_{BH} of the black hole. Indeed if this were not so, then mass loss by Hawking radiation would tend to compete with mass gain by accretion, an entirely different situation from the one envisaged here.

Charge is being envisaged here as a surrogate for rotation, and electrical conduction should be interpreted as a substitute for angular momentum transport. Angular momentum transport is a much weaker process than electrical conduction (if angular momentum transport were as strong as electrical conduction, then accretion disks would shed angular momentum as quickly as they shed charge, and accretion disks would not rotate). In the next Chapter 21, the conductivity is treated as a phenomenological free parameter, greatly suppressed compared to any realistic conductivity, but nevertheless possibly consistent with what might be a reasonable rate for the analogous angular momentum transport in a rotating black hole.

20.18.11 Tetrad connections

The expressions for the tetrad connections for the self-similar spacetime (20.140) are the same as those (20.23) for a general spherically symmetric spacetime. Expressions (20.24) and (20.25) for the proper radial acceleration h_0 and the radial Hubble parameter h_1 translate in the self-similar spacetime to

$$h_0 \equiv \partial_1 \ln(R \xi^0) , \quad h_1 \equiv \partial_0 \ln(R \xi^1) . \quad (20.178)$$

Comparing equations (20.178) to equations (20.24) and (20.29) shows that the lapse α and scale factor λ translate in the self-similar spacetime to

$$\alpha = R \xi^0 , \quad \lambda = R \xi^1 . \quad (20.179)$$

20.18.12 Spherical equations carry over to the self-similar case

The tetrad-frame Riemann, Weyl, and Einstein tensors in the self-similar spacetime take the same form as in the general spherical case, equations (20.30)–(20.35).

Likewise, the equations for the interior mass in §20.9, for energy-momentum conservation in §20.10, for the first law in §20.10.1, and the various equations for the electromagnetic field in §20.13, all carry through unchanged.

20.18.13 From partial to ordinary differential equations

The central simplifying feature of self-similar solutions is that they turn a system of partial differential equations into a system of ordinary differential equations.

By definition, a dimensionless quantity $A(r)$ is independent of conformal time t . It follows that the partial derivative of any dimensionless quantity $A(r)$ with respect to conformal time t vanishes,

$$0 = \frac{\partial A(r)}{\partial t} = \xi^m \partial_m A(r) = (\xi^0 \partial_0 + \xi^1 \partial_1) A(r) . \quad (20.180)$$

Consequently the directed radial derivative $\partial_1 F$ of a dimensionless quantity $A(r)$ is related to its directed time derivative $\partial_0 F$ by

$$\partial_1 A(r) = -\frac{\xi^0}{\xi^1} \partial_0 A(r) . \quad (20.181)$$

Equation (20.181) allows radial derivatives to be converted to time derivatives.

20.18.14 Integration variable

It is desirable to choose an integration variable that varies monotonically. A natural choice is the proper time τ in some tetrad frame, since this is guaranteed to increase monotonically. The 4-velocity at rest in the tetrad frame is by definition $u^m = \{1, 0, 0, 0\}$, so the proper time derivative is related to the directed conformal time derivative in the tetrad frame by $d/d\tau = u^m \partial_m = \partial_0$.

However, there is another choice of integration variable, the ray-tracing variable x defined by equation (20.155), that is not specifically tied to any tetrad frame, and that has a desirable (tetrad and coordinate) gauge-invariant meaning. The proper time derivative of any dimensionless function $A(r)$ in the tetrad frame is related to its derivative dA/dx with respect to the ray-tracing variable x by

$$\partial_0 A = u^m \partial_m A = (u^1 \partial_1)_{\text{sim}} A = -\frac{\xi^1}{R} \frac{dA}{dx} . \quad (20.182)$$

In the third expression, $(u^1 \partial_1)_{\text{sim}} A$ is $u^m \partial_m A$ expressed in the similarity frame (20.156), where the directed time and radial derivatives are $(\partial_0)_{\text{sim}} = (1/(R\sqrt{\Delta})) \partial/\partial t_{\times}$ and $(\partial_x)_{\text{sim}} = (\sqrt{\Delta}/R) \partial/\partial x$. The partial time derivative $\partial/\partial t_{\times}|_x = \partial/\partial t|_r$ vanishes acting on any dimensionless quantity $A(r)$. The last expression of (20.182) comes from $u^1_{\text{sim}} = -\xi^1/\sqrt{\Delta}$ in view of equation (20.152), the minus sign coming from the fact

that u_{sim}^1 is tetrad relative to similarity frame, while u^1 in equation (20.152) is similarity relative to tetrad frame.

In summary, the chosen integration variable is the dimensionless ray-tracing variable $-x$ (with a minus because $-x$ increases monotonically with proper time), the derivative with respect to which, acting on any dimensionless function, is related to the proper time derivative ∂_0 in any tetrad frame by

$$-\frac{d}{dx} = \frac{R}{\xi^1} \partial_0 . \quad (20.183)$$

Equation (20.183) involves ξ^1 , which is proportional to the proper velocity of the tetrad frame through the similarity frame, equation (20.154), and which therefore, being initially positive, must always remain positive in any tetrad frame attached to a fluid, as long as the fluid does not turn back on itself, as must be true for the self-similar solution to be consistent.

20.18.15 Integrals of motion

As remarked above, equation (20.180), in self-similar solutions $\xi^m \partial_m A(r) = 0$ holds for any dimensionless function $A(r)$. If both the directed derivatives $\partial_0 A(r)$ and $\partial_1 A(r)$ are known from the Einstein equations or elsewhere, then the result will be an integral of motion.

The spherically symmetric, self-similar Einstein equations admit two integrals of motion,

$$0 = R \xi^m \partial_m \beta_0 = R \beta_1 (\xi^0 h_0 + \xi^1 h_1) - \xi^0 \left(\frac{M}{R} + 4\pi R^2 p \right) + \xi^1 4\pi R^2 f , \quad (20.184a)$$

$$0 = R \xi^m \partial_m \beta_1 = R \beta_0 (\xi^0 h_0 + \xi^1 h_1) + \xi^1 \left(\frac{M}{R} - 4\pi R^2 \rho \right) + \xi^0 4\pi R^2 f . \quad (20.184b)$$

Taking ξ^1 times (20.184a) plus ξ^0 times (20.184b), and then β_0 times (20.184a) minus β_1 times (20.184b), gives

$$0 = v R (\xi^0 h_0 + \xi^1 h_1) - 4\pi R^2 [\xi^0 \xi^1 (\rho + p) - ((\xi^0)^2 + (\xi^1)^2) f] , \quad (20.185a)$$

$$0 = R \xi^m \partial_m \frac{M}{R} = -v \frac{M}{R} + 4\pi R^2 [\beta_1 \xi^1 \rho - \beta_0 \xi^0 p + (\beta_0 \xi^1 - \beta_1 \xi^0) f] . \quad (20.185b)$$

The quantities in square brackets on the right hand sides of equations (20.185) are scalars for each species x , so equations (20.185) can also be written

$$v R (\xi^0 h_0 + \xi^1 h_1) = 4\pi R^2 \sum_{\text{species } x} \xi_x^0 \xi_x^1 (\rho_x + p_x) , \quad (20.186a)$$

$$v \frac{M}{R} = 4\pi R^2 \sum_{\text{species } x} (\beta_{x,1} \xi_x^1 \rho_x - \beta_{x,0} \xi_x^0 p_x) , \quad (20.186b)$$

where the sum is over all species x , and $\beta_{x,m}$ and ξ_x^m are the 4-vectors β_m and ξ^m expressed in the rest frame of species x . Equations (20.186) are scalar equations, valid in any frame of reference.

For any fluid with equation of state $p/\rho = w = \text{constant}$, a further integral comes from considering

$$0 = R \xi^m \partial_m (R^2 p) = R [w \xi^0 \partial_0 (R^2 \rho) + \xi^1 \partial_1 (R^2 p)] , \quad (20.187)$$

and simplifying using the energy conservation equation for $\partial_0 \rho$ and the momentum conservation equation for $\partial_1 p$.

In the particular case of the electromagnetic field, equation (20.187) reduces to

$$0 = R \xi^m \partial_m \frac{Q}{R} = -v \frac{Q}{R} + 4\pi R^2 (\xi^1 q - \xi^0 j) , \quad (20.188)$$

which is valid in any radial tetrad frame.

The energy-momentum conservation equations (20.55) with fluxes (20.57) are

$$\partial_0 \rho + \frac{2\beta_0}{R} (\rho + p_\perp) + h_1 (\rho + p) = F^0 , \quad (20.189a)$$

$$\partial_1 p + \frac{2\beta_1}{R} (p - p_\perp) + h_0 (\rho + p) = F^1 . \quad (20.189b)$$

If a species is charged, then the energy flux into the charged species from the electromagnetic field is, equations (20.78),

$$F^0 = jE , \quad F^1 = qE . \quad (20.190)$$

There may be other contributions to the energy-momentum fluxes F^m if the species exchanges energy-momentum with another species, for example through collisions. Inserting equations (20.189) into equation (20.187) yields, in the centre-of-mass frame of a species,

$$(1 + w)R(\xi^1 h_0 + w \xi^0 h_1) - 2w_\perp (\xi^1 \beta_1 - w \xi^0 \beta_0) = \frac{R}{\rho} (\xi^1 F^1 + w \xi^0 F^0) . \quad (20.191)$$

Equation (20.191) rearranges to

$$R h_0 = \frac{2w_\perp \xi^1 (\xi^1 \beta_1 - w \xi^0 \beta_0) - w(1 + w) \xi^0 (\varepsilon/v) + (R/\rho) \xi^1 (\xi^1 F^1 + w \xi^0 F^0)}{(1 + w) [(\xi^1)^2 - w(\xi^0)^2]} , \quad (20.192)$$

where

$$\varepsilon \equiv 4\pi R^2 \sum_{\text{species } x} \xi_x^0 \xi_x^1 (1 + w_x) \rho_x \quad (20.193)$$

summed over all species x (including the one under consideration), where ξ_x^m is in the rest frame of species x .

20.18.16 Entropy

Substituting the self-similar expression (20.179) for the scale factor λ into the energy conservation equation (20.59) for a species in its own centre-of-mass frame gives

$$\partial_0 \ln [\rho R^{2(1+w_\perp)} (R \xi^1)^{1+w}] = \frac{F^0}{\rho} . \quad (20.194)$$

For a fluid with isotropic equation of state $w = w_\perp$, equation (20.194) becomes

$$\partial_0 \ln S = \frac{F^0}{(1+w)\rho} , \quad (20.195)$$

where S is (up to an arbitrary constant) the entropy of a comoving volume element $V \propto R^3 \xi^1$ of the fluid,

$$S \equiv R^3 \xi^1 \rho^{1/(1+w)} . \quad (20.196)$$

20.18.17 Summary of equations for accreting, self-similar, spherical, charged black holes

This section summarizes the equations used in Chapter 21 to compute the evolution of self-similar, spherical, charged black holes accreting a variety of fluids. For brevity, the index x labelling a fluid species is omitted. Equations (20.199)–(20.204) and (20.208) are valid in any tetrad frame governed by the self-similar line-element (20.140). Equations (20.197), (20.198), and (20.205)–(20.207) hold in the rest frame of the fluid in question, the frame where the energy flux f of the fluid is zero. For equations holding in the fluid rest frame, the quantities ξ^m , β_m , and h_m should be interpreted as evaluated in the fluid rest frame. Some quantities, notably v , M/R , Q/R , and Δ are (dimensionless) scalars, taking the same value in any tetrad frame. Equations (20.200)–(20.208) are dimensionless, factors of R appearing so as to make them so; for example Rh_m , $R^2\rho$, $R\sigma$ are dimensionless.

Self-similarity requires that each fluid have an equation of state with constant w and $w_\perp$, equations (20.171),

$$w \equiv p/\rho , \quad w_\perp \equiv p_\perp/\rho . \quad (20.197)$$

If the fluid is charged, then self-similarity requires that its conductivity σ be proportional to the square root of the proper energy density, equation (20.175),

$$\sigma = \kappa \rho^{1/2} , \quad (20.198)$$

with constant dimensionless conductivity coefficient κ .

The proper time τ in any tetrad frame evolves as

$$-\frac{d\tau}{dx} = \frac{R}{\xi^1} , \quad (20.199)$$

which follows from $dx/d\tau = \partial_0 x$ and equation (20.183). The circumferential radius R in any tetrad frame evolves as

$$-\frac{d \ln R}{dx} = \frac{\beta_0}{\xi^1} , \quad (20.200)$$

which follows from $dR/d\tau = \partial_0 R = \beta_0$ and equation (20.199).

The defining equations (20.178) for the proper acceleration h_0 and Hubble parameter h_1 yield equations

for the evolution of the time and radial components of the conformal Killing vector ξ^m in any tetrad frame,

$$-\frac{d\xi^0}{dx} = \beta_1 - Rh_0 , \quad (20.201a)$$

$$-\frac{d\xi^1}{dx} = -\beta_0 + Rh_1 . \quad (20.201b)$$

In the evolution equation (20.201a) for ξ^0 , equation (20.181) has been used to convert the conformal radial derivative ∂_1 to the conformal time derivative ∂_0 , and thence to $-d/dx$ by equation (20.183).

The Einstein equations (20.38) applied to the two expressions (20.35c) for G^{01} yield evolution equations for the time and radial components of the vierbein coefficients β_m in any tetrad frame,

$$-\frac{d\beta_0}{dx} = -\frac{1}{\xi^0} (\beta_1 Rh_1 + 4\pi R^2 T^{01}) , \quad (20.202a)$$

$$-\frac{d\beta_1}{dx} = \frac{1}{\xi^1} (\beta_0 Rh_0 + 4\pi R^2 T^{01}) . \quad (20.202b)$$

Again, in the evolution equation (20.202a) for β_0 , equation (20.181) has been used to convert the conformal radial derivative ∂_1 to the conformal time derivative ∂_0 . The energy flux T^{01} in equations (20.202) is the total energy flux summed over all species. The 4 evolution equations (20.201) and (20.202) for ξ^m and β_m are not independent: they are related by $\xi^m \beta_m = v$, a constant, equation (20.144b). To maintain numerical precision, it is important to avoid expressing small quantities as differences of large quantities. In practice, a suitable choice of variables to integrate proves to be $\xi^0 + \xi^1$, $\beta_0 - \beta_1$, and β_1 , each of which can be tiny in some circumstances. Starting from these variables, the following equations yield $\xi^0 - \xi^1$, along with the interior mass M and the horizon function Δ , equations (20.144a) and (20.144c), in a fashion that ensures numerical stability:

$$\xi^0 - \xi^1 = \frac{2v - (\xi^0 + \xi^1)(\beta_0 + \beta_1)}{\beta_0 - \beta_1} , \quad (20.203a)$$

$$\frac{2M}{R} = 1 + (\beta_0 + \beta_1)(\beta_0 - \beta_1) , \quad (20.203b)$$

$$\Delta = (\xi^0 + \xi^1)(\xi^0 - \xi^1) . \quad (20.203c)$$

Equation (20.203b) is numerically preferable to equation (20.186b), which can suffer loss of precision from cancellation of large quantities; equation (20.186b) can be used as a check.

The evolution equations (20.201) and (20.202) involve h_0 and h_1 . The integrals of motion considered in §20.18.15 yield explicit expressions for h_0 and h_1 not involving any derivatives. For the Hubble parameter h_1 , equation (20.186a) gives

$$Rh_1 = -\frac{\xi^0}{\xi^1} Rh_0 + \frac{\varepsilon}{v} , \quad (20.204)$$

where ε is given by equation (20.193). For the proper acceleration h_0 , a simple case is that of non-interacting (collisionless), pressureless, neutral “dark matter,” for which the acceleration vanishes,

$$h_0 = 0 \quad \text{dark matter} . \quad (20.205)$$

For a more general fluid, the integral of motion (20.192) yields an expression for h_0 . If the fluid exchanges energy-momentum only with the electromagnetic field, so that the fluxes F^m are given by equations (20.190), then the integral of motion (20.192), simplified using the integral of motion (20.188) for Q and the conductivity (20.198) in Ohm's law (20.174), reduces to

$$Rh_0 = \frac{\xi^1 \{8\pi w_\perp (\beta_1 \xi^1 - w\beta_0 \xi^0) R^2 \rho + [v + (1+w)4\pi R\sigma \xi^0] Q^2/R^2 - w(4\pi \xi^0 \varepsilon)^2/v\}}{4\pi \varepsilon [(\xi^1)^2 - w(\xi^0)^2]} . \quad (20.206)$$

Finally, equations are needed governing the evolution of the energy densities ρ of the fluids. If a fluid has isotropic equation of state, $w = w_\perp$, then the energy conservation equation translates into a conservation equation (20.195) for entropy (20.196). If the fluid exchanges energy-momentum only with the electromagnetic field, so that the flux F^0 is given by equations (20.190), then the entropy conservation equation (20.195) is

$$-\frac{d \ln S}{dx} = \frac{\sigma Q^2}{\xi^1 R^3 (1+w) \rho} . \quad (20.207)$$

The right hand side of equation (20.207) vanishes if the fluid is uncharged or nonconducting.

For the electromagnetic field, the energy conservation equation (20.78a) becomes

$$-\frac{d \ln Q}{dx} = -\frac{4\pi R\sigma}{\xi^1} . \quad (20.208)$$

If there is more than one charged conducting fluid, then the right hand side of equation (20.208) should be summed over the charged conducting fluids. Equation (20.208) says that (free) energy coming out of the electromagnetic field is going into (heat) energy of dissipation of charged conducting fluids. Equation (20.208) is numerically preferable to equation (20.188), which can suffer loss of precision from cancellation of large quantities; equation (20.188) can be used as a check.

20.19 General cylindrically symmetric spacetimes

The equations governing cylindrically symmetric spacetimes share much in common with those for spherically symmetric spacetimes.

The line-element of a general cylindrically symmetric spacetime can be taken to be, in cylindrical coordinates $x^\mu \equiv \{t, z, x, \phi\}$,

$$ds^2 = -\alpha^2 dt^2 + \frac{1}{b_1^2} (dz - \alpha b_0 dt)^2 + r^2 (dx^2 + x^2 d\phi^2) , \quad (20.209)$$

in which the metric coefficients α , b_0 , b_1 , and r are functions of time t and vertical position z . The cylindrical line-element (20.209) is similar but not identical to the spherical line-element (20.1). The radius $r(t, z)$ in the line-element (20.209) can be thought of as a cylindrical cosmic scale factor, and the coordinate x as a comoving cylindrical coordinate. Unlike the spherical case, the coefficients $b_0(t, z)$ and $b_1(t, z)$ are *not* equal

to $\beta_m \equiv \partial_m r$. Quantities β_m are defined to be directed derivatives of the radius r , the same as in the spherical line-element, equation (20.9),

$$\beta_m \equiv \partial_m r = \left\{ \frac{1}{\alpha} \frac{\partial r}{\partial t} + b_0 \frac{\partial r}{\partial z}, b_1 \frac{\partial r}{\partial z}, 0, 0 \right\} . \quad (20.210)$$

As in the spherical case, β_m is a tetrad 4-vector, and its scalar product with itself is a scalar, which defines the interior mass M ,

$$\frac{2M}{r} \equiv \beta_0^2 - \beta_1^2 . \quad (20.211)$$

The expression (20.211) for the mass M interior to z differs from the spherical case $2M/r = 1 + \beta_0^2 - \beta_1^2$, equation (20.11), because the flat line-element $dx^2 + x^2 d\phi^2$ in (20.209) replaces the spherical line-element $do^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$ in (20.1).

The tetrad connections are

$$\Gamma_{100} = h_0 \equiv \partial_1 \ln \alpha = b_1 \frac{\partial \ln \alpha}{\partial z} , \quad (20.212a)$$

$$\Gamma_{101} = h_1 \equiv b_0 \frac{\partial \ln \alpha b_0}{\partial z} - \partial_0 \ln b_1 , \quad (20.212b)$$

$$\Gamma_{202} = \Gamma_{303} = \frac{\beta_0}{r} , \quad (20.212c)$$

$$\Gamma_{212} = \Gamma_{313} = \frac{\beta_1}{r} , \quad (20.212d)$$

$$\Gamma_{323} = \frac{1}{rx} , \quad (20.212e)$$

which differ from the spherical connections (20.23) in h_0 , h_1 , and Γ_{323} .

With the changes to the interior mass M from equation (20.211), and to the connections h_0 , h_1 , and Γ_{323} from equations (20.212), all the equations in §20.6 for the Riemann, Ricci, Einstein, and Weyl tensors in the spherical case hold unchanged. In particular, the evolution equations (20.62), which follow from the Einstein equations (20.35c) and (20.35b), remain unchanged,

$$D_0 \beta_0 = - \frac{M}{r^2} - 4\pi r p , \quad (20.213a)$$

$$D_0 \beta_1 = 4\pi r f , \quad (20.213b)$$

which are valid for any choice of tetrad frame, not just the centre-of-mass frame.

20.20 An infinite thin wall?

The final problem considered in this Chapter is that of an infinite thin plane in vacuo, not because the problem is soluble, but rather because such a thing cannot exist in general relativity.

In Newtonian gravity, an infinite uniform wall produces a uniform gravitational force towards the wall.

If the wall has mass per unit area of $\hat{\rho}$, then solving Laplace's equation $\nabla^2\phi = 4\pi\rho$ with a delta-function source $\rho = \hat{\rho}\delta(z)$ implies that the gravitational force is the constant $g \equiv -\partial\phi/\partial z = -4\pi\hat{\rho}$ at any distance z from the wall. This is not what happens in general relativity (Jones, 2008).

Consider an infinite uniform thin wall in otherwise empty space. The symmetries of the situation imply that the line-element must take the cylindrically symmetric form (20.209), with z the vertical coordinate. As remarked at the end of §20.19, all the equations in §20.6 in the spherical case hold also for the planar line-element (20.209), provided that β_m , M , and h_m are interpreted as being given by equations (20.210), (20.211), and (20.212). For the cylindrical line-element (20.209), the mass equations (20.44) in the centre-of-mass frame become

$$\frac{\partial_0 M}{\partial_0 r} = -4\pi r^2 p, \quad (20.214a)$$

$$\frac{\partial M/\partial z}{\partial r/\partial z} = 4\pi r^2 \rho. \quad (20.214b)$$

The density and pressure vanish in the vacuum region outside the wall, $\rho = p = 0$. The mass equations (20.214) then imply that all derivatives of M vanish, so the interior mass M is constant everywhere outside the wall.

The vacuum region outside the wall defines no preferred frame, so there is freedom to Lorentz boost the spacelike 4-vector β_m in the $\gamma_0\gamma_1$ plane (the tz plane), such that $\beta_1 = 0$. In accordance with the definition (20.210) of β_1 , the vanishing of β_1 requires $\partial r/\partial z = 0$, that is, r is a function only of t , independent of z . Solving Einstein's equations in vacuo leads to the result that not only r but all the metric coefficients in the line-element (20.209) are functions only of t , independent of z . The resulting vacuum line-element is

$$ds^2 = -\frac{t}{2M} dt^2 + \frac{dz^2}{t} + t^2(dx^2 + x^2 d\phi^2). \quad (20.215)$$

The spacetime described by the line-element (20.215) has vanishing energy-momentum tensor, but a Weyl scalar C of

$$C = -\frac{M}{t^3}. \quad (20.216)$$

The line-element (20.215) is the Kasner (1921) spacetime (Exercise 17.5) with $q_a = \{-\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\}$. Which in turn looks like the Schwarzschild geometry near its singular surface (Exercise 17.6).

It is now apparent why there are difficulties in general relativity in finding a thin wall solution analogous to that in Newtonian gravity. The putative thin wall solution is actually the superluminally infalling region near the singular surface of the Schwarzschild geometry. Singularity theorems, Chapter 18, imply that, as long as the energy-momentum satisfies a positive energy condition, there are geodesics whose future terminates in such a geometry.

The interiors of accreting, spherical black holes

As discussed in Chapter 8, the Reissner-Nordström geometry for an ideal charged spherical black hole contains mathematical wormhole and white hole extensions to other universes. In reality, these extensions are not expected to occur, thanks to the mass inflation instability discovered by Poisson and Israel (1990). This Chapter explores how accretion modifies the internal structure of a spherical black hole. A charged black hole is not astronomically realistic, but it has an inner horizon like a rotating black hole, and may be considered a surrogate for a rotating black hole.

Two important lessons emerge from the investigations in this Chapter. The first is that the inner horizon of an accreting black hole is subject to the inflationary instability discovered by Poisson and Israel (1990). The instability is called inflation because it grows exponentially. The inflationary instability destroys the inner horizon, preventing the wormhole and white hole extensions to other universes that occur in the Reissner-Nordström geometry for an ideal charged spherical black hole. Poisson & Israel dubbed the instability “mass inflation,” but I tend to prefer the term “inflationary instability” since although the interior mass indeed increases exponentially during inflation, it is relativistic counter-streaming, not mass, that drives inflation (Hamilton and Avelino, 2010).

The second important lesson of this Chapter is that dissipation inside a black hole can create a lot of entropy inside a black hole, seemingly causing a problem with the second law of thermodynamics. Classically, entropy is additive along spacelike surfaces. For example, classically you would calculate the entropy of the air in the room you are sitting in by chopping up the volume of the room into small pieces and adding up the entropies of each piece. But inside a (conformally) stationary black hole, surfaces of constant (conformal) stationary time are spacelike, and the volume of a spacelike 3-surface over the age T of a black hole since it first collapsed is of order TR_+^2 , which for black holes that collapsed long ago is vastly larger than a naive estimate R_+^3 of the volume of a sphere of horizon radius R_+ . As shown in §21.10, if entropy is accumulated over this vast volume TR_+^2 , the cumulative entropy can vastly exceed the Bekenstein-Hawking (Bekenstein, 1973; Hawking, 1974) entropy, which is $1/4$ the area of the horizon in Planck units $c = G = \hbar = 1$. Which is problematic if the black hole cannot accommodate more than the Bekenstein-Hawking entropy.

Part of the solution to the problem is that quantum field theory does *not* demand that entropy be additive over a spacelike surface. In a sense that can be made more precise, entropy counts the logarithm of the number of quantum states accessed by the system. To the extent that states of different regions of a spacelike surface

are independent of each other, the number of states is multiplicative over those regions, and the entropy is additive. But quantum field theory allows states to be entangled between spacelike separated regions, in which case the regions are not independent, and their entropy is not additive. The numerical experiments reported in this Chapter are purely classical, and do not address the question of what might cause entanglement. But they do establish that (wholesale) entanglement between spacelike-separated regions inside a black hole seems necessary to avoid overcounting entropy.

The idea that the entropy of a black hole cannot exceed its Bekenstein-Hawking entropy has motivated **holographic** conjectures that the degrees of freedom of a volume are somehow encoded on its boundary, and consequently that the entropy of a volume is bounded by those degrees of freedom. Various counter-examples dispose of most simple-minded versions of holographic entropy bounds. The most successful entropy bound, with no known counter-examples, is Bousso’s (2002) **covariant entropy bound**. The covariant entropy bound concerns not just any old 3-dimensional volume, but rather the 3-dimensional volume formed by a null hypersurface, a lightsheet. For example, the horizon of a black hole is a null hypersurface, a lightsheet. The covariant entropy bound asserts that the entropy that passes (inward or outward) through a lightsheet that is everywhere converging cannot exceed $1/4$ of the 2-dimensional area of the boundary of the lightsheet. The numerical experiments reported in this Chapter are all consistent with Bousso’s covariant entropy bound, §21.11.

All the models presented in this Chapter are spherical and self-similar, as detailed in the previous Chapter 20. The strategy is to start from the simplest models, and to proceed to more complicated models. Section 21.1 defines the parameters of the models. Section 21.2 considers an uncharged black hole accreting a neutral relativistic “baryonic” plasma, while §21.3 generalizes to a charged black hole accreting a charged plasma. Section 21.4 introduces a second “dark matter” component, which counterstreams through the charged plasma, initiating and driving the mass inflation instability. Subsequent sections up to §21.9 explore various aspects of the inflationary instability, notably its cause, relativistic counterstreaming, in §21.6. Then, §21.10 gives the charged plasma a finite conductivity, which allows dissipation to occur, and entropy to be created inside the black hole. The greatest entropy creation occurs when the charge-to-mass of the black hole is large, the accretion rate is tiny, and the conductivity is tuned to a critical value at the borderline between models that do and do not collapse to a black hole, Figures 21.8 or 21.9. Critical models exhibit a periodic self-similar behaviour first discovered by Choptuik (1993). Section 21.11 shows that all the models satisfy (just barely, in the critical cases) Bousso’s covariant entropy bound.

The various models are from Hamilton and Pollack (2005), Hamilton and Pollack (2005), Wallace, Hamilton, and Polhemus (2008), Hamilton and Avelino (2010), which contain further detail.

21.1 Boundary conditions, mass, charge, and equation of state

The previous Chapter 20 set forward the equations governing spherical spacetimes. This section sets out the boundary conditions, mass and charge of the black hole, and equation of state adopted for the accreting spherical black hole models in the remainder of the Chapter.

21.1.1 Boundary conditions at an outer sonic point

Because information can propagate only inward inside the horizon of a black hole, it is natural to set boundary conditions outside the horizon of an accreting black hole. The policy adopted here is to set boundary conditions at a sonic point, where the infalling baryonic (subscripted b) fluid accelerates from subsonic to supersonic. The proper 3-velocity of the baryons through the self-similar frame is ξ_b^1/ξ_b^0 , equation (20.154) (the velocity ξ_b^1/ξ_b^0 is positive falling inward), and the sound speed is

$$\text{sound speed} = \sqrt{\frac{p_b}{\rho_b}} = \sqrt{w_b} , \quad (21.1)$$

and sonic points occur where the velocity equals the sound speed

$$\frac{\xi_b^1}{\xi_b^0} = \pm \sqrt{w_b} \quad \text{at sonic points} . \quad (21.2)$$

The denominator of the expression (20.206) for the proper acceleration $h_{b,0}$ of the baryonic fluid is zero at sonic points, indicating that the acceleration will diverge unless the numerator is also zero. Generically, what happens at a sonic point depends on whether the fluid transitions from subsonic upstream to supersonic downstream (as here) or vice versa. If (as here) the fluid transitions from subsonic to supersonic, then sound waves generated by discontinuities near the sonic point can propagate upstream, plausibly modifying the flow so as to ensure a smooth transition through the sonic point, effectively forcing the numerator, like the denominator, of the expression (20.206) to pass through zero at the sonic point. Conversely, if the fluid transitions from supersonic to subsonic, then sound waves cannot propagate upstream to warn the incoming fluid that a divergent acceleration is coming, and the result is a shock wave, where the fluid accelerates discontinuously, is heated, and thereby passes from supersonic to subsonic.

The solutions considered here assume that the acceleration $h_{b,0}$ at the sonic point is not only continuous (so the numerator of (20.206) is zero) but also differentiable. Such a sonic point is said to be regular, and the assumption imposes two boundary conditions at the sonic point.

The accretion in real black holes is likely to be much more complicated, but the assumption of a regular sonic point is the simplest physically reasonable one.

21.1.2 Mass and charge of the black hole

The mass $M_\bullet$ and charge $Q_\bullet$ of the black hole at any instant are defined here to be those that would be measured by a distant observer if there were no mass or charge outside the sonic point,

$$M_\bullet = M + \frac{Q^2}{2r} , \quad Q_\bullet = Q \quad \text{at the sonic point} . \quad (21.3)$$

The mass $M_\bullet$ in equation (21.3) includes the mass-energy $Q^2/2r$ that would be in the electric field outside the sonic point if there were no charge outside the sonic point, but it does not include mass-energy from any additional mass or charge that might be outside the sonic point.

In self-similar evolution, the black hole mass $M_\bullet$ increases linearly with proper time at rest far from the

black hole. The proper time is recorded on dark matter clocks that free-fall radially from rest far away. In the approximation that there is vanishing energy-momentum outside the sonic point other than that in the electric field, the solution outside the sonic point is Gullstrand-Painlevé. The Gullstrand-Painlevé line-element for dark matter that free falls radially from rest at infinity is equation (20.148) with

$$\beta_{1,d} = 1 \quad (21.4)$$

and unit lapse, the latter implying, from equation (20.149) with time T replaced by the dark matter time τ_d ,

$$1 = \alpha_d = \frac{\xi_d^0 R_d}{v \tau_d} . \quad (21.5)$$

The sonic point is at fixed conformal radius, and equation (21.5) shows that the dark matter time $\tau_d = R_d \xi_d^0 / v$ at that point increases in proportion to the conformal factor R_d . The mass accretion rate $\dot{M}_\bullet$ is

$$\dot{M}_\bullet \equiv \frac{dM_\bullet}{d\tau_d} = \frac{M_\bullet}{\tau_d} = \frac{v M_\bullet}{R_d \xi_d^0} \quad \text{at the sonic point} . \quad (21.6)$$

As remarked following equation (20.144), the residual gauge freedom in the global rescaling of conformal time allows the expansion rate v to be adjusted at will. One choice suggested by equation (21.6) is to set

$$\dot{M}_\bullet = v , \quad (21.7)$$

which is equivalent to scaling v such that

$$\xi_d^0 = \frac{M_\bullet}{R_d} \quad \text{at the sonic point} . \quad (21.8)$$

Equation (21.8) and the boundary condition (21.4) coupled with the scalar relations (20.144a) and (20.144b) fully determine the dark matter 4-vectors $\beta_{d,m}$ and ξ_d^m at the sonic point.

21.1.3 Equation of state

The density ρ_b and temperature T_b of an ideal relativistic baryonic fluid in thermodynamic equilibrium are related by

$$\rho_b = \frac{\pi^2 g_b}{30} T_b^4 , \quad (21.9)$$

where

$$g_b = g_B + \frac{7}{8} g_F \quad (21.10)$$

is the effective number of relativistic particle species, with g_B and g_F being the number of bosonic and fermionic species. If the expected increase in g with temperature T is modelled (so as not to spoil self-similarity) as a weak power law $g_b/g_P = T_b^\epsilon$, with g_P the effective number of relativistic species at the Planck

temperature, then the relation between density ρ_b and temperature T_b is

$$\rho_b = \frac{\pi^2 g_P}{30} T_b^{(1+w)/w} , \quad (21.11)$$

with equation of state parameter $w_b = 1/(3 + \epsilon)$ slightly less than the standard relativistic value $w = 1/3$. In the models considered here, the baryonic equation of state is taken to be

$$w_b = 0.32 . \quad (21.12)$$

The effective number g_P is fixed by setting the number of relativistic particles species to $g_b = 5.5$ at $T_b = 10 \text{ MeV}$, corresponding to a plasma of relativistic photons, electrons, and positrons. This corresponds to choosing the effective number of relativistic species at the Planck temperature to be $g_P \approx 2,400$, which is perhaps not unreasonable. The precise choices of g_b and w_b are not crucial.

The chemical potential of the relativistic baryonic fluid is likely to be close to zero, corresponding to equal numbers of particles and anti-particles. The entropy S_b of a proper Lagrangian volume element V of the fluid is then

$$S_b = \frac{(\rho_b + p_b)V}{T_b} , \quad (21.13)$$

which agrees with the earlier expression (20.196), but now has the correct normalization.

21.2 Black hole accreting a neutral relativistic plasma

Perhaps the simplest model of an accreting black hole that one could think of is that of a spherical black hole accreting a neutral relativistic “baryonic” plasma. In self-similar solutions, the charge of the black hole is produced self-consistently by the accreted charge of the baryonic fluid, so a neutral fluid produces an uncharged black hole.

Figure 21.1 shows the baryonic density ρ_b and Weyl curvature C inside the uncharged black hole. The mass and accretion rate have been taken to be

$$M_\bullet = 4 \times 10^6 M_\odot , \quad \dot{M}_\bullet = 10^{-16} , \quad (21.14)$$

which are motivated by the fact that the mass of the supermassive black hole at the centre of the Milky Way is $4 \times 10^6 M_\odot$, and its accretion rate is of order (Planck units are $c = G = \hbar = 1$)

$$\frac{\text{Mass of MW black hole}}{\text{age of Universe}} \approx \frac{4 \times 10^6 M_\odot}{10^{10} \text{ yr}} \approx \frac{4 \times 10^{44} \text{ Planck units}}{6 \times 10^{60} \text{ Planck units}} \approx 10^{-16} . \quad (21.15)$$

Figure 21.1 shows that the baryonic plasma plunges uneventfully to a central singularity, just as in the Schwarzschild solution. When the Weyl curvature scalar hits the Planck scale, $|C| = 1$, the baryonic proper density ρ_b is still well below the Planck density, so this singularity is curvature-dominated.

Figure 21.1 also shows the rate dS_b/dS_{BH} of increase of the plasma entropy per unit increase in the Bekenstein-Hawking entropy of the black hole, equation (21.45). The relevance of this quantity is discussed

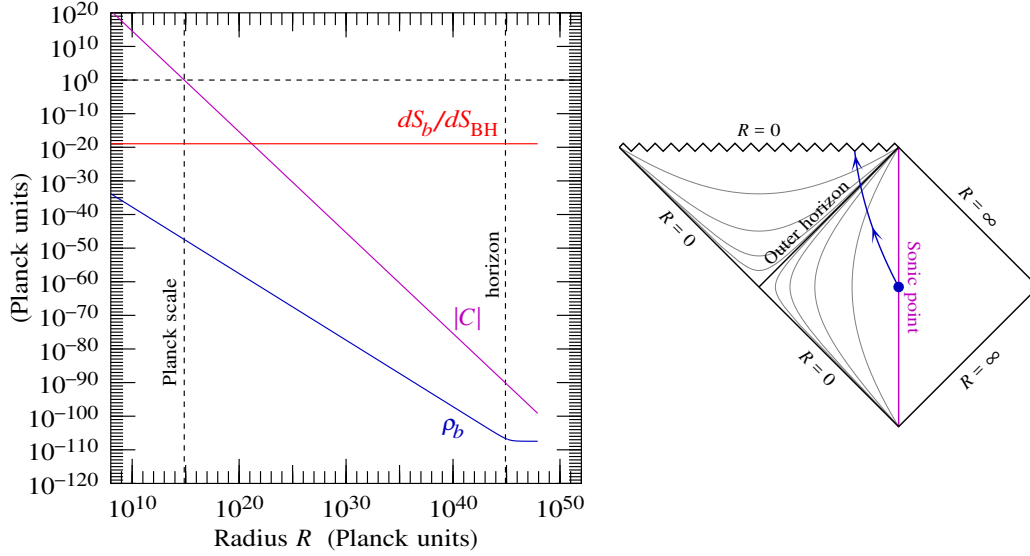


Figure 21.1 An uncharged baryonic plasma falls into an uncharged spherical black hole. The left panel shows in Planck units, as a function of circumferential radius, the plasma density ρ_b , the Weyl curvature scalar C (which is negative), and the rate dS_b/dS_{BH} of increase of the plasma entropy per unit increase in the Bekenstein-Hawking entropy of the black hole, equation (21.45). The mass is $M_\bullet = 4 \times 10^6 M_\odot$, the accretion rate is $\dot{M}_\bullet = 10^{-16}$, and the equation of state is $w_b = 0.32$. The right panel shows a Penrose diagram of the model.

in §21.10. The constancy of dS_b/dS_{BH} in Figure 21.1 reflects the fact that there is no dissipation in this model, so no additional entropy is created inside the black hole.

21.3 Black hole accreting a charged relativistic plasma

The next simplest model one can think of is that of a black hole accreting a charged relativistic plasma. Because the plasma is charged, the resulting black hole is also charged.

Figure 21.2 shows a black hole with charge-to-mass $Q_\bullet/M_\bullet = 10^{-5}$, but otherwise the same parameters as in the uncharged black hole of §21.2: $M_\bullet = 4 \times 10^6 M_\odot$, $\dot{M}_\bullet = 10^{-16}$, and $w_b = 0.32$. Inside the outer horizon, the baryonic plasma, repelled by the electric charge of the black hole self-consistently generated by the accretion of the charged baryons, becomes outgoing. Like the Reissner-Nordström geometry, the black hole has an (outgoing) inner horizon. The baryons drop through the inner horizon, shortly after which the self-similar solution terminates at an irregular sonic point, where the proper acceleration diverges. Normally this is a signal that a shock must form, but even if a shock is introduced, the plasma still terminates at an irregular sonic point shortly downstream of the shock. The failure of the self-similar solution to continue does not invalidate the solution to the past of the inner horizon, because the failure is hidden beneath the inner horizon, and cannot be communicated to infalling matter above it.

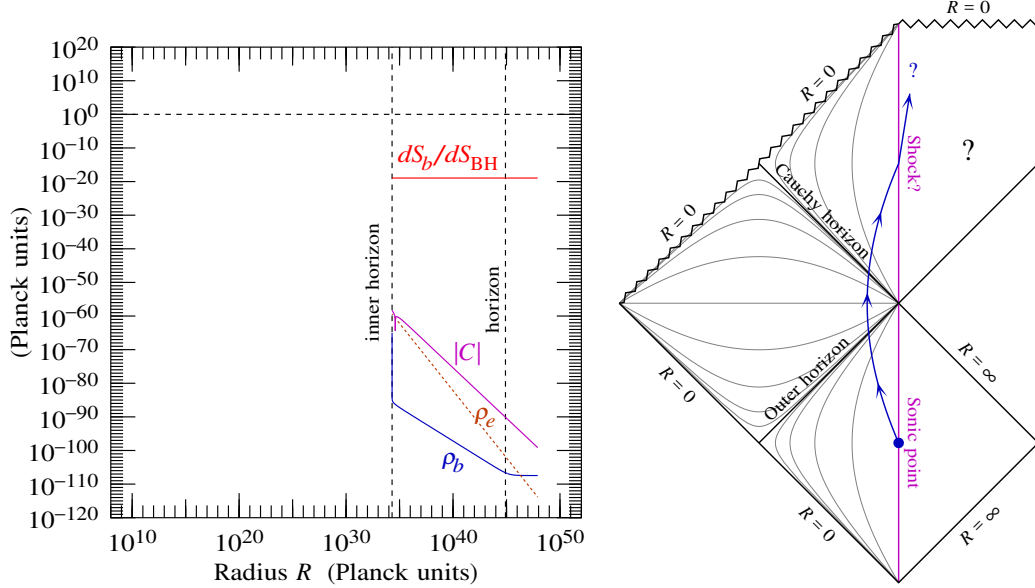


Figure 21.2 A charged, nonconducting, baryonic plasma falls into a charged black hole. The black hole has an inner horizon like the Reissner-Nordström geometry. The self-similar solution terminates at an irregular sonic point just beneath the inner horizon. The mass is $M_\bullet = 4 \times 10^6 M_\odot$, accretion rate $\dot{M}_\bullet = 10^{-16}$, equation of state $w_b = 0.32$, and black hole charge-to-mass $Q_\bullet/M_\bullet = 10^{-5}$. The right panel shows a Penrose diagram. The inner horizon is a Cauchy horizon: what happens in the spacetime to the future of the Cauchy horizon is unpredictable.

The inner horizon is a Cauchy horizon, meaning that the spacetime to the future of the inner horizon cannot be predicted uniquely from the past. The ambiguity in the possible presence and location of a shock the other side of the Cauchy horizon is a symptom of this unpredictability. Hamilton and Pollack (2005) give further details.

This solution, in which baryonic matter falls through an outgoing inner horizon, is nevertheless not realistic, because it assumes that there is no ingoing matter whatsoever, whereas even the tiniest amount of ingoing energy-momentum, in gravitational waves if nothing else, would suffice to trigger the inflationary instability. Such ingoing energy-momentum would appear infinitely blueshifted to the outgoing baryons falling through the inner horizon, which would produce inflation, as in §21.4.

21.4 Black hole accreting charged baryons and dark matter

One way to allow mass inflation in simple models is to admit not one but two fluids that can counter-stream relativistically through each other. A natural possibility is to feed the black hole not only with a charged relativistic fluid of baryons but also with neutral pressureless dark matter that streams freely through the

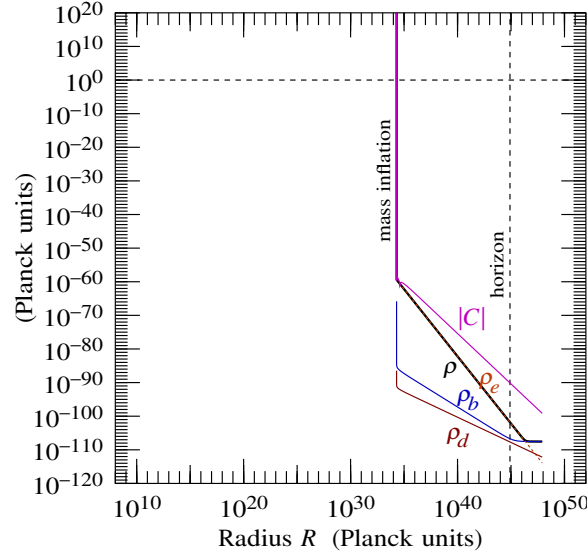


Figure 21.3 Not only charged baryonic plasma but also neutral pressureless dark matter fall into a black hole. The dark matter streams freely through the baryonic plasma. The relativistic counter-streaming produces mass inflation just above the erstwhile inner horizon, where the centre-of-mass density ρ (thick black line) and curvature C inflate rapidly to the Planck scale and beyond. The mass is $M_{\bullet} = 4 \times 10^6 M_{\odot}$, the accretion rate $\dot{M}_{\bullet} = 10^{-16}$, the baryonic equation of state $w_b = 0.32$, the charge-to-mass $Q_{\bullet}/M_{\bullet} = 10^{-5}$, the conductivity is zero, and the ratio of dark matter to baryonic density at the outer sonic point is $\rho_d/\rho_b = 0.1$.

baryons. The charged baryons, being repelled by the electric charge of the black hole, become outgoing, while the neutral dark matter remains ingoing.

Figure 21.3 shows that relativistic counter-streaming between the baryons and the dark matter causes the centre-of-mass density ρ and the Weyl curvature scalar C to inflate quickly up to the Planck scale and beyond. The ratio of dark matter to baryonic density at the sonic point is $\rho_d/\rho_b = 0.1$, but otherwise the parameters are the generic parameters of the previous two sections: $M_{\bullet} = 4 \times 10^6 M_{\odot}$, $\dot{M}_{\bullet} = 10^{-16}$, $w_b = 0.32$, $Q_{\bullet}/M_{\bullet} = 10^{-5}$, and zero conductivity. Almost all the centre-of-mass energy ρ is in the counter-streaming energy between the outgoing baryonic and ingoing dark matter. The individual densities ρ_b of baryons and ρ_d of dark matter (and ρ_e of electromagnetic energy) increase only modestly.

A striking feature of mass inflation is that the smaller the accretion rate, the shorter the length scale of inflation. Not only that, but the smaller one of the outgoing or ingoing streams is relative to the other, the shorter the length scale of inflation. Figure 21.4 shows black holes with three different accretion rates $\dot{M}_{\bullet} = 0.03, 0.01$, and 0.003 , all with the same ratio $\rho_d/\rho_b = 0.1$ of the dark-matter-to-baryon density ratio at the sonic point. The smaller the accretion rate, the faster is inflation. The accretion rates $\dot{M}_{\bullet}$ have been chosen to be relatively large so that the inflationary growth rate is discernible easily on the graph. The centre-of-mass density ρ and Weyl scalar C exponentiate along with, and in proportion to, the interior mass

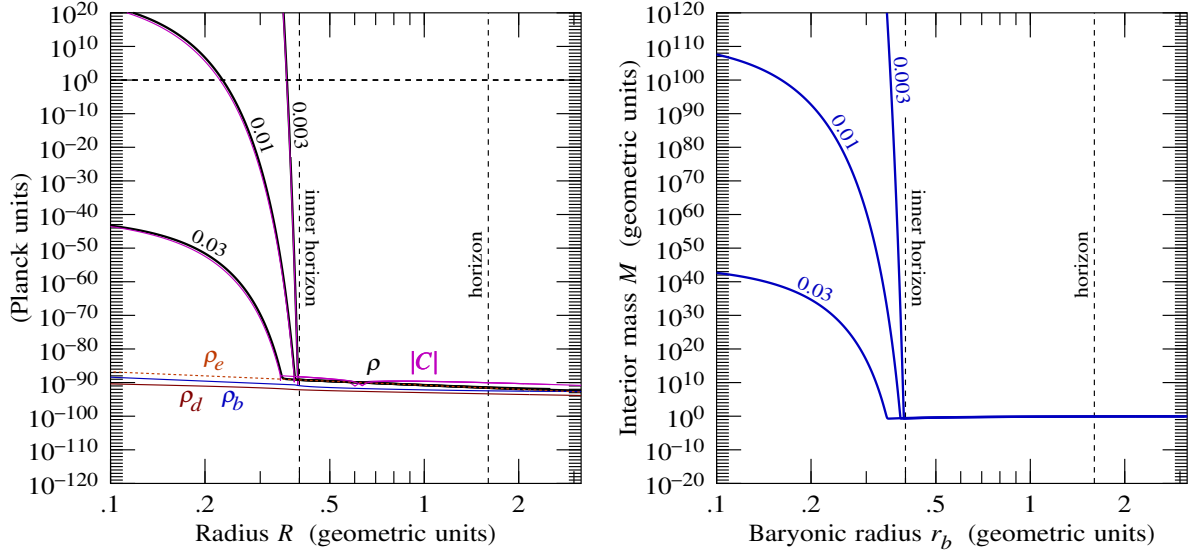


Figure 21.4 (Left panel) The centre-of-mass density ρ and Weyl curvature $|C|$, and (right panel) interior mass M , inside three black holes accreting baryons and dark matter at three different rates $\dot{M}_\bullet = 0.03, 0.01$, and 0.003 . In all three cases the dark-matter-to-baryon ratio at the sonic point is $\rho_d/\rho_b = 0.1$. The smaller the accretion rate, the faster the centre-of-mass density ρ , curvature C , and interior mass M inflate; note that the centre-of-mass energy ρ (thick black line) and the curvature $|C|$ almost coincide here. For the middle accretion rate $\dot{M}_\bullet = 0.01$ (to avoid confusion, only this case is plotted), the graph also shows the individual proper densities ρ_b of baryons, ρ_d of dark matter, and ρ_e of electromagnetic energy. During mass inflation, almost all the centre-of-mass energy ρ is in the streaming energy: the proper densities of individual components remain small. The black hole mass is $M_\bullet = 4 \times 10^6 M_\odot$, the baryonic equation of state is $w_b = 0.32$, the charge-to-mass is $Q_\bullet/M_\bullet = 0.8$, and the conductivity is zero. The position where the inner horizon would be for a Reissner-Nordström black hole of $Q_\bullet/M_\bullet = 0.8$ is marked, but in fact the inner horizon is destroyed by the inflationary instability.

M , which increases as the radius r decreases approximately as (see Hamilton and Avelino (2010) for more precise estimates)

$$\rho \propto C \propto M \propto \exp(-\ln r/\dot{M}_\bullet) . \quad (21.16)$$

Physically, the scale of length of inflation is set by how close to the inner horizon infalling material approaches before mass inflation begins. The smaller the accretion rate $\dot{M}_\bullet$, the closer the approach, and consequently the shorter the length scale of inflation.

Figure 21.4 shows that, as in Figure 21.3, almost all the centre-of-mass energy density ρ is in the streaming energy between the baryons and the dark matter. For one case, $\dot{M}_\bullet = 0.01$, Figure 21.4 shows the individual densities ρ_b of baryons and ρ_d of dark matter in their own frames, and ρ_e of electromagnetic energy, all of which remain tiny compared to the streaming energy.

Figure 21.4 also shows that inflation in due course comes to an end, whereupon the spacetime collapses to a spacelike singularity at zero radius. Hamilton and Avelino (2010) show that the maximum interior mass

attained is approximately the exponential of the reciprocal of the mass accretion rate,

$$M_{\max} \sim \exp(1/\dot{M}_{\bullet}) . \quad (21.17)$$

For small accretion rates, this interior mass is absurdly huge. For example, for the “realistic” accretion rate of $\dot{M}_{\bullet} = 10^{-16}$ adopted in the model of Figure 21.3, the maximum interior mass attained is $M_{\max} \sim e^{10^{16}}$, and the maximum proper streaming density ρ and curvature C are similarly ridiculously vast. The density and curvature vastly exceed the Planck scale.

Curvature is synonymous with tidal force. It seems entirely likely that the tidal force will result in pair creation that grows intense as the curvature approaches the Planck scale. Frolov, Kristjansson, and Thorlacius (2006) show that in the case a charged black hole in 2 spacetime dimensions, such pair creation does in fact occur. However, there have been no studies of what happens in the realistic case of 4 spacetime dimensions.

21.5 The black hole collider

The previous section, §21.4, showed that almost all the centre-of-mass energy during mass inflation is in the energy of counter-streaming. Thus the black hole acts like an extravagantly powerful particle accelerator (Hamilton and Avelino, 2010).

Each baryon in the black hole collider sees a flux $n_d u^1$ of dark matter particles per unit area per unit time, where $n_d = \rho_d/m_d$ is the proper number density of dark matter particles in their own frame, and u^1 is the radial component of the proper 4-velocity, the γv , of the dark matter through the baryons. The γ factor in u^1 is the relativistic beaming factor: all frequencies, including the collision frequency, are speeded up by the relativistic beaming factor γ . As the baryons accelerate through the collider, they spend a proper time interval $d\tau/d\ln u^1$ in each e -fold of Lorentz factor u^1 . The number of collisions per baryon per e -fold of u^1 is the dark matter flux $(\rho_d/m_d)u^1$, multiplied by the time $d\tau/d\ln u^1$, multiplied by the collision cross-section σ . The total cumulative number of collisions that have happened in the black hole particle collider equals this multiplied by the total number of baryons that have fallen into the black hole, which is approximately equal to the black hole mass $M_{\bullet}$ divided by the mass m_b per baryon. Thus the total cumulative number of collisions in the black hole collider is

$$\frac{\text{number of collisions}}{e\text{-fold of } u^1} = \frac{M_{\bullet}}{m_b} \frac{\rho_d}{m_d} \sigma u^1 \frac{d\tau}{d\ln u^1} . \quad (21.18)$$

Figure 21.5 shows, for several different accretion rates $\dot{M}_{\bullet}$, the collision rate $M_{\bullet} \rho_d u^1 d\tau/d\ln u^1$ of the black hole collider, expressed in units of the black hole accretion rate $\dot{M}_{\bullet}$. This collision rate, multiplied by $\dot{M}_{\bullet} \sigma/(m_d m_b)$, gives the number of collisions (21.18) in the black hole. In the units $c = G = 1$ being used here, the mass of a baryon (proton) is $1 \text{ GeV} \approx 10^{-54} \text{ m}$. If the cross-section σ is expressed in canonical accelerator units of femtobarns ($1 \text{ fb} = 10^{-43} \text{ m}^2$) then the number of collisions (21.18) is

$$\frac{\text{number of collisions}}{e\text{-fold of } u^1} = 10^{45} \left(\frac{\sigma}{1 \text{ fb}} \right) \left(\frac{300 \text{ GeV}^2}{m_b m_d} \right) \left(\frac{\dot{M}_{\bullet}}{10^{-16}} \right) \left(\frac{\rho_d u^1 d\tau/d\ln u^1}{0.03 \dot{M}_{\bullet}/M_{\bullet}} \right) . \quad (21.19)$$

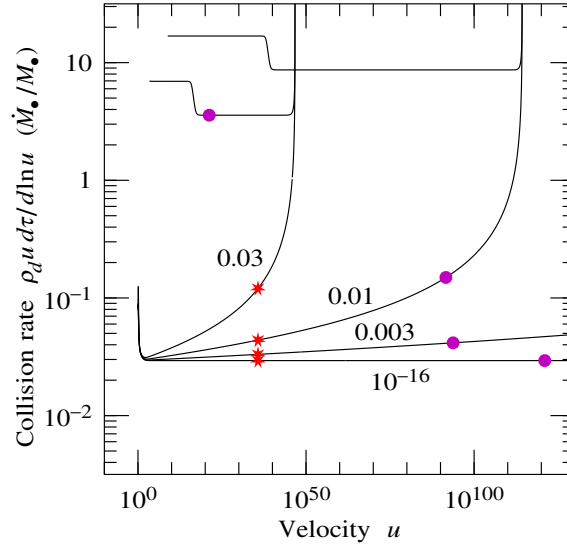


Figure 21.5 Collision rate of the black hole collider per e -fold of velocity u (meaning γv), expressed in units of the inverse black hole accretion time $\dot{M}_\bullet/M_\bullet$. The curves are labelled with their mass accretion rates: $\dot{M}_\bullet = 0.03, 0.01, 0.003$, and 10^{-16} (the three models with the larger accretion rates are the same as those in Figure 21.4). Stars mark where the centre-of-mass energy of colliding baryons and dark matter particles exceeds the Planck energy, while disks show where the Weyl curvature scalar C exceeds the Planck scale.

Particle accelerators measure their cumulative luminosities in inverse femtobarns. Equation (21.19) shows that the black hole accelerator delivers about 10^{45} femtobarns $^{-1}$ ($= 100 \text{ m}^2$), and it does so in each e -fold of collision energy up to the Planck energy and beyond.

To quote the final sentences of Hamilton (2011): “It appears inescapable that Nature is conducting vast numbers of collision experiments over a broad range of peri- and super-Planckian energies in large numbers of black holes throughout our Universe. Does Nature do anything interesting with this extravagance—such as create baby universes—or is it merely a final hurrah en route to nothingness?”

21.6 The mechanism of mass inflation

This section explains why mass inflation occurs, and why it is inevitable as long as even the tiniest streams of outgoing and ingoing energy-momentum impinge on the inner horizon. The arguments are from Hamilton and Avelino (2010), which gives more detail. For a taste of how this works out mathematically, Exercise 21.1 takes you through the case of equal pressureless streams.

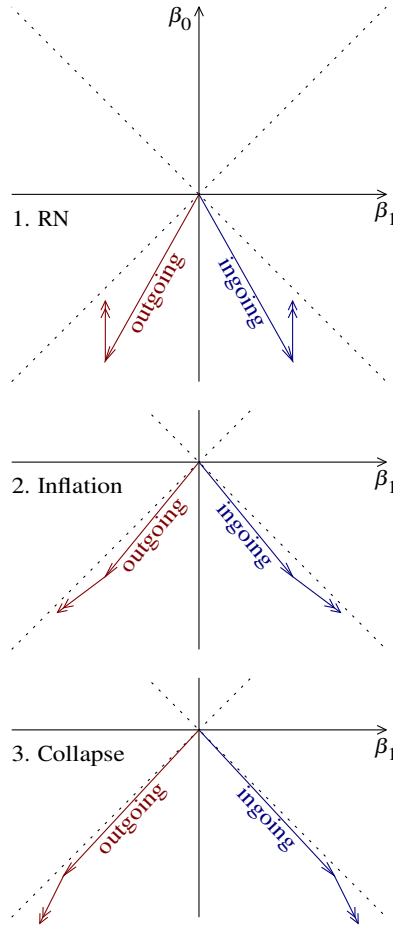


Figure 21.6 Spacetime diagrams of the tetrad-frame 4-vector β_m , equation (20.9), illustrating qualitatively the three successive phases of mass inflation: 1. (top) the Reissner-Nordström phase, where inflation ignites; 2. (middle) the inflationary phase itself; and 3. (bottom) the collapse phase, where inflation comes to an end. In each diagram, the arrowed lines labelled outgoing and ingoing illustrate two representative examples of the 4-vector $\{\beta_0, \beta_1\}$, while the double-arrowed lines illustrate the rate of change of these 4-vectors implied by Einstein's equations (20.62). Inside the horizon of a black hole, all locally inertial frames necessarily fall inward, so the radial velocity $\beta_0 \equiv \partial_0 r$ is always negative. A locally inertial frame is outgoing or ingoing depending on whether the proper radial gradient $\beta_1 \equiv \partial_1 r$ measured in that frame is negative or positive.

21.6.1 Reissner-Nordström phase

Figure 21.6 illustrates how the two Einstein equations (20.62) produce the three phases of mass inflation inside a charged spherical black hole.

During the initial phase, illustrated in the top panel of Figure 21.6, the spacetime geometry is well-approximated by the vacuum, Reissner-Nordström geometry. During this phase the radial energy flux f is effectively zero, so β_1 remains constant, according to equation (20.62b). The change in the radial velocity β_0 , equation (20.62a), depends on the competition between the Newtonian gravitational force $-M/r^2$, which is always attractive (tending to make the radial velocity β_0 more negative), and the gravitational force $-4\pi rp$ sourced by the radial pressure p . In the Reissner-Nordström geometry, the static electric field produces a negative radial pressure, or tension, $p = -Q^2/(8\pi r^4)$, which produces a gravitational repulsion $-4\pi rp = Q^2/(2r^3)$. At some point (depending on the charge-to-mass ratio) inside the outer horizon, the gravitational repulsion produced by the tension of the electric field exceeds the attraction produced by the interior mass M , so that the radial velocity β_0 slows down. This regime, where the (negative) radial velocity β_0 is slowing down (becoming less negative), while β_1 remains constant, is illustrated in the top panel of Figure 21.6.

If the initial Reissner-Nordström phase were to continue, then the radial 4-gradient β_m would become lightlike. In the Reissner-Nordström geometry this does in fact happen, and where it happens defines the inner horizon. The problem with this is that the lightlike 4-vector β_m points in one direction for outgoing frames, and in the opposite direction for ingoing frames. If β_m becomes lightlike, then outgoing and ingoing frames are streaming through each other at the speed of light. This is the infinite blueshift at the inner horizon first pointed out by Penrose (1968).

If there were no matter present, or if there were only one stream of matter, either outgoing or ingoing but not both, then β_m could indeed become lightlike. But if both outgoing and ingoing matter are present, even in the tiniest amount, then it is physically impossible for the outgoing and ingoing frames to stream through each other at the speed of light.

If both outgoing and ingoing streams are present, then as they race through each other ever faster, they generate a radial pressure p , and an energy flux f , which begin to take over as the main source on the right hand side of the Einstein equations (20.62). This is how mass inflation is ignited.

21.6.2 Inflationary phase

The infalling matter now enters the second, mass inflationary phase, illustrated in the middle panel of Figure 21.6.

During this phase, the gravitational force on the right hand side of the Einstein equation (20.62a) is dominated by the pressure p produced by the counter-streaming outgoing and ingoing matter. The mass M is completely sub-dominant during this phase (in this respect, the designation “mass inflation” is misleading, since although the mass inflates, it does not drive inflation). The counter-streaming pressure p is positive, and so accelerates the radial velocity β_0 (makes it more negative). At the same time, the radial gradient β_1 is being driven by the energy flux f , equation (20.62b). For typically low accretion rates, the streams are cold, in the sense that the streaming energy density greatly exceeds the thermal energy density, even if the accreted material is at relativistic temperatures. This follows from the fact that for mass inflation to begin, the gravitational force produced by the counter-streaming pressure p must become comparable to that produced by the mass M , which for streams of low proper density requires a hyper-relativistic streaming velocity. For a cold stream of proper density ρ moving at 4-velocity $u^m \equiv \{u^0, u^1, 0, 0\}$, the streaming energy

flux would be $f \sim \rho u^0 u^1$, while the streaming pressure would be $p \sim \rho (u^1)^2$. Thus their ratio $f/p \sim u^0/u^1$ is slightly greater than one. It follows that, as illustrated in the middle panel of Figure 21.6, the change in β_1 slightly exceeds the change in β_0 , which drives the 4-vector β_m , already nearly lightlike, to be even more nearly lightlike. This is mass inflation.

Inflation feeds on itself. The radial pressure p and energy flux f generated by the counter-streaming outgoing and ingoing streams increase the gravitational force. But, as illustrated in the middle panel of Figure 21.6, the gravitational force acts in opposite directions for outgoing and ingoing streams, tending to accelerate the streams faster through each other. An intuitive way to understand this is that the gravitational force is always inwards, meaning in the direction of smaller radius, but the inward direction is towards the black hole for ingoing streams, and away from the black hole for outgoing streams.

The feedback loop in which the streaming pressure and flux increase the gravitational force, which accelerates the streams faster through each other, which increases the streaming pressure and flux, is what drives mass inflation. Inflation produces an exponential growth in the streaming energy, and along with it the interior mass, and the Weyl curvature.

21.6.3 Collapse phase

It might seem that inflation is locked into an exponential growth from which there is no exit. But the Einstein equations (20.62) have one more trick up their sleeve.

For the counter-streaming velocity to continue to increase requires that the change in β_1 from equation (20.62b) continues to exceed the change in β_0 from equation (20.62a). This remains true as long as the counter-streaming pressure p and energy flux f continue to dominate the source on the right hand side of the equations. But the mass term $-M/r^2$ also makes a contribution to the change in β_0 , equation (20.62a). It turns out (Hamilton and Avelino, 2010) that, at least in the case of collisionless streams, the mass term exponentiates slightly faster than the pressure term (in Exercise 21.1, for example, this occurs because in equation (21.28) there is a $+\beta^2$ term in the numerator and a $-\beta^2$ term in the denominator). At a certain point, the additional acceleration produced by the mass means that the combined gravitational force $M/r^2 + 4\pi r p$ exceeds $4\pi r f$. Once this happens, the 4-vector β_m , instead of being driven to becoming more lightlike, starts to become less lightlike. That is, the counter-streaming velocity starts to slow. At that point inflation ceases, and the streams quickly collapse to zero radius.

It is ironic that it is the increase of mass that brings mass inflation to an end. Not only does mass not drive mass inflation, but as soon as mass begins to contribute significantly to the gravitational force, it brings mass inflation to an end.

21.7 The far future?

The Penrose diagram of a Reissner-Nordström, Figure 8.7, or Kerr-Newman black hole indicates that an observer who passes through the outgoing inner horizon sees the entire future of the outside universe go by. In a sense, this is “why” the outside universe appears infinitely blueshifted.

This raises the question of whether what happens at the outgoing inner horizon of a real black hole indeed depends on what happens in the far future. If it did, then the conclusions of §21.6, which are based in part on the proposition that the accretion rate is approximately constant, would be suspect. A lot can happen in the far future, such as black hole mergers, the Universe ending in a big crunch, Hawking evaporation, or something else beyond our current ken.

Outgoing and ingoing observers both see each other highly blueshifted near the inner horizon. An outgoing observer sees ingoing observers from the future, while an ingoing observer sees outgoing observers from the past. Each stream sees approximately one black hole crossing time elapse on the opposing stream for each e -fold increase in blueshift (Hamilton and Avelino, 2010).

For astronomically realistic black holes, exponentiating the Weyl curvature up to the Planck scale will take typically a few hundred e -folds of blueshift, as illustrated for example in Figure 21.4. Thus what happens at the inner horizon of a realistic black hole before quantum gravity intervenes depends only on the immediate past and future of the black hole, a few hundred black hole crossing times, not on the distant future or past. This conclusion holds even if the accretion rate of one of the outgoing or ingoing streams is tiny compared to the other.

From a stream's own point of view on the other hand, the entire inflationary episode goes by in a flash.

21.8 Weak null singularity on the Cauchy horizon?

It is commonly stated in the literature that the generic outcome of inflation is a “weak null singularity on the Cauchy horizon.” Weak means that the tidal force, the Weyl curvature, exponentiates to infinity in a finite amount of proper time. Null refers to the fact that the streaming velocity between outgoing and ingoing streams reaches the speed of light.

In my view this conclusion is incorrect. The conclusion is an artefact of assuming that after collapsing, a black hole remains isolated for ever, whereas real astronomical black holes accrete, cosmic microwave background photons if nothing else. Moreover the conclusion of a weak null singularity ignores the fact that the diverging tidal force is likely to result in diverging pair creation, and such pairs would surely act as an effective source of accretion, again precipitating collapse.

The fact that a volume element remains little distorted during inflation even though the tidal force, as measured by the Weyl curvature scalar, exponentiates to huge values was first pointed out by Ori (1991). The physical reason for the small tidal distortion despite the huge tidal force is that the proper time over which the force operates is tiny.

Dafermos (2005) has proved a number of mathematical theorems that establish that a null singularity forms on the Cauchy horizon of a charged spherical black hole accreting a massless scalar field. The situation envisaged by the theorems is that of a black hole that collapses and thereafter remains isolated. The collapse generates an outgoing Price tail of radiation. The theorems assume that the outgoing Price radiation falls off sufficiently rapidly along outgoing null geodesics, and Dafermos and Rodnianski (2005) have proved that the required condition on the Price radiation holds for an isolated spherical black hole accreting a massless scalar field. The theorems confirm the several analytic and numerical studies that have found a null singularity on

the Cauchy horizon (Ori, 1991; Bonanno et al., 1994b; Brady and Smith, 1995; Burko, 1997; Burko and Ori, 1998; Hod and Piran, 1998a; Hod and Piran, 1998b; Ori, 1999; Hansen, Khokhlov, and Novikov, 2005).

Burko (2002; 2003) finds numerically that a null singularity forms only if the scalar field set up outside the horizon falls off sufficiently rapidly, the required degree of rapidity depending on the parameters of the problem, such as the charge-to-mass ratio of the black hole. If too much scalar field continues to be accreted, then no null singularity forms, and the field collapses to a central singularity.

All the results are consistent with the estimate (21.16) that the interior mass inflates exponentially with an exponent inversely proportional to the mass accretion rate $\dot{M}_\bullet$. If the accretion rate goes to zero, $\dot{M}_\bullet \rightarrow 0$, then the exponential growth rate becomes infinite, leading to a weak null singularity.

Frolov, Kristjansson, and Thorlacius (2006) have shown that in the simplified case of a 1+1-dimensional charged black hole, if the effects of pair creation of charged particles are taken into account, then the result is collapse to a spacelike singularity rather than a null singularity on the Cauchy horizon. The result is consistent with the argument of the present paper that as long as there is any source that continues to replenish outgoing and ingoing streams near the inner horizon, the ultimate result will be collapse to a spacelike singularity. The results of Frolov, Kristjansson, and Thorlacius (2006) suggest that even without any direct accretion, pair creation provides a sufficient source of outgoing and ingoing streams.

Exercise 21.1 A collisionless two-stream model of inflation. This problem is from Hamilton and Avelino (2010). Some of the equations below repeat equations elsewhere in this book, but they are left as is so that the problem remains self-contained.

Einstein's equations in a spherically symmetric spacetime imply that the covariant rate of change of the radial 4-gradient $\beta_m \equiv \partial_m r = \{\partial_0 r, \partial_1 r, 0, 0\}$ in the frame of any radially moving orthonormal tetrad is (these are equations (20.62))

$$D_0 \beta_0 = -\frac{M}{r^2} - 4\pi r p, \quad (21.20a)$$

$$D_0 \beta_1 = 4\pi r f, \quad (21.20b)$$

where D_0 is the tetrad-frame covariant time derivative, p is the radial pressure, f is the radial energy flux, and M is the interior mass defined by (this is equation (20.11))

$$\frac{2M}{r} - 1 \equiv \beta^2 \equiv -\beta_m \beta^m = \beta_0^2 - \beta_1^2. \quad (21.21)$$

1. **Freely-falling stream.** Consider a stream of matter that is freely falling radially inside the horizon of a spherically symmetric black hole. Let u be the radial component of the tetrad-frame 4-velocity u^m of the stream relative to the “no-going” frame where $\beta_1 = 0$ (the frame of reference that divides outgoing frames $\beta_1 < 0$ from ingoing frames $\beta_1 > 0$):

$$u^m \equiv \{-\beta_0/\beta, -\beta_1/\beta, 0, 0\} = \{\sqrt{1+u^2}, u, 0, 0\}. \quad (21.22)$$

Note that β_0 is negative inside the horizon for both outgoing and ingoing frames. The time component

$u^0 \equiv -\beta_0/\beta = \sqrt{1+u^2}$ of the tetrad-frame 4-velocity is positive (as it should be for a proper 4-velocity), while the radial component $u \equiv u^1 \equiv -\beta_1/\beta$ of the tetrad-frame 4-velocity is positive outgoing, negative ingoing. Show that along the worldline of the stream

$$\frac{d \ln \beta}{d \ln r} = \frac{1}{\beta^2} \left[-\frac{M}{r} - 4\pi r^2 \left(p + \frac{\beta_1}{\beta_0} f \right) \right] , \quad (21.23a)$$

$$\frac{d \ln u}{d \ln r} = \frac{1}{\beta^2} \left[\frac{M}{r} + 4\pi r^2 \left(p + \frac{\beta_0}{\beta_1} f \right) \right] . \quad (21.23b)$$

[Hint: If the stream is freely falling, then the proper time derivative ∂_0 in the tetrad frame of the stream equals the covariant time derivative D_0 . Thus the proper rates of change of $\ln \beta$ and $\ln u$ with respect to $\ln r$ along the worldline of the stream are

$$\frac{d \ln \beta}{d \ln r} = \frac{\partial_0 \ln \beta}{\partial_0 \ln r} , \quad \frac{d \ln u}{d \ln r} = \frac{\partial_0 \ln u}{\partial_0 \ln r} . \quad (21.24)$$

These can be evaluated through

$$\begin{aligned} \partial_0 \ln \beta &= D_0 \ln \beta = \frac{1}{2\beta^2} D_0 \beta^2 = \frac{1}{2\beta^2} D_0 (\beta_0^2 - \beta_1^2) \\ &= \frac{1}{\beta^2} (\beta_0 D_0 \beta_0 - \beta_1 D_0 \beta_1) , \end{aligned} \quad (21.25a)$$

$$\begin{aligned} \partial_0 \ln u &= D_0 \ln u = D_0 \ln \beta_1 - D_0 \ln \beta \\ &= \frac{1}{\beta_1} D_0 \beta_1 - D_0 \ln \beta , \end{aligned} \quad (21.25b)$$

$$\partial_0 \ln r = \frac{1}{r} \partial_0 r = \frac{\beta_0}{r} , \quad (21.25c)$$

with Einstein's equations (21.20) substituted into equations (21.25a) and (21.25b).]

2. **Equal outgoing and ingoing streams.** Consider the symmetrical case of two equal streams of radially outgoing ($\beta_1 < 0$) and ingoing ($\beta_1 > 0$) neutral, pressureless, non-interacting matter ("dust"), each of proper density ρ in their own frames, freely falling into a charged black hole. Show that

$$\frac{d \ln \beta}{d \ln r} = -\frac{1}{2\beta^2} (-\lambda + \beta^2 + \mu u^2) , \quad (21.26a)$$

$$\frac{d \ln u}{d \ln r} = -\frac{1}{2\beta^2} (\lambda - \beta^2 + \mu + \mu u^2) , \quad (21.26b)$$

where

$$\lambda \equiv Q^2/r^2 - 1 , \quad \mu \equiv 16\pi r^2 \rho . \quad (21.27)$$

Hence conclude that

$$\frac{d \ln \beta}{d \ln u} = \frac{-\lambda + \beta^2 + \mu u^2}{\lambda - \beta^2 + \mu + \mu u^2} . \quad (21.28)$$

[Hint: The assumption that the streams are neutral, pressureless, and non-interacting is needed to make

the streams freely falling, so that equations (21.23) are valid. The pressure p in the tetrad frame of each stream is the sum of the electromagnetic pressure p_e and the streaming pressure p_s

$$p = p_e + p_s . \quad (21.29)$$

The electromagnetic pressure p_e is

$$p_e = -\frac{Q^2}{8\pi r^4} , \quad (21.30)$$

with Q the charge of the black hole, which is constant because the infalling streams are neutral. The streaming pressure p_s that each stream sees is

$$p_s = \rho(u_s^1)^2 , \quad (21.31)$$

where the streaming 4-velocity u_s^m between the two streams is the 4-velocity of the observed stream Lorentz-boosted by the 4-velocity of the observing stream (the radial velocities u^1 of the observed and observing streams have opposite signs)

$$u_s^0 = (u^0)^2 + (u^1)^2 = 1 + 2u^2 , \quad (21.32a)$$

$$u_s^1 = -2u^0 u^1 = -2u\sqrt{1+u^2} . \quad (21.32b)$$

The energy flux f in the tetrad frame of each stream is the streaming flux f_s

$$f = f_s = \rho u_s^0 u_s^1 . \quad (21.33)$$

You should find that the combinations of streaming pressure and flux that go into equations (21.23) are

$$p_s + \frac{\beta_1}{\beta_0} f_s = 2\rho u^2 , \quad (21.34a)$$

$$p_s + \frac{\beta_0}{\beta_1} f_s = -2\rho(1+u^2) . \quad (21.34b)$$

]

3. **Reissner-Nordström phase.** If the accretion rate is small, then initially the stream density ρ is small, and consequently μ is small. Argue that in this regime equation (21.28) simplifies to

$$\frac{d \ln \beta}{d \ln u} = \frac{-\lambda + \beta^2}{\lambda - \beta^2} . \quad (21.35)$$

Hence conclude that

$$\beta = \frac{C}{u} , \quad (21.36)$$

where C is some constant set by initial conditions (generically, C will be of order unity).

4. **Transition to mass inflation.** Argue that in the Reissner-Nordström phase, β becomes small, and u

grows large, as the streams fall to smaller radius r . Argue that in due course equation (21.28) becomes well-approximated by

$$\frac{d \ln \beta}{d \ln u} = \frac{-\lambda + \mu u^2}{\lambda + \mu u^2}. \quad (21.37)$$

Treating λ and μ as constants (which is a good approximation), show that the solution to equation (21.37) subject to the initial condition set by equation (21.36) is

$$\beta = \frac{C(\lambda + \mu u^2)}{\lambda u}. \quad (21.38)$$

[Hint: λ is positive. In the Reissner-Nordström solution, β would go to zero at the inner horizon.]

5. **Sketch.** Sketch the solution (21.38), plotting u against β on logarithmic axes. Mark the regime where mass inflation is occurring.
6. **Inflationary growth rate.** Argue that during mass inflation the inflationary growth rate $d \ln \beta / d \ln r$ is

$$\frac{d \ln \beta}{d \ln r} = -\frac{\lambda^2}{2C^2 \mu}. \quad (21.39)$$

Comment on how the inflationary growth rate depends on accretion rate (on ρ).

21.9 Black hole accreting a fluid with an ultrahard equation of state

Poisson & Israel's (1990) original proposal was that mass inflation would be driven by a "Price tail" (Price, 1972) of gravitational radiation generated during the initial collapse of a black hole. But gravitational radiation is spin 2, which cannot be accommodated by a spherically symmetric spacetime. There are no spherical gravitational waves; the lowest order harmonic of gravitational waves is quadrupole ($\ell = 2$).

This has motivated the most common approach in the literature to modeling inflation in spherical spacetimes, which is to allow the black hole to accrete a massless scalar (spin 0) field, which does admit spherical ($\ell = 0$) waves moving at the speed of light (Christodoulou, 1986; Goldwirth and Tsvi, 1987; Gnedin and Gnedin, 1993; Bonanno et al., 1994a; Brady, 1995; Brady and Smith, 1995; Burko, 1997; Burko and Ori, 1998; Burko, 1999; Husain and Olivier, 2001; Burko, 2002; Burko, 2003; Martín-García and Gundlach, 2003; Dafermos, 2005; Hansen, Khokhlov, and Novikov, 2005; Hod and Piran, 1997; Hod and Piran, 1998a; Hod and Piran, 1998b; Sorkin and Piran, 2001; Oren and Piran, 2003; Dafermos, 2005; Dafermos and Rodnianski, 2005).

No massless scalar field has been observed in nature, although a massive scalar field, the Higgs boson, has been observed by the Large Hadron Collider with a mass of ≈ 125 GeV (ATLAS Collaboration, 2012; CMS Collaboration, 2012), and it is likely that cosmological inflation was driven by a scalar field at GUT density.

An alternative way to model inflation with a single fluid is with a perfect fluid with sound speed equal to the speed of light, $\sqrt{w} = 1$. This kind of fluid is called ultrahard. An ultrahard fluid is not the same as

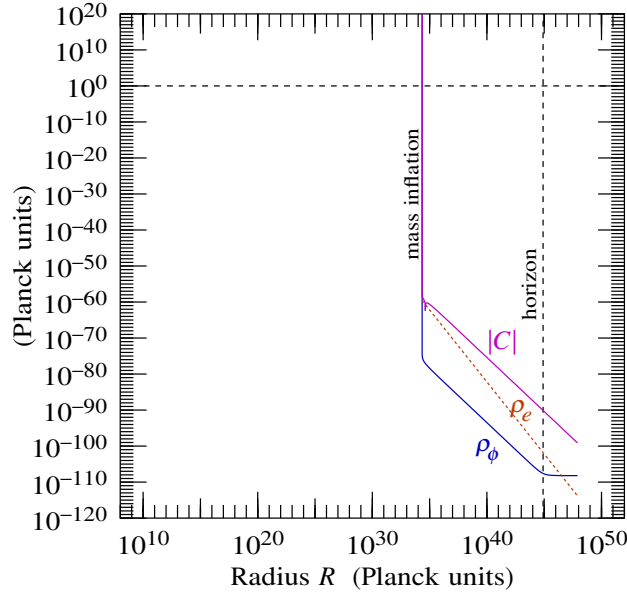


Figure 21.7 Similar to Figure 21.2, but instead of a relativistic fluid, the black hole accretes a charged fluid ϕ with an ultrahard equation of state $w = 1$, which means that the speed of sound equals the speed of light. The fluid therefore supports relativistic counter-streaming, as a result of which mass inflation occurs just above the erstwhile inner horizon. The mass is $M_{\bullet} = 4 \times 10^6 M_{\odot}$, the accretion rate $\dot{M}_{\bullet} = 10^{-16}$, the charge-to-mass $Q_{\bullet}/M_{\bullet} = 10^{-5}$, and the conductivity is zero.

a scalar field, but shares some of its properties (Babichev et al., 2008), notably that it supports spherical waves moving at the speed of light.

Figure 21.7 shows a black hole that accretes a charged, nonconducting fluid with this ultrahard equation of state. The parameters are otherwise the same as in Figure 21.2: a mass of $M_{\bullet} = 4 \times 10^6 M_{\odot}$, an accretion rate of $\dot{M}_{\bullet} = 10^{-16}$, and a black hole charge-to-mass of $Q_{\bullet}/M_{\bullet} = 10^{-5}$. As the Figure shows, mass inflation takes place just above the place where the inner horizon would be. During mass inflation, the density ρ_{ϕ} and the Weyl scalar C exponentiate rapidly up to the Planck scale and beyond. The outcome is quite similar to that of the two-fluid accretion model of Figure 21.3.

21.10 Black hole accreting a conducting charged plasma

This section shows that, if the plasma falling through the horizon is allowed to dissipate inside the black hole, then the ratio dS_b/dS_{BH} , equation (21.45), of the entropy created inside the black hole to the increase

in the Bekenstein-Hawking entropy from the black hole's area increase can (vastly) exceed 1,

$$\frac{dS_b}{dS_{\text{BH}}} \gg 1. \quad (21.40)$$

As discussed in the introduction to this Chapter, this raises the fundamental question about how to count entropy inside a black hole. This section illustrates the problem with a toy model, §21.10.2, in which a spherical black hole accretes a plasma that not only is charged but also has a finite conductivity, so that dissipation can occur, creating entropy inside the horizon. The model is not realistic, but the problem it illustrates is a real one.

21.10.1 Entropy creation

Bekenstein (1973) first argued that a black hole should have a quantum entropy proportional to its horizon area A , and Hawking (1974) supplied the constant of proportionality $1/4$ in Planck units. The Bekenstein-Hawking entropy S_{BH} is, in Planck units $c = G = \hbar = 1$,

$$S_{\text{BH}} = \frac{A}{4}. \quad (21.41)$$

For a spherical black hole of horizon radius R_+ , the area is $A = 4\pi R_+^2$. Hawking showed that a black hole has a temperature T_{H} equal to $1/(2\pi)$ times the surface gravity κ_+ at its horizon, again in Planck units,

$$T_{\text{H}} = \frac{\kappa_+}{2\pi}. \quad (21.42)$$

For a spherical black hole, the surface gravity is $\kappa_+ = -D_0\beta_0 = M/r^2 + 4\pi r\rho$ evaluated at the horizon, equation (20.62a).

The proper velocity of the baryonic fluid through the similarity frame equals ξ_b^1/ξ_b^0 , equation (20.154). Thus the entropy S_b , equation (21.13), accreted through the horizon, at conformal radius r_+ , per unit proper time of the fluid is

$$\frac{dS_b}{d\tau_b} = \frac{4\pi R_b^2 \xi_b^1}{\xi_b^0} \frac{(1 + w_b)\rho_b}{T_b} \Big|_{r_b=r_+}. \quad (21.43)$$

Meanwhile the horizon radius R_+ expands in proportion to the conformal factor, $R_+ \propto e^{v t_b}$, and $dt_b/d\tau_b = \partial_0 t_b = 1/(R_b \xi_b^0)$, so the Bekenstein-Hawking entropy $S_{\text{BH}} = \pi R_+^2$ increases as

$$\frac{dS_{\text{BH}}}{d\tau_b} = \frac{2\pi R_+^2 v}{R_b \xi_b^0}. \quad (21.44)$$

Putting (21.43) and (21.44) together implies that the entropy S_b accreted through the horizon per unit increase of the Bekenstein-Hawking entropy S_{BH} is

$$\frac{dS_b}{dS_{\text{BH}}} = \frac{2R_b^3 \xi_b^1 (1 + w_b)\rho_b}{R_+^2 v T_b} \Big|_{r=r_+}. \quad (21.45)$$

Inside the sonic point, dissipation increases the entropy according to equation (20.207). The entropy varies

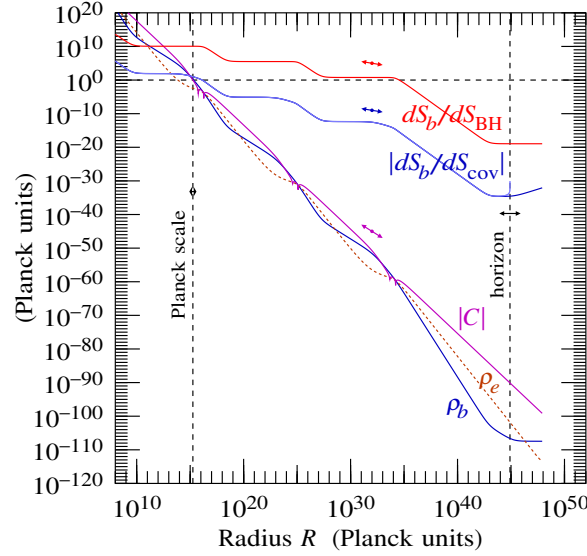


Figure 21.8 Here the baryonic plasma falling into the black hole is charged, and electrically conducting. The conductivity is set equal (within numerical accuracy) to the critical conductivity above which the plasma plunges to a central singularity, since this leads to maximum entropy production inside the horizon. The mass is $M_\bullet = 4 \times 10^6 M_\odot$, the accretion rate $\dot{M}_\bullet = 10^{-16}$, the equation of state $w_b = 0.32$, the charge-to-mass $Q_\bullet/M_\bullet = 10^{-5}$, and the conductivity parameter $\kappa_b = 1.24$. Arrows show how quantities vary a factor of 10 into the past and future.

as $S_b \propto R_b^3 \xi_b^1 (1 + w_b) \rho_b / T_b$, equation (21.13) with volume $V \propto R_b^3 \xi_b^1$, so the rate of increase of the entropy of the black hole, evaluated down to any radius, per unit increase of its Bekenstein-Hawking entropy, is

$$\frac{dS_b}{dS_{\text{BH}}} = \frac{2R_b^3 \xi_b^1 (1 + w_b) \rho_b}{R_+^2 v T_b}, \quad (21.46)$$

which looks the same as equation (21.45) but now evaluated at any radius.

21.10.2 Black hole accreting a conducting relativistic plasma

If the electrical conductivity of the plasma is small, then the solutions resemble the nonconducting solutions of §21.3. But if the conductivity is large enough effectively to neutralize the plasma as it approaches the centre, then the plasma can plunge all the way to the central singularity, as in the uncharged case in §21.2. The most entropy is created inside the black hole when the conductivity is tuned to equal, within numerical accuracy, the critical conductivity above which the plasma collapses to a central singularity.

Figure 21.8 shows the case where the conductivity equals the critical conductivity, here $\kappa_b = 1.24$. The parameters are otherwise the same as in §21.3, a mass of $M_\bullet = 4 \times 10^6 M_\odot$, an accretion rate $\dot{M}_\bullet = 10^{-16}$, an equation of state $w_b = 0.32$, and a black hole charge-to-mass of $Q_\bullet/M_\bullet = 10^{-5}$. The model is from Wallace, Hamilton, and Polhemus (2008).

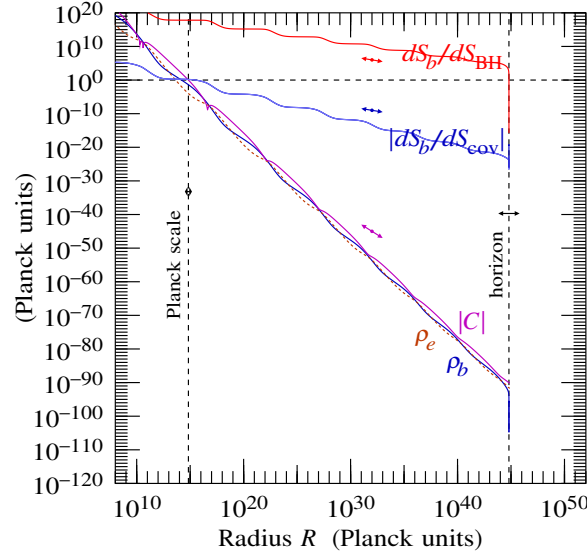


Figure 21.9 This black hole creates a lot of entropy by having a large charge-to-mass $Q_\bullet/M_\bullet = 0.8$ and a low accretion rate $\dot{M}_\bullet = 10^{-28}$, but otherwise the same parameters as in Figure 21.8. The conductivity parameter $\kappa_b = 1.24$ is again at the critical value above which the plasma plunges to a central singularity.

The solution at the critical conductivity exhibits the periodic self-similar behaviour first discovered in numerical simulations by Choptuik (1993), and known as “critical collapse” because it happens at the borderline between solutions that do and do not collapse to a black hole. The ringing of curves in Figure 21.8 is a manifestation of the self-similar periodicity, not a numerical error.

These solutions are not subject to the mass inflation instability, and they could potentially be prototypical of the behaviour inside realistic rotating black holes. For this to work, the outward transport of angular momentum inside a rotating black hole must be large enough effectively to produce zero angular momentum at the centre. Given that angular momentum transport is a rather weak process (Balbus and Hawley, 1998), it seems likely that real rotating black holes do not dissipate all their spin, and that inflation does occur in reality.

Figure 21.8 shows that the entropy produced by Ohmic dissipation inside the black hole can potentially exceed the Bekenstein-Hawking entropy of the black hole by a large factor. The Figure shows the rate dS_b/dS_{BH} of increase of entropy per unit increase in its Bekenstein-Hawking entropy. The rate includes entropy generated down to radius R ; the entropy increases inward because of dissipation. The rate hits unity, $dS_b/dS_{BH} \approx 1$, at a radius of about 10^{-10} of the horizon radius. If the increase of entropy is followed to where the curvature hits the Planck scale, $|C| \approx 1$, then the entropy relative to Bekenstein-Hawking is $dS_b/dS_{BH} \approx 10^{10}$.

Since the model is self-similar, the shape of the curves in Figure 21.8 is fixed with respect to conformal units, but the conversion to proper (in this case Planck) units varies; the arrows show how the curves vary a

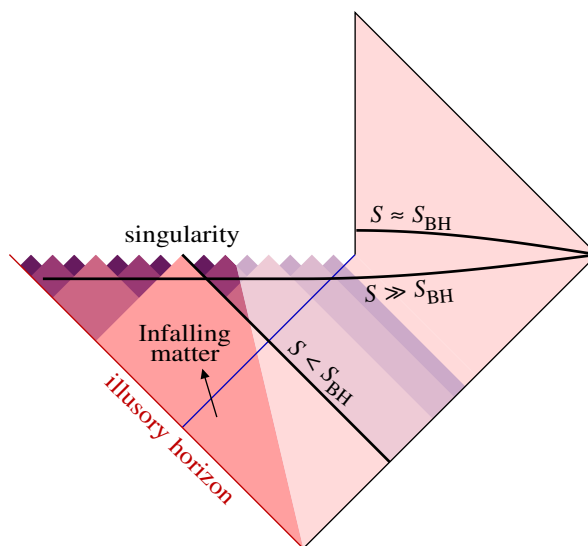


Figure 21.10 Penrose diagram of the accreting, dissipating black hole of Figures 21.8 or 21.9. The entropy passing through the spacelike slice before the black hole evaporates ($S \gg S_{\text{BH}}$) exceeds that passing through the spacelike slice after the black hole evaporates ($S \approx S_{\text{BH}}$), apparently violating the second law of thermodynamics. However, the entropy passing through any null slice respects the second law ($S < S_{\text{BH}}$), consistent with Bousso's (2002) covariant entropy bound. Near the singularity there is a proliferation of spacelike-separated patches of spacetime that cease to be in causal contact because their future lightcones cease to intersect.

factor of 10 into the past and future. If the entropy accumulates additively, then instantaneous rate dS_b/dS_{BH} shown in the Figure can be interpreted as approximately the cumulative entropy created inside the black hole relative to the Bekenstein-Hawking entropy.

If the entropy created inside a black hole exceeds the Bekenstein-Hawking entropy, here by a factor of $\sim 10^{10}$, and the black hole later evaporates radiating only the Bekenstein-Hawking entropy, then entropy is destroyed, violating the second law of thermodynamics.

Generally, the smaller the accretion rate $\dot{M}_\bullet$, the more entropy is produced. If moreover the charge-to-mass $Q_\bullet/M_\bullet$ is large, then the entropy can be produced closer to the outer horizon. Figure 21.9 shows a model with a relatively large charge-to-mass $Q_\bullet/M_\bullet = 0.8$, and a low accretion rate $\dot{M}_\bullet = 10^{-28}$. The large charge-to-mass ratio in spite of the relatively high conductivity requires force-feeding the black hole: the sonic point must be pushed to just above the horizon. The large charge and high conductivity lead to a burst of entropy production just beneath the horizon.

21.11 Holography

Bousso's (2002) covariant entropy bound postulates that the entropy that passes (inward or outward) through a lightsheet that is everywhere converging cannot exceed $1/4$ of the 2-dimensional area of the boundary of the lightsheet.

In the self-similar black holes under consideration, the horizon is expanding, and outgoing lightrays that sit on the horizon do not constitute a converging lightsheet. However, a spherical shell of ingoing lightrays that starts on the horizon falls inwards and therefore does form a converging lightsheet, and a spherical shell of outgoing lightrays that starts just slightly inside the horizon also falls inward and forms a converging lightsheet. The rate at which entropy S_b passes through such outgoing or ingoing spherical lightsheets per unit decrease in the area $S_{\text{cov}} \equiv \pi R_b^2$ of the lightsheet is

$$\left| \frac{dS_b}{dS_{\text{cov}}} \right| = \frac{dS_b}{dS_{\text{BH}}} \frac{R_+^2}{R_b^2} \frac{v}{\xi_b^1 |\beta_{b,0} \pm \beta_{b,1}|} = \frac{2R_b(1+w_b)\rho_b}{|\beta_{b,0} \pm \beta_{b,1}|T_b}, \quad (21.47)$$

in which the $\pm$ sign is $+$ for outgoing, $-$ for ingoing lightsheets. A sufficient condition for Bousso's covariant entropy bound to be satisfied is

$$|dS_b/dS_{\text{cov}}| \leq 1. \quad (21.48)$$

Figures 21.8 and 21.9 show that the rate $|dS_b/dS_{\text{cov}}|$ at which entropy passes through outgoing or ingoing spherical lightsheets is less than one at all scales below the Planck scale. This shows not only that the black holes obey Bousso's covariant entropy bound, but also that no individual observer inside the black hole sees more than the Bekenstein-Hawking entropy on their lightcone, provided that infall is terminated when the curvature reaches the Planck scale.

The Penrose diagram 21.10 illustrates the proliferation of spacetime patches near the singularity that become causally disconnected because their future lightcones cease to intersect. Holography requires that patches are quantum entangled with each other so that the quantum degrees of freedom of volumes inside the black hole are the same Bekenstein-Hawking degrees of freedom regardless of who is observing them.

21.12 Weird stuff at the outer horizon?

A number of papers have suggested that a magical phase transition at, or just outside, the outer horizon prevents any horizon from forming. Is it true?

For example, could there be there a mass inflation instability at the outer horizon? If there were a White Hole on the other side of the outer horizon, then indeed an object entering the outer horizon would encounter an inflationary instability. But in real astronomical black holes formed from the collapse of matter, there is no White Hole, and no inflationary instability at the outer horizon.

Some have argued that quantum field theory may somehow blow up at the horizon. Invariably these arguments confuse the true (event) horizon with the illusory horizon, §7.27. General relativity is unambiguous about what happens at horizons. At least in the macroscopic black holes that exist in our Universe, free-fall

frames at the horizon of a black hole are locally inertial, and quantum field theory should remain well-behaved there.

Others have argued that it takes an infinite time for an infalling observer to reach the horizon, and the black hole evaporates before the observer reaches the horizon, so in effect no horizon ever forms. Again this is incorrect. The reason an outsider sees an infaller take an infinite time to reach the horizon is a light-travel-time effect: light emitted at the horizon remains at the horizon for ever, so it takes an infinite time for light to lift off the horizon, §7.27. In their own frame, an infaller falls through the horizon and reaches the singular surface in a finite proper time. If a second infaller falls in some time after the first infaller, the second infaller does not catch up with the first infaller at the horizon. Rather the second infaller sees the first infaller frozen on the illusory horizon still ahead, still dimming and redshifting away.

Ideal rotating black holes

Among the remarkable mathematical properties of the Kerr-Newman line-element is the fact that, as first shown by Carter (1968), the equations of motion of test particles, massive or massless, neutral or charged, are Hamilton-Jacobi separable. The trajectories of test particles are thus described by a complete set of four integrals of motion. Line-elements with this property are called **separable**. The physically interesting separable spacetimes are Λ -Kerr-Newman black holes, which are ideal charged rotating black holes in a background with a cosmological constant Λ .

The proposition of separability imposes certain conditions on the line-element, §22.3, that would be difficult to guess a priori. In this Chapter, the Kerr solution and its electrovac cousins are derived by separating systematically the Einstein and Maxwell equations. Although conceptually simple, separating the Einstein and Maxwell equations is laborious.

Mathematically, the properties of the Kerr-Newman geometry can be traced to symmetries expressed by the existence of two Killing vectors, associated with stationarity and axisymmetry, and a Killing tensor, associated with separability, §23.3. It is extraordinary that so simple a set of propositions should lead to so intricate a web of implications.

There are other ingenious mathematical ways to arrive at the Kerr solution (Stephani et al., 2003). I like the separable approach not only because of its conceptual simplicity, but also because a generalization of separability to conformal separability yields solutions for rotating black holes that undergo inflation at their inner horizon, Chapter 24, as astronomically realistic black holes must.

22.1 Separable geometries

22.1.1 Separable line-element

The Kerr geometry is stationary, axisymmetric, and separable. Choose coordinates $x^\mu \equiv \{t, x, y, \phi\}$ in which t is the time with respect to which the spacetime is stationary, ϕ is the azimuthal angle with respect to which the spacetime is axisymmetric, and x and y are radial and angular coordinates. In §22.3 it is shown that the

line-element may be taken to be

$$ds^2 = \rho^2 \left[-\frac{\Delta_x}{(1 - \omega_x \omega_y)^2} (dt - \omega_y d\phi)^2 + \frac{dx^2}{\Delta_x} + \frac{dy^2}{\Delta_y} + \frac{\Delta_y}{(1 - \omega_x \omega_y)^2} (d\phi - \omega_x dt)^2 \right], \quad (22.1)$$

where the conditions of stationarity, axisymmetry, and separability imply that the conformal factor ρ is separable

$$\rho = \sqrt{\rho_x^2 + \rho_y^2}, \quad (22.2)$$

and that

$$\begin{aligned} \rho_x, \omega_x, \Delta_x & \text{ are functions of } x \text{ only,} \\ \rho_y, \omega_y, \Delta_y & \text{ are functions of } y \text{ only.} \end{aligned} \quad (22.3)$$

Thanks to the invariant character of the coordinates t and ϕ , the metric coefficients g_{tt} , $g_{t\phi}$, and $g_{\phi\phi}$ all have a gauge-invariant significance,

$$g_{tt} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (-\Delta_x + \omega_x^2 \Delta_y), \quad (22.4a)$$

$$g_{t\phi} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (\omega_y \Delta_x - \omega_x \Delta_y), \quad (22.4b)$$

$$g_{\phi\phi} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (-\omega_y^2 \Delta_x + \Delta_y). \quad (22.4c)$$

The condition $g_{tt} = 0$ defines the boundary of ergospheres, $g_{t\phi} = 0$ defines the turnaround radius, and $g_{\phi\phi} = 0$ defines the boundary of the sisytube. The determinant of the 2×2 submatrix of t - ϕ coefficients is

$$g_{tt}g_{\phi\phi} - g_{t\phi}^2 = -\frac{\rho^4}{(1 - \omega_x \omega_y)^2} \Delta_x \Delta_y. \quad (22.5)$$

The quantity Δ_x is the horizon function. Horizons occur where the horizon function vanishes Δ_x vanishes. The quantity Δ_y is the polar function, whose vanishing defines not a horizon, but rather the location of the (north and south) poles of the geometry. As shown in §23.4, whereas trajectories can pass through a horizon into a region where Δ_x has opposite sign, trajectories cannot pass through $\Delta_y = 0$ into a region where Δ_y has opposite sign. Without loss of generality, the polar function Δ_y can be taken to be positive, since the line-element (22.1) with both Δ_x and Δ_y flipped in sign describes the same geometry with flipped signature.

22.1.2 Λ -Kerr-Newman

As shown in §22.6, the Λ -Kerr-Newman line-element is obtained by imposing boundary conditions that, at least for vanishing cosmological constant, are asymptotically flat far from the black hole, and are nonsingular at the north and south poles, $\theta = 0$ and π . For Λ -Kerr-Newman, the radial and angular parts ρ_x and ρ_y of the separable conformal factor are

$$\rho_x \equiv r = a \cot(ax), \quad \rho_y \equiv a \cos \theta = -ay, \quad (22.6)$$

where r and θ are oblate spheroidal coordinates, §9.2, and a is the spin parameter of the black hole. Why use coordinates x and y in place of r and θ ? Because the coordinate derivatives that arise when separating the Einstein and Maxwell equations, §22.4 and §22.5, are simplest when expressed with respect to x and y . The derivative of x is related to that of r by

$$\frac{\partial}{\partial x} = -R^2 \frac{\partial}{\partial r}, \quad R \equiv \sqrt{r^2 + a^2} = \frac{a}{\sin(ax)}. \quad (22.7)$$

For Λ -Kerr-Newman, the coefficients ω_x and ω_y in the line-element (22.1) are

$$\omega_x = \frac{a}{R^2}, \quad \omega_y = a \sin^2 \theta, \quad (22.8)$$

and the horizon and polar functions Δ_x and Δ_y are (the horizon function Δ_x here is related to the earlier horizon function Δ , equation (9.3), by $\Delta_x = R^{-2}\Delta$)

$$\Delta_x = \frac{1}{R^2} \left(1 - \frac{2M_\bullet r}{R^2} + \frac{Q_\bullet^2 + \mathcal{Q}_\bullet^2}{R^2} - \frac{\Lambda r^2}{3} \right), \quad (22.9a)$$

$$\Delta_y = \sin^2 \theta \left(1 + \frac{\Lambda a^2 \cos^2 \theta}{3} \right), \quad (22.9b)$$

where $M_\bullet$ is the black hole's mass, $Q_\bullet$ and $\mathcal{Q}_\bullet$ are its electric and magnetic charge, and Λ is the cosmological constant. By themselves, Maxwell's equations preclude magnetic charge, in which case $\mathcal{Q}_\bullet = 0$. However, any grand unified theory large enough to predict the quantization of charge (as observed) necessarily contains magnetic charges (magnetic monopoles) as topological defects. In any case, magnetic charge is retained here to bring out the symmetry between electric and magnetic charge. The electromagnetic field is purely radial. The covariant tetrad-frame electromagnetic potential A_m is

$$A_m = \frac{1}{\rho} \left\{ -\frac{Q_\bullet r}{R^2 \sqrt{\Delta_x}}, 0, 0, -\frac{\mathcal{Q}_\bullet \cos \theta}{\sqrt{\Delta_y}} \right\}, \quad (22.10)$$

and the radial electric and magnetic fields E and B are given by

$$E + IB \equiv F_{10} + IF_{23} = \frac{Q_\bullet + I\mathcal{Q}_\bullet}{(\rho_x - I\rho_y)^2}, \quad (22.11)$$

where I is the pseudoscalar of the spacetime algebra, satisfying $I^2 = -1$. The Weyl tensor (12.27) has only a (complex) spin 0 component, and is

$$C = \psi_0 = -\frac{1}{(\rho_x - I\rho_y)^3} \left(M_\bullet - \frac{Q_\bullet^2 + \mathcal{Q}_\bullet^2}{\rho_x + I\rho_y} \right). \quad (22.12)$$

The spacetime is singular at

$$\rho_x = \rho_y = 0, \quad (22.13)$$

which is a ring at $r = 0$ and $\theta = \pi/2$.

22.2 Horizons

Horizons occur where the horizon function Δ_x vanishes,

$$\Delta_x = 0 . \quad (22.14)$$

For Kerr-Newman with vanishing cosmological constant, there are outer and inner horizons $r_{\pm}$ at

$$r_{\pm} = M_{\bullet} \pm \sqrt{M_{\bullet}^2 - a^2 - Q_{\bullet}^2 - \mathcal{Q}_{\bullet}^2} . \quad (22.15)$$

If there is a nonzero cosmological constant Λ , then the horizon condition (22.14) is a quartic in r , and there may be as many as 4 horizons. If there is a small positive cosmological constant, then in addition to the usual outer and inner black hole horizons, there are cosmological horizons at large positive and negative radii. If the cosmological constant is larger and positive, then there are cosmological horizons with no black hole. If the cosmological constant is zero or negative, then there are no cosmological horizons. If the cosmological constant is sufficiently negative, then there is no black hole.

22.3 Conditions from Hamilton-Jacobi separability

This section derives the form (22.1) of the separable line-element from the condition of the separability of the Hamilton-Jacobi equation, coupled with the assumptions of stationarity and axisymmetry. The Hamilton-Jacobi equation is solved in Chapter 23 to obtain the trajectories of neutral or charged particles in rotating charged black holes.

With respect to an orthonormal tetrad, the Hamilton-Jacobi equation for a test particle of mass m and electric charge q moving in a spacetime with vierbein $e_m{}^{\mu}$ and electromagnetic potential A_m is, equation (4.110) or (4.111),

$$\eta^{mn} \left(e_m{}^{\mu} \frac{\partial S}{\partial x^{\mu}} - q A_m \right) \left(e_n{}^{\nu} \frac{\partial S}{\partial x^{\nu}} - q A_n \right) = -m^2 . \quad (22.16)$$

The Hamilton-Jacobi equation (22.16) is a partial differential equation in the particle action S , equation (4.36). Let $\hat{e}_m{}^{\mu}$ and $\hat{A}_m$ denote the inverse vierbein coefficients and tetrad-frame electromagnetic potential with an overall conformal factor ρ factored out:

$$\hat{e}_m{}^{\mu} \equiv \rho e_m{}^{\mu} , \quad \hat{A}_m \equiv \rho A_m . \quad (22.17)$$

With respect to the scaled inverse vierbein $\hat{e}_m{}^{\mu}$ and electromagnetic potential $\hat{A}_m$, the Hamilton-Jacobi equation (22.16) can be rewritten

$$\eta^{mn} \left(\hat{e}_m{}^{\mu} \frac{\partial S}{\partial x^{\mu}} - q \hat{A}_m \right) \left(\hat{e}_n{}^{\nu} \frac{\partial S}{\partial x^{\nu}} - q \hat{A}_n \right) = -m^2 \rho^2 . \quad (22.18)$$

To separate the Hamilton-Jacobi equation (22.18), one demands that the left and right hand sides of the equation be sums of terms each of which depends only on a single coordinate. The “simplest possible way” (Carter, 1968b) to separate the left hand side of the Hamilton-Jacobi equation (22.18) is to impose that each

of the individual factors, comprising the scaled inverse vierbein coefficients $\hat{e}_m{}^\mu$, the derivatives $\partial S/\partial x^\mu$ of the action, and the scaled potentials $\hat{A}_m$, is a function of a single coordinate, and that products of factors are nonvanishing only when all factors are functions of the same coordinate. The derivatives $\partial S/\partial x^\mu$ of the action are each functions of a single coordinate provided that the action S is itself a sum of terms S_μ each depending on a single coordinate x^μ ,

$$S = \sum_{\mu} S_{\mu}(x^{\mu}) . \quad (22.19)$$

Canonical momenta are equal to derivatives of the action, equation (4.105), so the condition (22.19) imposes that each canonical momentum π_μ be a function only of the corresponding coordinate x^μ ,

$$\pi_\mu = \frac{\partial S_\mu}{\partial x^\mu} = \text{function of } x^\mu . \quad (22.20)$$

A special case of the condition (22.20) occurs when a canonical momentum π_μ is a constant, which occurs when the metric is independent of the coordinate x^μ , equation (4.50). In the case of the Kerr geometry and its cousins, the spacetime is stationary and axisymmetric. Stationary means that the geometry is invariant with respect to some time coordinate t , while axisymmetry means that the geometry is invariant with respect to some azimuthal angular coordinate ϕ . The corresponding canonical momenta π_t and π_ϕ are constants of motion, defining respectively the constant energy E and the azimuthal angular momentum L of the trajectory,

$$\pi_t = \frac{\partial S}{\partial t} = -E , \quad \pi_\phi = \frac{\partial S}{\partial \phi} = L . \quad (22.21)$$

If the two remaining coordinates are denoted x and y , then the particle action S , equation (22.19), is the sum

$$S = -Et + L\phi + S_x(x) + S_y(y) , \quad (22.22)$$

where $S_x(x) = \int \pi_x dx$ and $S_y(y) = \int \pi_y dy$, equation (22.20), are respectively functions only of x and y .

Given that π_t and π_ϕ are constants, while π_x and π_y are respectively functions of x and y , the left hand side of the Hamilton-Jacobi equation (22.18) separates as a sum of terms each of which depends only on x or only on y provided that

$$\text{for each } m, \begin{cases} \text{either} & \hat{e}_m{}^\mu \text{ for all } \mu, \text{ and } \hat{A}_m, \text{ are functions of } x \text{ only, and } \hat{e}_m{}^y = 0 , \\ \text{or} & \hat{e}_m{}^\mu \text{ for all } \mu, \text{ and } \hat{A}_m, \text{ are functions of } y \text{ only, and } \hat{e}_m{}^x = 0 . \end{cases} \quad (22.23)$$

The case that matches the Kerr and related geometries is the 2+2 choice

$$\text{the } \begin{cases} \text{top} \\ \text{bottom} \end{cases} \text{ condition of (22.23) holds for } \begin{cases} m = 0 \text{ and } 1 \\ m = 2 \text{ and } 3 \end{cases} . \quad (22.24)$$

Thus separability consistent with Kerr requires that

$$\hat{e}_0{}^y = \hat{e}_1{}^y = \hat{e}_2{}^x = \hat{e}_3{}^x = 0 . \quad (22.25)$$

Given the separability conditions (22.25), the inverse vierbein coefficients $\hat{e}_0{}^x$ and $\hat{e}_3{}^y$ can be transformed

to zero by a tetrad gauge transformation consisting of a Lorentz boost by velocity $\hat{e}_0^x/\hat{e}_1^x$ between tetrad axes γ_0 and γ_1 , and a (commuting) spatial rotation by angle $\text{atan}(\hat{e}_3^y/\hat{e}_2^y)$ between tetrad axes γ_2 and γ_3 . Thus without loss of generality,

$$\hat{e}_0^x = \hat{e}_3^y = 0. \quad (22.26)$$

The gauge conditions (22.26) having been effected, the inverse vierbein coefficients $\hat{e}_1^t$, $\hat{e}_2^t$, $\hat{e}_1^\phi$, and $\hat{e}_2^\phi$ can be eliminated by coordinate gauge transformations $t \rightarrow t'$ and $\phi \rightarrow \phi'$ defined by

$$dt = dt' + \frac{\hat{e}_1^t}{\hat{e}_1^x} dx + \frac{\hat{e}_2^t}{\hat{e}_2^y} dy, \quad d\phi = d\phi' + \frac{\hat{e}_1^\phi}{\hat{e}_1^x} dx + \frac{\hat{e}_2^\phi}{\hat{e}_2^y} dy. \quad (22.27)$$

Equations (22.27) are integrable because $\hat{e}_1^\mu$ and $\hat{e}_2^\mu$ are respectively functions of x and y only. The transformations (22.27) of t and ϕ are admissible because they preserve the Killing vectors $\partial/\partial t$ and $\partial/\partial \phi$,

$$\left. \frac{\partial}{\partial t} \right|_{x,y,\phi} = \left. \frac{\partial}{\partial t'} \right|_{x,y,\phi}, \quad \left. \frac{\partial}{\partial \phi} \right|_{t,x,y} = \left. \frac{\partial}{\partial \phi'} \right|_{t,x,y}. \quad (22.28)$$

Thus without loss of generality

$$\hat{e}_1^t = \hat{e}_2^t = \hat{e}_1^\phi = \hat{e}_2^\phi = 0. \quad (22.29)$$

Finally, coordinate transformations of the x and y coordinates

$$x \rightarrow x', \quad y \rightarrow y', \quad (22.30)$$

can be chosen such that $\hat{e}_1^x$ is any function of x , and $\hat{e}_2^y$ is any function of y . A choice that proves advantageous in separating the Einstein and Maxwell equations is

$$\hat{e}_1^x \hat{e}_0^t = \hat{e}_2^y \hat{e}_3^\phi = \pm 1. \quad (22.31)$$

The separability conditions (22.23) with the 2+2 choice (22.24), which imply conditions (22.25), coupled with the gauge conditions (22.26), (22.29), and (22.31), bring the inverse vierbein e_m^μ to the form

$$e_m^\mu = \frac{1}{\rho} \begin{pmatrix} \frac{1}{\sqrt{\Delta_x}} & 0 & 0 & \frac{\omega_x}{\sqrt{\Delta_x}} \\ 0 & -\sqrt{\Delta_x} & 0 & 0 \\ 0 & 0 & \sqrt{\Delta_y} & 0 \\ \frac{\omega_y}{\sqrt{\Delta_y}} & 0 & 0 & \frac{1}{\sqrt{\Delta_y}} \end{pmatrix}, \quad (22.32)$$

where ω_x and Δ_x are some functions of x , and ω_y and Δ_y are some functions of y . The minus sign in e_1^x is chosen so that, for Λ -Kerr-Newman, the radial tetrad basis vector γ_1 points outward, the direction of

increasing radius r but decreasing x . The corresponding vierbein $e^m{}_\mu$ is

$$e^m{}_\mu = \rho \begin{pmatrix} \frac{\sqrt{\Delta_x}}{1 - \omega_x \omega_y} & 0 & 0 & -\frac{\omega_y \sqrt{\Delta_x}}{1 - \omega_x \omega_y} \\ 0 & -\frac{1}{\sqrt{\Delta_x}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{\Delta_y}} & 0 \\ -\frac{\omega_x \sqrt{\Delta_y}}{1 - \omega_x \omega_y} & 0 & 0 & \frac{\sqrt{\Delta_y}}{1 - \omega_x \omega_y} \end{pmatrix}, \quad (22.33)$$

which implies the line-element (22.1).

The above form (22.32) of the inverse vierbein was derived from the condition that the left hand side of the Hamilton-Jacobi equation (22.18) be separable. Given stationarity and axisymmetry, the left hand side is a sum of two terms, one depending on the radial coordinate x , the other on the angular coordinate y . If the mass m is nonzero, then the squared conformal factor ρ^2 on the right hand side of the Hamilton-Jacobi equation (22.18) must also separate as a sum of terms depending on x and y . This is the condition (22.2).

If Hamilton-Jacobi separability is demanded only for massless particles, $m = 0$, then a more general class of conformally separable solutions can be found, which are explored in Chapter 24.

Exercise 22.1 Explore other separable solutions. The above derivation of the form of the line-element assumed not only separability, but also stationarity and axisymmetry, and the 2+2 choice (22.24) that matches Kerr. Explore other possible choices (Carter, 1968b).

Exercise 22.2 Explore separable solutions in an arbitrary number N of spacetime dimensions.

22.4 Electrovac solutions from separation of Einstein's equations

As shown in §22.3, the assumptions of stationarity, axisymmetry, and separability, coupled with some other auxiliary assumptions (separability “in the simplest possible way” (Carter, 1968b)), and the 2+2 choice (22.24), imposes the form (22.1) of the line-element and the conditions (22.2) and (22.3). Given this form of the line-element, the Kerr solution and its electrovac cousins can be derived by separating the Einstein equations systematically.

22.4.1 Electrovac energy-momenta

The energy-momentum of a static radial electromagnetic field is

$$8\pi T_{mn}^e = \frac{Q_\bullet^2 + Q_\circ^2}{\rho^4} \text{diag}(1, -1, 1, 1). \quad (22.34)$$

The energy-momentum of a cosmological constant Λ is

$$8\pi T_{mn}^\Lambda = -\Lambda \eta_{mn} . \quad (22.35)$$

22.4.2 Separation of 8 Einstein equations with zero source

Given the form (22.1) of the line-element and the conditions (22.2) and (22.3), 4 of the 10 tetrad-frame Einstein components G_{mn} vanish identically:

$$G_{01} = G_{02} = G_{13} = G_{23} = 0 . \quad (22.36)$$

Of the remaining 6 Einstein components G_{mn} , the following 4 have zero electrovac source:

$$\rho^2 G_{12} = -2\sqrt{\Delta_x \Delta_y} \left[\rho \frac{\partial^2(1/\rho)}{\partial x \partial y} - \frac{3}{4(1 - \omega_x \omega_y)^2} \frac{d\omega_x}{dx} \frac{d\omega_y}{dy} \right] , \quad (22.37a)$$

$$\rho^2 G_{03} = \frac{\sqrt{\Delta_x \Delta_y}}{2\rho^2} \left[\frac{\partial}{\partial x} \left(\frac{\rho^2}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right) - \frac{\partial}{\partial y} \left(\frac{\rho^2}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right) \right] , \quad (22.37b)$$

$$\rho^2 (G_{00} + G_{11}) = \frac{2\Delta_x}{1 - \omega_x \omega_y} \left[\rho \frac{\partial}{\partial x} \left((1 - \omega_x \omega_y) \frac{\partial(1/\rho)}{\partial x} \right) + \frac{1}{4(1 - \omega_x \omega_y)} \left(\frac{d\omega_y}{dy} \right)^2 \right] , \quad (22.37c)$$

$$\rho^2 (G_{22} - G_{33}) = \frac{2\Delta_y}{1 - \omega_x \omega_y} \left[\rho \frac{\partial}{\partial y} \left((1 - \omega_x \omega_y) \frac{\partial(1/\rho)}{\partial y} \right) + \frac{1}{4(1 - \omega_x \omega_y)} \left(\frac{d\omega_x}{dx} \right)^2 \right] . \quad (22.37d)$$

If the conformal factor ρ^2 is supposed to separate as a sum of radial and angular parts, equation (22.2), then the homogeneous version of equation (22.37a) reduces to

$$\frac{d(\rho_x^2)}{d\omega_x} \frac{d(\rho_y^2)}{d\omega_y} - \frac{(\rho_x^2 + \rho_y^2)^2}{(1 - \omega_x \omega_y)^2} = 0 . \quad (22.38)$$

Series expansion of (22.38) leads to the result that

$$\rho^2 \equiv \rho_x^2 + \rho_y^2 = \frac{1 - \omega_x \omega_y}{(f_0 + f_1 \omega_x)(f_1 + f_0 \omega_y)} , \quad (22.39a)$$

$$\rho_x = \sqrt{\frac{g_0 - g_1 \omega_x}{(f_0 g_1 + f_1 g_0)(f_0 + f_1 \omega_x)}} , \quad \rho_y = \sqrt{\frac{g_1 - g_0 \omega_y}{(f_0 g_1 + f_1 g_0)(f_1 + f_0 \omega_y)}} , \quad (22.39b)$$

where f_0, f_1, g_0 , and g_1 are constants. At this point the constants g_0 and g_1 can be adjusted arbitrarily without affecting ρ : the overall normalization of g_0 and g_1 is cancelled by the normalizing factor of $1/\sqrt{f_0 g_1 + f_1 g_0}$ in ρ_x and ρ_y , and the relative sizes of g_0 and g_1 can be changed by adjusting an arbitrary constant in the split between ρ_x^2 and ρ_y^2 . Given the expression (22.39) for the conformal factor ρ , the Einstein component G_{03} , equation (22.37b), reduces to

$$\rho^2 G_{03} = \frac{\sqrt{\Delta_x \Delta_y}}{2(1 - \omega_x \omega_y)} \left[\frac{d\omega_x}{dx} \frac{d}{dx} \ln \left(\frac{d\omega_x/dx}{f_0 + f_1 \omega_x} \right) - \frac{d\omega_y}{dy} \frac{d}{dy} \ln \left(\frac{d\omega_y/dy}{f_1 + f_0 \omega_y} \right) \right] . \quad (22.40)$$

Homogeneous solution of this equation can be accomplished by separation of variables, setting each of the two terms inside square brackets, the first of which is a function only of x , while the second is a function only of y , to the same separation constant $2f_2$. The result is

$$\frac{d\omega_x}{dx} = 2\sqrt{(f_0 + f_1\omega_x) \left[g_0 + \frac{1}{f_0}(f_1g_0 + f_2)\omega_x \right]}, \quad (22.41a)$$

$$\frac{d\omega_y}{dy} = 2\sqrt{(f_1 + f_0\omega_y) \left[g_1 + \frac{1}{f_1}(f_0g_1 + f_2)\omega_y \right]}, \quad (22.41b)$$

for some constants g_0 and g_1 , which can be taken without loss of generality to equal those in the conformal factor (22.39). With the conformal factor ρ given by equation (22.39) and $d\omega_x/dx$ and $d\omega_y/dy$ given by equations (22.41), the Einstein components $G_{00} + G_{11}$ and $G_{22} - G_{33}$ reduce to

$$\rho^2 (G_{00} + G_{11}) = 2\Delta_x \frac{(f_2 + f_0g_1 + f_1g_0)(f_1 + f_0\omega_y)^2}{f_0f_1(1 - \omega_x\omega_y)^2}, \quad (22.42a)$$

$$\rho^2 (G_{22} - G_{33}) = 2\Delta_y \frac{(f_2 + f_0g_1 + f_1g_0)(f_0 + f_1\omega_x)^2}{f_0f_1(1 - \omega_x\omega_y)^2}. \quad (22.42b)$$

These vanish provided that the constant f_2 satisfies

$$f_2 = -(f_0g_1 + f_1g_0). \quad (22.43)$$

Inserting this value into equations (22.41) implies

$$\boxed{\frac{d\omega_x}{dx} = 2\sqrt{(f_0 + f_1\omega_x)(g_0 - g_1\omega_x)}}, \quad (22.44a)$$

$$\boxed{\frac{d\omega_y}{dy} = 2\sqrt{(f_1 + f_0\omega_y)(g_1 - g_0\omega_y)}}. \quad (22.44b)$$

The sign of the square root for $d\omega_x/dx$ is the same as that for ρ_x , while the sign of the square root for $d\omega_y/dy$ is the same as that for ρ_y .

22.4.3 Separation of the remaining 2 Einstein equations

Define Y_x and Y_y by

$$Y_x \equiv \frac{d\Delta_x}{dx} - \Delta_x \frac{d}{dx} \ln \left[(f_0 + f_1\omega_x) \frac{d\omega_x}{dx} \right], \quad (22.45a)$$

$$Y_y \equiv \frac{d\Delta_y}{dy} - \Delta_y \frac{d}{dy} \ln \left[(f_1 + f_0\omega_y) \frac{d\omega_y}{dy} \right]. \quad (22.45b)$$

In terms of Y_x and Y_y , the Einstein components $G_{00} - G_{11}$ and $G_{22} + G_{33}$ are

$$\begin{aligned} & \rho^2 (G_{00} - G_{11}) \\ &= \frac{1}{1 - \omega_x \omega_y} \left(Y_x \frac{d \ln \omega_x}{dx} - Y_y \frac{d \ln \omega_y}{dy} \right) + Y_x \frac{d}{dx} \ln \left(\frac{f_0 + f_1 \omega_x}{\omega_x} \right) - \frac{\partial Y_y}{\partial y} + Y_y \frac{d}{dy} \ln \left[\frac{\omega_y (f_1 + f_0 \omega_y)}{d\omega_y/dy} \right] , \end{aligned} \quad (22.46a)$$

$$\begin{aligned} & \rho^2 (G_{22} + G_{33}) \\ &= \frac{1}{1 - \omega_x \omega_y} \left(Y_x \frac{d \ln \omega_x}{dx} - Y_y \frac{d \ln \omega_y}{dy} \right) - Y_y \frac{d}{dy} \ln \left(\frac{f_1 + f_0 \omega_y}{\omega_y} \right) + \frac{\partial Y_x}{\partial x} - Y_x \frac{d}{dx} \ln \left[\frac{\omega_x (f_0 + f_1 \omega_x)}{d\omega_x/dx} \right] . \end{aligned} \quad (22.46b)$$

Homogeneous solutions of these equations can be found by supposing that Y_x is a function only of the radial coordinate radius x , while Y_y is a function only of the angular coordinate y , and by separating each of the equations as

$$\frac{1}{(1 - \omega_x \omega_y)} \left(\frac{f_0 h_0 + h_2 \omega_x + f_1 h_1 \omega_x^2}{\omega_x} - \frac{f_1 h_1 + h_2 \omega_y + f_0 h_0 \omega_y^2}{\omega_y} \right) - \frac{f_0 h_0 + h_3 \omega_x}{\omega_x} + \frac{f_1 h_1 + h_3 \omega_y}{\omega_y} = 0 , \quad (22.47)$$

for some constants h_0 , h_1 , h_2 , and h_3 . Separating each of equations (22.46) according to the pattern of equation (22.47) leads to the homogeneous solutions

$$Y_x = \frac{(f_0 + f_1 \omega_x)(h_0 + h_1 \omega_x)}{d\omega_x/dx} , \quad Y_y = \frac{(f_1 + f_0 \omega_y)(h_1 + h_0 \omega_y)}{d\omega_y/dy} . \quad (22.48)$$

Solutions including the energy-momentum of a static electromagnetic field fall out with little extra work. With appropriate boundary conditions, this is the Kerr-Newman solution. Solutions with $G_{00} = -G_{11} = G_{22} = G_{33}$, as is true for a static radial electromagnetic field, are found by taking the difference of equations (22.46) and separating that difference in the pattern of equation (22.47). The solution is a sum of a homogeneous solution (22.48) and a particular solution

$$Y_x = \frac{2(Q_\bullet^2 + \mathcal{Q}_\bullet^2)(f_0 + f_1 \omega_x)^2}{d\omega_x/dx} , \quad Y_y = 0 . \quad (22.49)$$

Inserting equations (22.49) into the Einstein expressions (22.46) yields Einstein components that have precisely the form (22.34) of the tetrad-frame energy-momentum tensor of a static radial electromagnetic field.

Similarly, solutions including vacuum energy, which has $G_{00} = -G_{11} = -G_{22} = -G_{33}$, can be found by separating the sum of equations (22.46) in the pattern of equation (22.47). A particular solution is

$$Y_x = \frac{2\Lambda}{f_1^2 d\omega_x/dx} , \quad Y_y = \frac{2\Lambda \omega_y^2}{f_1^2 d\omega_y/dy} . \quad (22.50)$$

Inserting equations (22.50) into the Einstein expressions (22.46) yields Einstein components that have precisely the form of a cosmological constant, $G_{mn} = -\Lambda \eta_{mn}$.

Solving equations (22.45a) and (22.45b) with Y_x and Y_y given by a sum of the homogeneous, electromagnetic, and vacuum contributions, equations (22.48), (22.49), and (22.50), yields the general electrovac

solution for the horizon and polar functions Δ_x and Δ_y ,

$$\Delta_x = (f_0 + f_1\omega_x) \left[(k_0 + k_1\omega_x) - \frac{2M_\bullet \sqrt{(f_0 + f_1\omega_x)(g_0 - g_1\omega_x)}}{(f_0g_1 + f_1g_0)^{3/2}} + \frac{(Q_\bullet^2 + \mathcal{Q}_\bullet^2)(f_0 + f_1\omega_x)}{f_0g_1 + f_1g_0} \right] - \frac{\Lambda(g_0 - g_1\omega_x)}{3f_1(f_0g_1 + f_1g_0)^2} , \quad (22.51a)$$

$$\Delta_y = (f_1 + f_0\omega_y) \left[(k_1 + k_0\omega_y) - \frac{2N_\bullet \sqrt{(f_1 + f_0\omega_y)(g_1 - g_0\omega_y)}}{(f_0g_1 + f_1g_0)^{3/2}} \right] + \frac{\Lambda\omega_y(g_1 - g_0\omega_y)}{3f_1(f_0g_1 + f_1g_0)^2} , \quad (22.51b)$$

where k_0 and k_1 are arbitrarily adjustable constants arising from the freedom of choice in the constants h_0 and h_1 of the homogeneous solution. The constant $M_\bullet$ in the expression (22.51a) for Δ_x is the black hole's mass. The constant $N_\bullet$ in the expression (22.51b) for Δ_y is the NUT parameter (Taub, 1951; Newman, Tamburino, and Unti, 1963; Stephani et al., 2003; Kagramanova et al., 2010), which is to the mass $M_\bullet$ as magnetic charge $\mathcal{Q}_\bullet$ is to electric charge $Q_\bullet$.

22.5 Electrovac solutions of Maxwell's equations

22.5.1 Solution of Maxwell's equations

Write the electromagnetic potential A_m in terms of a scaled electromagnetic potential $\mathcal{A}_m$, equation (23.2). Separability of the Hamilton-Jacobi equations requires, equations (22.23) and (22.24), that

$$\begin{aligned} \mathcal{A}_t, \quad \mathcal{A}_x & \text{ are functions of } x \text{ only,} \\ \mathcal{A}_y, \quad \mathcal{A}_\phi & \text{ are functions of } y \text{ only.} \end{aligned} \quad (22.52)$$

For the line-element (22.1), and with the conditions (22.52), the nonvanishing components of the tetrad-frame electromagnetic field F_{mn} are the radial electric E and magnetic B fields

$$E \equiv F_{10} = -\frac{1}{\rho^2} \left(\frac{d\mathcal{A}_t}{dx} + \frac{\omega_y \mathcal{A}_t - \mathcal{A}_\phi}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right) , \quad (22.53a)$$

$$B \equiv F_{23} = \frac{1}{\rho^2} \left(\frac{d\mathcal{A}_\phi}{dy} + \frac{\omega_x \mathcal{A}_\phi - \mathcal{A}_t}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right) . \quad (22.53b)$$

The remaining components of the electromagnetic field vanish identically,

$$F_{02} = F_{03} = F_{12} = F_{13} = 0 . \quad (22.54)$$

Since the electromagnetic field F_{mn} does not depend on either $\mathcal{A}_x$ or $\mathcal{A}_y$, these components are pure gauge, and can be set to zero,

$$\mathcal{A}_x = \mathcal{A}_y = 0 . \quad (22.55)$$

Since the electromagnetic field given by equations (22.53) and (22.54) is the curl of the potential, the field

automatically satisfies the source-free Maxwell's equations, The sourced Maxwell's equations are

$$D^m F_{mn} = 4\pi j_n . \quad (22.56)$$

Stationary solutions require vanishing current, $j_n = 0$. Two of the sourced Maxwell equations vanish identically, and the corresponding currents vanish automatically:

$$j_1 = j_2 = 0 . \quad (22.57)$$

Given the expressions (22.44) for $d\omega_x/dx$ and $d\omega_y/dy$, the remaining two sourced Maxwell's equations can be written

$$-\frac{\sqrt{\Delta_x}}{\rho^3} \left[\frac{\partial Z_t}{\partial x} + Z_t \frac{\partial}{\partial x} \ln \left(\frac{1}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right) - Z_\phi \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right] = 4\pi j_0 , \quad (22.58a)$$

$$-\frac{\sqrt{\Delta_y}}{\rho^3} \left[\frac{\partial Z_\phi}{\partial y} + Z_\phi \frac{\partial}{\partial y} \ln \left(\frac{1}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right) - Z_t \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right] = 4\pi j_3 , \quad (22.58b)$$

where Z_t and Z_ϕ are defined to be

$$Z_t \equiv \frac{d\omega_x}{dx} \frac{\partial}{\partial x} \left(\frac{\mathcal{A}_t}{d\omega_x/dx} \right) , \quad (22.59a)$$

$$Z_\phi \equiv \frac{d\omega_y}{dy} \frac{\partial}{\partial y} \left(\frac{\mathcal{A}_\phi}{d\omega_y/dy} \right) . \quad (22.59b)$$

The homogeneous solutions of equations (22.58) are

$$Z_t = Z_\phi = 0 . \quad (22.60)$$

Homogeneous solution of equations (22.59) yields

$$\frac{\mathcal{A}_t}{d\omega_x/dx} \equiv -\frac{Q_\bullet}{2(f_0 g_1 + f_1 g_0)} , \quad (22.61a)$$

$$\frac{\mathcal{A}_\phi}{d\omega_y/dy} = -\frac{\mathcal{Q}_\bullet}{2(f_0 g_1 + f_1 g_0)} , \quad (22.61b)$$

where $Q_\bullet$ and $\mathcal{Q}_\bullet$ are constants of integration, which can be interpreted as respectively the enclosed electric charge within radius x , and the enclosed magnetic charge above latitude y . Inserting the solutions (22.61) for $\mathcal{A}_m$ into the expressions (22.53) yields the electric and magnetic fields (22.11).

22.5.2 Separation of Maxwell's equations

The form (22.58) of the Maxwell equations for j_0 and j_3 assumed that $d\omega_x/dx$ and $d\omega_y/dy$ satisfy the equations (22.44) obtained by separating Einstein's equations. However, the Maxwell equations can also be separated directly, and the conditions (22.66) that result are consistent with the Einstein conditions (22.44).

If one provisionally supposes that $\mathcal{A}_\phi = 0$, then the Maxwell equation for the angular current j_3 is

$$\frac{\mathcal{A}_t \sqrt{\Delta_y}}{\rho^3 (1 - \omega_x \omega_y)^2} \left\{ \frac{d\omega_x}{dx} \left[(1 - \omega_x \omega_y) \frac{d \ln(\mathcal{A}_t / \omega_x)}{dx} + \frac{d \ln \omega_x}{dx} \right] + \frac{d\omega_y}{dy} \left[(1 - \omega_x \omega_y) \frac{d}{dy} \ln \left(\frac{d \ln \omega_y}{dy} \right) + \frac{d \ln \omega_y}{dy} \right] \right\} = 4\pi j_3 . \quad (22.62)$$

Conversely, if one provisionally supposes that $\mathcal{A}_t = 0$, then the Maxwell equation for the radial current j_0 is

$$\frac{\mathcal{A}_\phi \sqrt{\Delta_x}}{\rho^3 (1 - \omega_x \omega_y)^2} \left\{ \frac{d\omega_y}{dy} \left[(1 - \omega_x \omega_y) \frac{d \ln(\mathcal{A}_\phi / \omega_y)}{dy} + \frac{d \ln \omega_y}{dy} \right] + \frac{d\omega_x}{dx} \left[(1 - \omega_x \omega_y) \frac{d}{dx} \ln \left(\frac{d \ln \omega_x}{dx} \right) + \frac{d \ln \omega_x}{dx} \right] \right\} = 4\pi j_0 . \quad (22.63)$$

The homogeneous solutions of equations (22.62) and (22.63) prove to be the homogeneous solutions of the full equations without any restriction on $\mathcal{A}_t$ or $\mathcal{A}_\phi$. Equation (22.62) separates with 4 separation constants q_i as

$$(1 - \omega_x \omega_y) \left(\frac{-q_0 + q_3 \omega_x}{\omega_x} \right) + \frac{q_0 - 2q_1 \omega_x - q_2 \omega_x^2}{\omega_x} + (1 - \omega_x \omega_y) \left(\frac{-q_2 - q_3 \omega_y}{\omega_y} \right) + \frac{q_2 + 2q_1 \omega_y - q_0 \omega_y^2}{\omega_y} = 0 , \quad (22.64)$$

and equation (22.63) separates in a similar fashion, the vanishing of the second term inside braces in either of equations (22.62) or (22.63) requiring that

$$q_3 = q_1 . \quad (22.65)$$

The separated solutions for ω_x and ω_y are

$$\frac{d\omega_x}{dx} = 2\sqrt{q_0 - 2q_1 \omega_x - q_2 \omega_x^2} , \quad (22.66a)$$

$$\frac{d\omega_y}{dy} = 2\sqrt{q_2 + 2q_1 \omega_y - q_0 \omega_y^2} . \quad (22.66b)$$

These are consistent with the separated solution (22.44) found for Einstein's equations provided that

$$q_0 = f_0 g_0 , \quad 2q_1 = f_0 g_1 - f_1 g_0 , \quad q_2 = f_1 g_1 . \quad (22.67)$$

22.6 Λ -Kerr-Newman boundary conditions

The electrovac solutions of physical interest are those that go over to asymptotically flat space far from the black hole, at least in the absence of a cosmological constant. The condition of being far from the black hole can be interpreted as meaning where the influence of the mass and charge of the black hole becomes negligible. Inspection of expression (22.51a) for the radial horizon function Δ_x shows that the effect of mass and charge becomes negligible where $f_0 + f_1 \omega_x \rightarrow 0$. Expression (22.39) for the separable conformal factor ρ shows that the conformal factor diverges where $f_0 + f_1 \omega_x \rightarrow 0$, confirming that this location is indeed “at infinity.”

The quantity ω_x is the angular velocity at which the tetrad frame (which has been chosen to align with the principal frame) moves through the coordinates. This follows from the fact that the tetrad-frame 4-velocity relative to itself is by definition $u^m = \{1, 0, 0, 0\}$, so the coordinate frame velocity of the tetrad frame is $u^\mu = e_m^\mu u^m = e_0^\mu$, so the angular velocity of the tetrad frame is $d\phi/dt = e_0^\phi/e_0^t = \omega_x$. If the tetrad frame is not rotating through the coordinates at infinity, then the angular velocity ω_x vanishes at infinity. Since infinity is where $f_0 + f_1\omega_x$ vanishes, a tetrad frame that is corotating with the coordinates at infinity corresponds to $f_0 = 0$. Below, equations (22.76), it is shown that the situation where f_0 is nonzero differs by a coordinate transformation from the case where f_0 is zero. Thus f_0 may be set equal to zero without loss of generality.

Further conditions follow from requiring that the metric coefficients $g_{t\phi}$ and $g_{\phi\phi}$, equations (22.4), vanish at the poles of the rotation axis, $\theta = 0$ and π , to avoid singular behaviour at the poles. The vanishing of $g_{t\phi}$ and $g_{\phi\phi}$ at the poles requires that both ω_y and Δ_y must vanish at the poles.

Connection with familiar polar coordinates $\{r, \theta, \phi\}$ may be established by requiring that the metric coefficients (22.4) go over to their asymptotic expressions in the absence of a cosmological constant or NUT parameter,

$$g_{tt} \rightarrow -1, \quad g_{t\phi} \rightarrow 0, \quad g_{\phi\phi} \rightarrow r^2 \sin^2\theta \quad \text{as } r \rightarrow \infty. \quad (22.68)$$

The expressions (22.51) for the horizon functions Δ_x and Δ_y then imply that, in the absence of a cosmological constant or NUT parameter,

$$\Delta_y = \frac{\omega_y}{a} = \sin^2\theta, \quad \Delta_x \rightarrow \frac{\omega_x}{a} \rightarrow \frac{1}{r^2} \quad \text{as } r \rightarrow \infty, \quad (22.69)$$

where $a = 1/(f_1 k_0)$ is some constant, which proves to be the familiar spin parameter, and an overall normalization has been fixed by scaling the conformal factor to $\rho \rightarrow r$ at infinity, a natural choice. The normalization of ρ fixes $f_1 = a^{-1/2}$.

Integrating the relation (22.44b) between ω_y and y with $f_0 = 0$ establishes that ω_y is quadratic in y . Requiring that the polar part of the metric $g_{yy} dy^2 \equiv \rho^2 dy^2 / \Delta_y$ be nonsingular at the poles implies that y is proportional to $\cos\theta$ plus a constant that can be set to zero without loss of generality. Requiring that the polar metric go over to its asymptotic expression $g_{yy} dy^2 \rightarrow r^2 d\theta^2$ as $r \rightarrow \infty$ fixes the normalization

$$y = -\cos\theta, \quad (22.70)$$

where the sign is chosen so that $d\omega/dy$ has the same sign as ρ_y (eq. 22.75), in accordance with equation (22.44b). Expression (22.70) can be imposed also in the presence of a cosmological constant and a NUT parameter. For Λ -Kerr-Newman with no NUT parameter,

$$\omega_y = a \sin^2\theta. \quad (22.71)$$

Equation (22.71) is not true if the NUT parameter is nonvanishing, a case deferred to §22.7. The expressions (22.70) and (22.71) are consistent with the relation (22.44b) between them provided that $g_1 = ag_0$ and $g_0 = a/f_1$. The complete set of constants in equations (22.51) is

$$f_0 = 0, \quad f_1 = a^{-1/2}, \quad g_0 = a^{3/2}, \quad g_1 = a^{5/2}, \quad k_0 = a^{-1/2}, \quad k_1 = 0. \quad (22.72)$$

The radial variable x analogous to the angular variable y of equation (22.70) comes from solving equation (22.44a), $d\omega_x/dx = 2\sqrt{a\omega_x(1-a\omega_x)}$, which gives

$$x = \frac{1}{a} \operatorname{asin} \sqrt{a\omega_x} . \quad (22.73)$$

A pair of radial and angular variables that emerged naturally from the analysis, besides x and y , are ρ_x and ρ_y defined by equation (22.39b). In terms of ω_x , the radial variable is $\rho_x = \sqrt{a(1-a\omega_x)/\omega_x}$. It is conventional to define the radial coordinate r to be equal to ρ_x , which is consistent with the asymptotic behaviour $\rho \rightarrow r$ as $r \rightarrow \infty$, in which case

$$\omega_x = \frac{a}{R^2} , \quad R = \sqrt{r^2 + a^2} . \quad (22.74)$$

The radial and angular variables ρ_x and ρ_y are

$$\rho_x = r , \quad \rho_y = a \cos \theta . \quad (22.75)$$

This completes the derivation of the Λ -Kerr-Newman solutions.

22.6.1 There are no separable electrovac solutions that rotate at infinity

The Λ -Kerr-Newman boundary conditions in §22.6 took $f_0 = 0$, corresponding to the situation where the tetrad-frame is corotating with the coordinates at infinity, $\omega_x \rightarrow 0$ as $r \rightarrow \infty$. What happens if f_0 is nonzero? If f_0 is nonzero, then the tetrad rotates through the coordinates with some constant finite angular velocity ω_∞ at infinity. As argued at the beginning of §22.6, infinity is where $f_0 + f_1\omega_x = 0$. Thus a finite angular velocity at infinity corresponds to $f_0 = -f_1\omega_\infty$. However, the apparent rotation at infinity can be removed by transforming the azimuthal coordinate ϕ so that it corotates at infinity. The line-element can then be brought to standard electrovac form with $f_0 = 0$ by a coordinate transformation of the angular coordinate y . Specifically, the coordinate transformations

$$\phi' = \phi + \omega_\infty t , \quad dy' = (1 - \omega_\infty \omega_y) dy , \quad (22.76)$$

bring the line-element with $\omega_\infty \neq 0$ to the standard separable electrovac form with $\omega_\infty = 0$,

$$ds^2 = \rho^2 \left[-\frac{\Delta_x}{(1 - \omega'_x \omega'_y)^2} (dt - \omega'_y d\phi)^2 + \frac{dx^2}{\Delta_x} + \frac{dy'^2}{\Delta'_y} + \frac{\Delta'_y}{(1 - \omega'_x \omega'_y)^2} (d\phi' - \omega'_x dt)^2 \right] , \quad (22.77)$$

with primed quantities

$$\omega'_x \equiv \omega_x - \omega_\infty , \quad \omega'_y \equiv \frac{\omega_y}{1 - \omega_\infty \omega_y} , \quad \Delta'_y \equiv \frac{\Delta_y}{(1 - \omega_\infty \omega_y)^2} . \quad (22.78)$$

Notice that the physical location of north and south poles, at $\omega_y = \Delta_y = 0$, is unchanged by the choice of coordinates.

Thus, among separable electrovac solutions, there are no solutions that physically rotate at infinity.

22.7 Taub-NUT geometry

Spacetimes with a finite NUT parameter $N_\bullet$ were discovered by Taub (1951) and Newman, Tamburino, and Unti (1963). A nice review of the problematic nature of such spacetimes is given by Kagramanova et al. (2010).

Black holes with a finite NUT parameter $N_\bullet$ have the property that the coefficient ω_y , equation (22.83), does not vanish at one or both poles. As usual for a well-behaved azimuthally symmetric spacetime, when a geodesic passes infinitesimally close to a pole, the azimuthal angle ϕ along the geodesic jumps by $\pm\pi$, the sign depending on which side of the pole the geodesic is deemed to pass; but since the angle ϕ is periodic in 2π , the ambiguity in sign does not lead to a net ambiguity in the angle ϕ . The problem for NUT spacetimes, where ω_y does not vanish, is that when a geodesic passes infinitesimally close to a pole, the time coordinate t also jumps by $\pm\omega_y\pi$. The value of ω_y at a pole, $\theta = 0$ or π , is

$$\omega_y^\pm = 2N_\bullet(c_\bullet \mp 1) , \quad (22.79)$$

where the $\mp$ sign on the right hand side is $-$ at the north pole $\theta = 0$, and $+$ at the south pole $\theta = \pi$. The jump in the time coordinate is problematic, because it means that a particle passing through a pole leaps forwards or backwards in time, the choice of forwards or backwards depending on which side of the pole the geodesic is deemed to pass. Misner (1963) argued that the discontinuity in time could be solved by making the time coordinate t periodic with period $2\pi/|\omega_y^\pm|$. A periodic time coordinate does not of course describe the real Universe.

Associated with the discontinuity in the time coordinate t around a pole, black holes with a finite NUT parameter $N_\bullet$ have closed timelike curves circulating around one (if $c_\bullet = \pm 1$) or both (if $c_\bullet \neq \pm 1$) polar axes, so violate causality. NUT black holes are fun, but not physically realistic.

What accounts for the singular behaviour along poles? The answer is that there is a string of torsion along each pole. As described in §2.19.2 and examined further in §16.17, the Riemann and torsion tensors are the fields associated with the two gauge groups of general relativity, the Lorentz group and the translation group. The Riemann and torsion tensors describe how a frame respectively Lorentz-transforms and translates when parallel-transported around an infinitesimal loop. Torsion is sourced by spin angular-momentum Σ_{lmn} , §16.11. Although classic general relativity assumes that torsion vanishes, torsion should not be dismissed summarily, because spinor fields do carry spin angular-momentum that generates torsion, Exercise 16.7. However, the spinors familiar in the real world, such as electrons, are point-like and massive, whereas NUT strings are string-like and massless.

22.7.1 Taub-NUT line-element

When the NUT parameter $N_\bullet$ is finite, and with the boundary conditions that there are (north and south) poles at $\Delta_y = 0$, the conditions (22.72) generalize to

$$\begin{aligned} f_0 &= 0 , \quad f_1 = a^{-1/2} , \quad g_0 = a^{3/2} , \quad g_1 = a^{1/2}b^2 , \\ k_0 &= a^{-1/2} \left[1 - \frac{1}{3}\Lambda(a^2 - 2c_\bullet N_\bullet + N_\bullet^2) \right] , \quad k_1 = -2a^{-3/2}N_\bullet \left[N_\bullet + ac_\bullet + \frac{2}{3}a^2\Lambda N_\bullet(c_\bullet^2 - 1) \right] , \end{aligned} \quad (22.80)$$

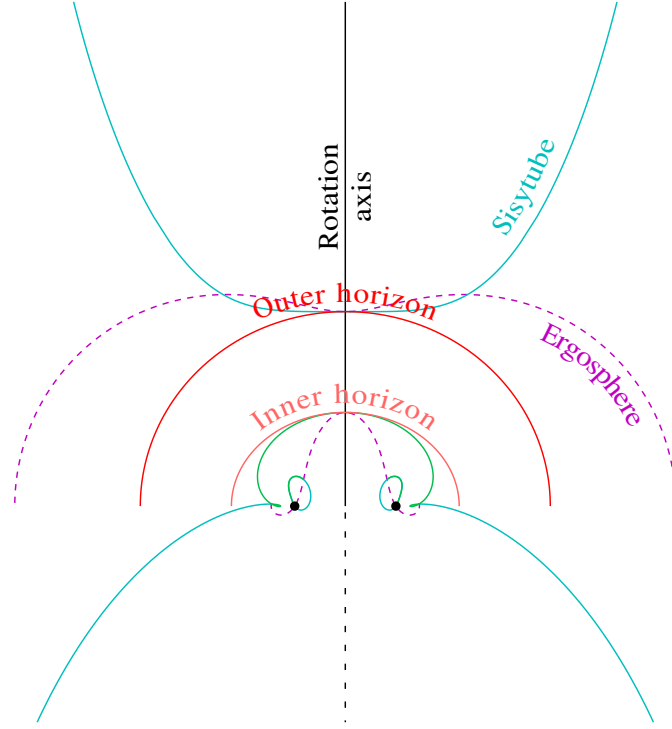


Figure 22.1 Geometry of a Kerr-NUT black hole, with NUT parameters $N_{\bullet} = 0.75M$ and $c_{\bullet} = -1$, and spin parameter $a = 1.2M$. The outer sisytube encircles the northern polar axis outside the outer ergosphere, while the inner sisytube encircles the extension of the northern axis into the Antiverse inside the inner ergosphere. The choice $c_{\bullet} = -1$ means that there is no sisytube around the southern polar axis. Dashed purple lines mark the boundaries $g_{tt} = 0$ of ergospheres, while green (between the ergospheres) and cyan (outside the ergospheres) lines mark $g_{\phi\phi} = 0$.

where

$$b \equiv \sqrt{a^2 + 2ac_{\bullet}N_{\bullet} + N_{\bullet}^2} . \quad (22.81)$$

Besides the NUT parameter $N_{\bullet}$, there is an additional constant, the auxiliary NUT parameter $c_{\bullet}$.

The resulting Taub-NUT line-element takes the separable form (22.1), with coefficients as follows. The radial and angular parts ρ_x and ρ_y of the conformal factor $\rho = \sqrt{\rho_x^2 + \rho_y^2}$ are

$$\rho_x \equiv r = b \cot(bx) , \quad \rho_y \equiv N_{\bullet} + a \cos \theta = N_{\bullet} - ay . \quad (22.82)$$

If $|N_{\bullet}| \leq |a|$, then there is a ring singularity where the Weyl curvature (22.88) diverges, at $\rho_x = \rho_y = 0$, corresponding to $r = 0$ and $\cos \theta = -N_{\bullet}/a$. There is no singularity if $|N_{\bullet}| > |a|$. The coefficients ω_x and ω_y

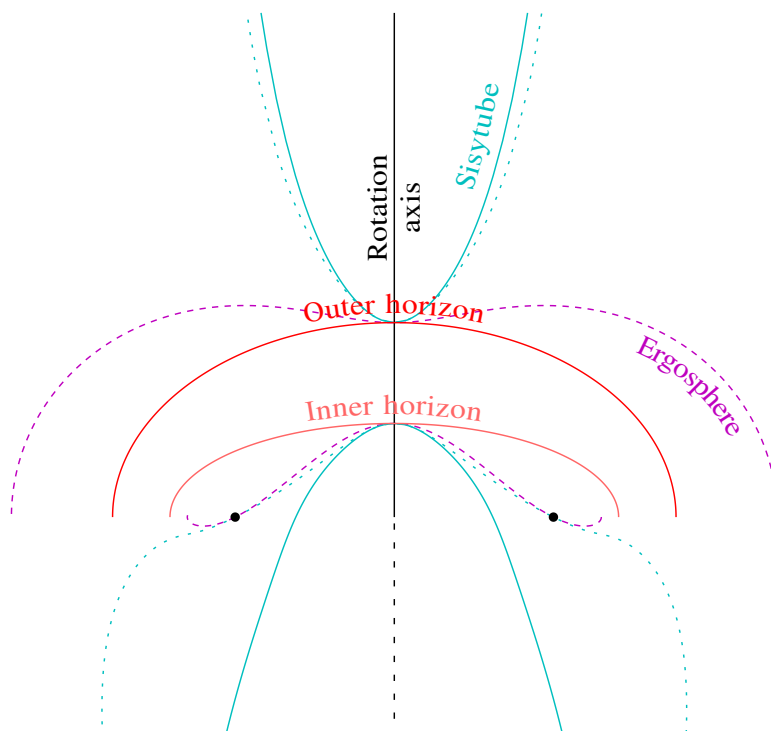


Figure 22.2 Similar to Figure 22.1, but with $c_{\bullet} = 0$ instead of $c_{\bullet} = -1$. Sisytubes encircle both the north (solid cyan lines) and south (dashed cyan lines) polar axes.

are

$$\omega_x = \frac{a}{R^2}, \quad R \equiv \sqrt{r^2 + b^2} = \frac{b}{\sin(bx)}, \quad \omega_y = a \sin^2 \theta + 2N_{\bullet}(c_{\bullet} - \cos \theta). \quad (22.83)$$

The radius R is everywhere positive, even if b is imaginary, so ω_x can also be taken to be everywhere positive (if a is negative, flip the poles, $y \rightarrow -y$, to make a positive). Notice that the NUT parameter $N_{\bullet}$ breaks spherical symmetry even if the black hole is nonrotating, $a = 0$, because ω_y , equation (22.83), cannot vanish at both poles as long as the NUT parameter $N_{\bullet}$ is nonzero. The combination $1 - \omega_x \omega_y$ satisfies

$$1 - \omega_x \omega_y = \frac{\rho^2}{R^2}, \quad (22.84)$$

which is always positive. The horizon and polar functions Δ_x and Δ_y are

$$\Delta_x = \frac{1}{R^4} \left[r^2 - 2M_{\bullet}r + a^2 + Q_{\bullet}^2 + \mathcal{Q}_{\bullet}^2 - N_{\bullet}^2 - \frac{1}{3}\Lambda \left(r^2 + (a - N_{\bullet})^2 \right) \left(r^2 + (a + N_{\bullet})^2 \right) \right], \quad (22.85a)$$

$$\Delta_y = \sin^2 \theta \left[1 - \frac{1}{3}a^2 \Lambda \sin^2 \theta + \frac{4}{3}\Lambda N_{\bullet}(N_{\bullet} + a \cos \theta) \right]. \quad (22.85b)$$

Poles occur where $\Delta_y = 0$, that is, at $\theta = 0$ or π . Horizons occur where $\Delta_x = 0$. For vanishing Λ , there are outer and inner horizons at

$$r_{\pm} = M_{\bullet} \pm \sqrt{M_{\bullet}^2 + N_{\bullet}^2 - Q_{\bullet}^2 - \mathcal{Q}_{\bullet}^2 - a^2} . \quad (22.86)$$

The horizon and polar functions Δ_x and Δ_y given by equations (22.85) do not agree with the earlier expressions (22.9) for nonzero Λ and vanishing $N_{\bullet}$, but this is not a misprint. The difference arises from an arbitrariness in the choice of homogeneous solution (the choice of k_0) when a cosmological constant Λ is present, equations (22.51). The tetrad-frame electromagnetic potential A_m is

$$A_m = \frac{1}{\rho} \left\{ -\frac{Q_{\bullet} r}{R^2 \sqrt{\Delta_x}}, 0, 0, -\frac{Q_{\bullet} (a \cos \theta + N_{\bullet})}{a \sqrt{\Delta_y}} \right\} . \quad (22.87)$$

The Weyl tensor has only a spin 0 component, and is, generalizing equation (22.12) for Λ -Kerr-Newman,

$$C = -\frac{1}{(\rho_x - I\rho_y)^3} \left(M_{\bullet} + IN_{\bullet} - \frac{Q_{\bullet}^2 + \mathcal{Q}_{\bullet}^2}{\rho_x + I\rho_y} \right) . \quad (22.88)$$

22.7.2 Sisytubes in Taub-NUT

Sisytubes, containing closed timelike curves, occur in regions where $g_{tt} \leq 0$ and $g_{\phi\phi} \leq 0$, Exercise 23.4. A sisytube encircles any pole where ω_y fails to vanish, since along poles, from equations (22.4) with $\Delta_y = 0$,

$$g_{\phi\phi} = -\frac{\rho^2 \omega_y^2 \Delta_x}{(1 - \omega_x \omega_y)^2} , \quad (22.89)$$

which is negative outside the horizon, $\Delta_x > 0$, unless ω_y vanishes. If the NUT parameter $N_{\bullet}$ is nonzero, then generically sisytubes encircle both poles, but for the special cases $c_{\bullet} = \pm 1$, a sisytube encircles only one of the two poles. The conclusion holds even when the black hole spin is zero, $a = 0$. For $\Lambda = 0$, the sisytube tends to a cylinder of constant radius at large distances from the black hole,

$$|r| \sin \theta \rightarrow |\omega_y| \rightarrow |2N_{\bullet} (c_{\bullet} \pm 1)| \quad \text{as } r \rightarrow \pm\infty , \quad (22.90)$$

in which the sign of ± 1 is the sign of y , namely $+$ at the south pole, $-$ at the north pole.

The critical velocity v_c at which there is a closed timelike curve is calculated in Exercise 23.4, equation (23.59). The sign of the critical velocity v_c is minus the sign of ω_y along its polar axes, which is the sign of $-N_{\bullet} (c_{\bullet} \pm 1)$.

Figure 22.1 illustrates the geometry for an uncharged Kerr-NUT black hole with $N_{\bullet} = 0.75M_{\bullet}$, $c_{\bullet} = -1$, and spin $a = 1.2M_{\bullet}$. Since $c_{\bullet} = -1$, a sisytube encircles the north pole but not the south pole. The sign of ω_y is negative along the north pole, so closed timelike curves circulate prograde.

Figure 22.2 is a Kerr-NUT black hole with the same parameters, except that $c_{\bullet} = 0$ in place of $c_{\bullet} = -1$. Here sisytubes enclose both north and south poles. The shapes of the sisytubes differ between north and south poles despite $c_{\bullet} = 0$. The north versus south asymmetry comes from the sign of the NUT parameter $N_{\bullet}$, which affects the polar function Δ_y , equation (22.85b). Closed timelike curves circulate the north pole prograde, the south pole retrograde.

22.7.3 Can the auxiliary NUT parameter $c_\bullet$ be adjusted by a coordinate transformation?

In §22.6.1 it was seen that, in the separable electrovac spacetimes being considered, any apparent rotation (of the tetrad frame through the coordinates) at infinity can be eliminated by a coordinate transformation (22.76) of the angular coordinates ϕ and y .

It might seem that a similar coordinate transformation of the t and x coordinates by

$$t' = t + \omega_0 \phi, \quad dx' = (1 - \omega_x \omega_0) dx, \quad (22.91)$$

would bring the line-element to the standard separable electrovac form (22.77) with primed quantities

$$\omega'_y \equiv \omega_y - \omega_0, \quad \omega'_x \equiv \frac{\omega_x}{1 - \omega_x \omega_0}, \quad \Delta'_x \equiv \frac{\Delta_x}{(1 - \omega_x \omega_0)^2}. \quad (22.92)$$

The coordinate transformation (22.91) would then allow the auxiliary NUT parameter $c_\bullet$ to be adjusted arbitrarily. For example, $c_\bullet$ could be set to ± 1 , or 0, or whatever other value one might prefer.

Ordinarily the choice of $c_\bullet$ would be dictated by physical reasons, which in the present cause would mean the absence of sisy tubes. Indeed, a sisy tube at the north pole can be eliminated by setting $c_\bullet = 1$; but then there is a sisy tube at the south pole. Likewise, a sisy tube at the south pole can be eliminated by setting $c_\bullet = -1$; but then there is a sisy tube at the north pole. One might perhaps choose $c_\bullet = 0$ as the most symmetric choice, but this still leaves the north-south asymmetry coming from the sign of $N_\bullet$, as illustrated in Figure 22.2. Evidently the problems of the Taub-NUT spacetime are fundamentally topological, and unavoidable.

Actually, the coordinate transformation (22.91) cannot be made freely, since it already encodes topological information. That is, axisymmetric identification $\phi \equiv \phi + 2\pi$ at fixed time t differs from axisymmetric identification at transformed time $t' = t + \omega_0 \phi$. Ordinarily the preferred time coordinate would be dictated by physical reasons, but again all choices are unphysical.

Trajectories in ideal rotating black holes

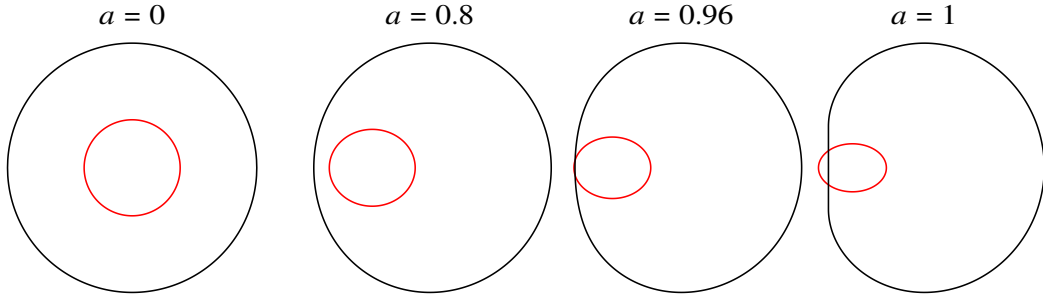


Figure 23.1 Silhouettes (black curves) of Kerr black holes with various spin parameters a , from left to right $a = 0$, 0.8, 0.96, and 1 (units $M = 1$), as observed in the equatorial plane from a far distance. The (red) ellipses show the horizons of the black holes as an indication of what the black holes would look like without any gravitational lensing. The silhouette is compressed on the approaching side and expanded on the receding side. See §23.14.

In the previous Chapter 22, the form of the Kerr-Newman line-element and its cousins was derived from the condition that geodesics are Hamilton-Jacobi separable. In this Chapter, the Hamilton-Jacobi equations are separated, and the trajectories of neutral and charged particles in the Kerr-Newman geometry are explored.

23.1 Hamilton-Jacobi equation

The Hamilton-Jacobi equation for a particle of mass m and electric charge q in the Λ -Kerr-Newman geometry can be brought to a simple form (23.8) by writing the covariant tetrad-frame momentum p_m of a particle in terms of a set of Hamilton-Jacobi parameters P_m ,

$$p_m \equiv \frac{1}{\rho} \left\{ \frac{P_t}{\sqrt{\Delta_x}}, \frac{P_x}{\sqrt{\Delta_x}}, \frac{P_y}{\sqrt{\Delta_y}}, \frac{P_\phi}{\sqrt{\Delta_y}} \right\}, \quad (23.1)$$

and the covariant tetrad-frame electromagnetic potential A_m in terms of a set of Hamilton-Jacobi potentials $\mathcal{A}_m$,

$$A_m \equiv \frac{1}{\rho} \left\{ \frac{\mathcal{A}_t}{\sqrt{\Delta_x}}, \frac{\mathcal{A}_x}{\sqrt{\Delta_x}}, \frac{\mathcal{A}_y}{\sqrt{\Delta_y}}, \frac{\mathcal{A}_\phi}{\sqrt{\Delta_y}} \right\}, \quad (23.2)$$

given by equation (22.10), which in turn follow from equations (22.55) and (22.61),

$$\mathcal{A}_m = \left\{ -\frac{Q_\bullet r}{R^2}, 0, 0, -Q_\bullet \cos \theta \right\}, \quad (23.3)$$

with $Q_\bullet$ and $Q_\bullet$ respectively the electric and magnetic charge of the black hole. The contravariant coordinate momenta $dx^\mu/d\lambda = d\{t, x, y, \phi\}/d\lambda = e_m^\mu p^k$ are related to the Hamilton-Jacobi parameters P_m by

$$\frac{dx^\mu}{d\lambda} = \frac{1}{\rho^2} \left\{ -\frac{P_t}{\Delta_x} + \frac{\omega_y P_\phi}{\Delta_y}, -P_x, P_y, -\frac{\omega_x P_t}{\Delta_x} + \frac{P_\phi}{\Delta_y} \right\}. \quad (23.4)$$

The tetrad-frame momenta p_m are related to the generalized momenta π_μ by $p_m = e_m^\mu \pi_\mu - qA_m$, which implies that the Hamilton-Jacobi parameters P_m are related to the canonical momenta π_μ by

$$P_t \equiv \pi_t + \pi_\phi \omega_x - q\mathcal{A}_t, \quad (23.5a)$$

$$P_x \equiv -\Delta_x \pi_x - q\mathcal{A}_x, \quad (23.5b)$$

$$P_y \equiv \Delta_y \pi_y - q\mathcal{A}_y, \quad (23.5c)$$

$$P_\phi \equiv \pi_\phi + \pi_t \omega_y - q\mathcal{A}_\phi. \quad (23.5d)$$

Time translation symmetry and axisymmetry imply that π_t and π_ϕ are constants of motion, equation (22.21),

$$\pi_t = -E, \quad \pi_\phi = L. \quad (23.6)$$

The separability conditions derived in §22.3 imply that

$$\begin{aligned} P_t, \quad P_x & \text{ are functions of } x \text{ only,} \\ P_y, \quad P_\phi & \text{ are functions of } y \text{ only.} \end{aligned} \quad (23.7)$$

In terms of the Hamilton-Jacobi parameters P_m , the Hamilton-Jacobi equation (22.18) is

$$-\frac{P_t^2 + P_x^2}{\Delta_x} + \frac{P_y^2 + P_\phi^2}{\Delta_y} = -m^2 \rho^2. \quad (23.8)$$

Separability for massive particles, $m \neq 0$, requires that the conformal factor ρ separate as equation (22.2).

The Hamilton-Jacobi equation (23.8) then separates as

$$-\left(\frac{P_t^2 + P_x^2}{\Delta_x} + m^2 \rho_x^2 \right) = \frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho_y^2 = \mathcal{K}, \quad (23.9)$$

with $\mathcal{K}$ a separation constant, the **Carter constant**. The separated Hamilton-Jacobi equations (23.9) imply that

$$P_x = \pm \sqrt{P_t^2 - (\mathcal{K} + m^2 \rho_x^2) \Delta_x} , \quad (23.10a)$$

$$P_y = \pm \sqrt{-P_\phi^2 + (\mathcal{K} - m^2 \rho_y^2) \Delta_y} . \quad (23.10b)$$

In spherical spacetimes, $a = 0$, the Carter constant equals the total angular momentum squared of the geodesic. Mathematically, the existence of the constant derives from a Killing tensor (23.16) associated with separability. From the expression (23.4) for the coordinate momenta $dx^\mu/d\lambda$, the trajectory of a freely falling particle follows from integrating $dy/dx = -P_y/P_x$, equivalent to the implicit equation

$$-\frac{dx}{P_x} = \frac{dy}{P_y} . \quad (23.11)$$

Again from expression (23.4), the time and azimuthal coordinates t and ϕ along the trajectory are then obtained by quadratures,

$$dt = \frac{P_t dx}{P_x \Delta_x} + \frac{\omega_y P_\phi dy}{P_y \Delta_y} , \quad d\phi = \frac{\omega_x P_t dx}{P_x \Delta_x} + \frac{P_\phi dy}{P_y \Delta_y} . \quad (23.12)$$

Again from expression (23.4), the affine parameter λ along the trajectory satisfies $d\lambda/\rho^2 = -dx/P_x = dy/P_y$, so similarly reduces to quadratures,

$$d\lambda = -\frac{\rho_x^2 dx}{P_x} + \frac{\rho_y^2 dy}{P_y} . \quad (23.13)$$

In the limiting case of trajectories at constant latitude y , where dy/P_y is zero divided by zero, expressions for t , ϕ , and λ along the trajectory are obtained by replacing $dy/P_y \rightarrow -dx/P_x$ in equations (23.12) and (23.13). Similarly for circular trajectories, where dx/P_x is zero divided by zero, expressions for t , ϕ , and λ along the trajectory are obtained by replacing $dx/P_x \rightarrow -dy/P_y$.

23.2 Particle with magnetic charge

The above Hamilton-Jacobi equations were for a test particle of mass m and electric charge q , but no magnetic charge. Whereas electric charge is a scalar, magnetic charge is a pseudoscalar. Equations of motion for a magnetic charge are obtained by taking the Hodge dual of those for an electric charge, effectively swapping the roles of the electric and magnetic fields. The Hodge dual of the electromagnetic field (22.11), obtained by multiplying by the pseudoscalar I , equation (13.24), is the same expression with the electric $Q_\bullet$ and magnetic $\mathcal{Q}_\bullet$ charges of the black hole exchanged according to

$$Q_\bullet \rightarrow -\mathcal{Q}_\bullet , \quad \mathcal{Q}_\bullet \rightarrow Q_\bullet . \quad (23.14)$$

Coupling the pseudoscalar magnetic charge to the dual electromagnetic field gives an extra minus sign, $I^2 = -1$. Thus the Hamilton-Jacobi equations generalize to a particle with both electric charge q_e and

magnetic charge q_m by replacing

$$qQ_\bullet \rightarrow q_e Q_\bullet + q_m Q_\bullet, \quad qQ_\bullet \rightarrow q_e Q_\bullet - q_m Q_\bullet, \quad (23.15)$$

in the expressions (23.5a) and (23.5d) for P_t and P_ϕ . The particle magnetic charge q_m is set to zero hereafter; it can be reincorporated by making the transformation (23.15) of constants.

23.3 Killing vectors and Killing tensor

The Kerr-Newman geometry is stationary and axisymmetric. As such it has two Killing vectors $\mathbf{e}_t$ and $\mathbf{e}_\phi$, §7.32. The symmetries imply conservation of energy $E = -\pi_t$ and azimuthal angular momentum $L = \pi_\phi$ of freely falling particles, equations (22.21).

The separability of the Kerr-Newman geometry means that it also has a Killing tensor K^{mn} . The Hamilton-Jacobi equation (23.9) can be written in terms of the tetrad-frame momenta p_m , equation (23.1), as

$$K^{mn} p_m p_n = \mathcal{K}, \quad (23.16)$$

where K^{mn} is

$$K^{mn} = \text{diag}(-\rho_y^2, \rho_y^2, \rho_x^2, \rho_x^2). \quad (23.17)$$

The Killing tensor K^{mn} satisfies Killing's equation

$$D_{(k} K_{mn)} = 0. \quad (23.18)$$

23.4 Turnaround

The squared Hamilton-Jacobi parameters P_x^2 and P_y^2 can be regarded as effective radial and angular potentials. The coordinates x and y of a freely falling particle are constrained to move within the regions where the potentials P_x^2 and P_y^2 are positive. The trajectory of a freely falling particle turns around in x where $P_x = 0$, and turns around in y where $P_y = 0$. That trajectories turn around at these points can be seen from equation (23.11), which with the expressions (23.10) for P_x and P_y can be written

$$\frac{d\lambda}{\rho^2} = -\frac{dx}{\sqrt{P_t^2 - (\mathcal{K} + m^2 \rho_x^2) \Delta_x}} = \frac{dy}{\sqrt{-P_\phi^2 + (\mathcal{K} - m^2 \rho_y^2) \Delta_y}}. \quad (23.19)$$

At points where the polar function vanishes, $\Delta_y = 0$, the Hamilton-Jacobi equation (23.9) implies that

$$P_y = P_\phi = 0 \quad \text{at } \Delta_y = 0. \quad (23.20)$$

Consequently trajectories must turn around in y if they hit $\Delta_y = 0$. Since the Weyl curvature is finite at $\Delta_y = 0$, there is no singularity at $\Delta_y = 0$. Rather, the points where Δ_y vanishes define the (north and south) poles of the geometry. Trajectories can pass through the poles, but they must turn around in latitude y when they do so.

Table 23.1: Signs of P_t and P_x in various regions of the Kerr-Newman geometry

Region	Sign
Universe, Wormhole, Antiverse	$P_t < 0$
Parallel Universe, Parallel Wormhole, Parallel Antiverse	$P_t > 0$
Black Hole	$P_x < 0$
White Hole	$P_x > 0$
Horizon, Inner Horizon	$P_t = P_x < 0$
Parallel Horizon, Parallel Inner Horizon	$-P_t = P_x < 0$
Antihorizon, Inner Antihorizon	$-P_t = P_x > 0$
Parallel Antihorizon, Parallel Inner Antihorizon	$P_t = P_x > 0$

23.5 Constraints on the Hamilton-Jacobi parameters P_t and P_x

Horizons divide the spacetime into regions where the Hamilton-Jacobi parameters P_t and P_x satisfy certain conditions. The Hamilton-Jacobi equation (23.8) rearranges to

$$P_t^2 - P_x^2 = \left(\frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho^2 \right) \Delta_x . \quad (23.21)$$

This shows that the Hamilton-Jacobi parameters P_t and P_x must satisfy

$$\begin{aligned} |P_t| &> |P_x| && \text{if } \Delta_x > 0 , \\ |P_t| &= |P_x| && \text{if } \Delta_x = 0 , \\ |P_t| &< |P_x| && \text{if } \Delta_x < 0 . \end{aligned} \quad (23.22)$$

The Hamilton-Jacobi parameters must be continuous, including across horizons. Thus P_t must have the same sign everywhere throughout any connected region where Δ_x is positive, which in the Kerr-Newman geometry means either outside the outer horizon or inside the inner horizon. Similarly P_x must have the same sign everywhere throughout any connected region where Δ_x is negative, which in the Kerr-Newman geometry means between the outer and inner horizons.

Outside the outer horizon, in the Universe region of the Kerr-Newman geometry, Figure 9.6, the time parameter P_t must be negative, reflecting the fact that the time coordinate t must be timelike and increasing with the proper time of any particle. The radial parameter P_x can be either positive (outfalling) or negative (infalling).

Inside the outer horizon, in the Black Hole region of the geometry, the radial parameter P_x must be negative, reflecting the fact that the radius is timelike and decreasing with the proper time of any particle. The time parameter P_t can be either positive (outgoing) or negative (ingoing).

Particles that cross the outer horizon are necessarily infalling and ingoing at the horizon, with $P_t = P_x$ negative. The Hamilton-Jacobi parameters are finite and continuous across the horizon. The expression (23.1)

shows that the tetrad-frame momenta p_0 and p_1 are proportional to $1/\sqrt{\Delta_x}$, and therefore diverge at the horizon, where $\Delta_x = 0$. The divergence is the origin of the inflationary instability at the inner horizon discussed in Chapter 24.

Table 23.1 lists the constraints on the Hamilton-Jacobi parameters P_t and P_x in each of the regions of the Penrose diagram of Figure 9.6.

23.6 Principal null congruences

The middle expression of equation (23.9) shows that the Carter constant $\mathcal{K}$ is necessarily positive. The vanishing of the Carter constant,

$$\mathcal{K} = 0 , \quad (23.23)$$

defines a special set of geodesics, called the **principal outgoing** and **ingoing null congruences**. A congruence is a space-filling, non-overlapping set of geodesics. The geodesics on the principal congruences are null, $m = 0$, and satisfy

$$P_y = P_\phi = 0 . \quad (23.24)$$

They further satisfy $P_t^2 = P_x^2$. Outgoing and ingoing geodesics are distinguished by the relative signs of P_t and P_x ,

$$\begin{aligned} P_t &= -P_x & \text{outgoing} \\ P_t &= P_x & \text{ingoing} . \end{aligned} \quad (23.25)$$

Photons that hold steady on the horizon are members of the outgoing principal null congruence.

The condition $P_\phi = 0$ implies that the ratio of angular momentum $L = \pi_\phi$ to energy $E = -\pi_t$ on the principal null congruences is

$$\frac{L}{E} = \omega_y . \quad (23.26)$$

The affine parameter λ along a principal null congruence satisfies

$$d\lambda \propto \frac{\rho^2 dx}{P_t/E} = \frac{\rho^2 dx}{1 - \omega_x \omega_y} = \frac{dr}{\sqrt{f_0 g_1 + f_1 g_0 (f_1 + f_0 \omega_y)}} , \quad (23.27)$$

where f_i and g_i are the constants of the general electrovac solution, §22.4. As argued in §22.6.1, a coordinate transformation allows the constant f_0 to be set to zero without loss of generality. Thus the affine parameter, which is defined only up to a normalization and a shift, along the principal null congruences can be taken to be

$$\lambda = \pm r . \quad (23.28)$$

The line-element (22.1) defines a tetrad (the Boyer-Lindquist tetrad) that is aligned with the principal null

congruences. By definition, an object at rest in the tetrad frame has tetrad-frame 4-velocity $u^m = \{1, 0, 0, 0\}$. The coordinate 4-velocity u^μ of the tetrad frame through the coordinates is

$$u^\mu = e_0^\mu = \frac{1}{\rho\sqrt{\Delta_x}}\{1, 0, 0, \omega_x\} . \quad (23.29)$$

Thus the principal tetrad frame is at rest in x and y , but rotates through the coordinates at angular velocity $d\phi/dt = \omega_x$ about the black hole.

23.7 Carter integral $\mathcal{Q}$

It is common to replace the Carter constant $\mathcal{K}$ by the Carter integral $\mathcal{Q}$ defined by

$$\mathcal{K} = \mathcal{Q} + \left. \frac{P_\phi^2}{\Delta_y} \right|_{\rho_y=0} , \quad (23.30)$$

which has the property that $\mathcal{Q} = 0$ for orbits in the equatorial plane, $\rho_y = 0$. For Λ -Kerr-Newman, the Carter integral is

$$\mathcal{Q} = \mathcal{K} - (L - aE)^2 . \quad (23.31)$$

Exercise 23.1 Boundary of the region between the horizons visible to an infaller at the inner horizon. Between the outer and inner horizons, all trajectories must fall inward. What geodesics have the largest angular motion between the horizons? Hence determine the boundary of the region between the horizons visible to an infaller who reaches the inner horizon.

Solution. The boundary between regions visible and invisible to an infaller between the horizons is set by photons that have the largest angular motion. The Carter constant $\mathcal{K}$, which for a spherical black hole equals the total angular momentum squared, is infinite for photons with the largest angular motion,

$$\mathcal{K} = \infty . \quad (23.32)$$

The ratio J of azimuthal to total angular momentum is

$$J \equiv \frac{L}{\sqrt{\mathcal{K}}} = \sin \theta_J = \pm \sqrt{1 - y_J^2} , \quad (23.33)$$

where $y_J \equiv -\cos \theta_J$ is the largest latitude reached on the geodesic ($J = 0$ and $y_J = \pm 1$ for geodesics that reach the poles, $J = \pm 1$ and $y_J = 0$ on the equator). In the limit of infinite $\mathcal{K}$, the Hamilton-Jacobi solution (23.11) simplifies to

$$\begin{aligned} - \int_{x_0} \frac{dx}{P_x} &= \frac{1}{\sqrt{\mathcal{K}}} \int_{r_0} \frac{dr}{\sqrt{(r - e_-)(e_+ - r)}} = \frac{2}{\sqrt{\mathcal{K}}} \left[\operatorname{atan} \sqrt{\frac{r - e_-}{e_+ - r}} \right]_{r_0}^r \\ &= \int_{y_0} \frac{dy}{P_y} = \frac{1}{\sqrt{\mathcal{K}}} \int_{y_0} \frac{dy}{\sqrt{(y + y_J)(y_J - y)}} = \frac{2}{\sqrt{\mathcal{K}}} \left[\operatorname{atan} \sqrt{\frac{y + y_J}{y_J - y}} \right]_{y_0}^y , \end{aligned} \quad (23.34)$$

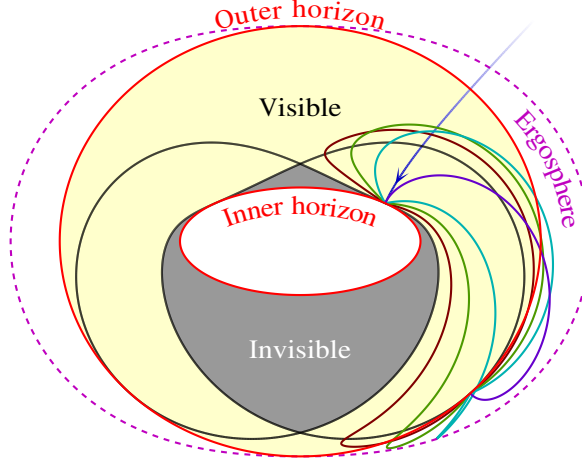


Figure 23.2 Regions between the outer and inner horizons visible and invisible to an infaller who reaches the inner horizon of a Kerr black hole. The black hole here has spin parameter $a = 0.8M_{\bullet}$, and the infaller (blue line) falls with zero angular momentum $L = 0$ along a geodesic at latitude $\theta_0 = 45^\circ$. The boundary of the (in)visible region is set by null trajectories with infinite Carter constant $\mathcal{K}$ and asymptotically small azimuthal angular momenta J ; such trajectories, dragged around by the rotation of the black hole, spiral around the black hole divergingly many times, hugging the $J = 0$ surface (black lines) all the way between the outer and inner horizons. Null geodesics with infinite $\mathcal{K}$ and various azimuthal angular momenta J are shown on the radius-latitude r - θ plane (projected to azimuth $\phi = 0$) as solid lines passing through the observer at the inner horizon. To reach the observer at latitude θ_0 , $|J|$ cannot exceed $J_{\max} = \sin \theta_0$; the trajectories shown have $|J|/J_{\max}$ equal to 0 (black), $\frac{1}{4}$ (brown), $\frac{1}{2}$ (green), $\frac{3}{4}$ (cyan), and 1 (purple). For each $|J| < J_{\max}$ there are two trajectories, which approach the observer from respectively lower and higher latitude. Null geodesics with infinite $\mathcal{K}$ start as outgoing at the outer horizon, all at the same latitude antipodal, $\theta = \pi/2 - \theta_0$, to the final latitude θ_0 at the inner horizon (but not antipodal in azimuth), switch to ingoing inside the ergosphere, fall back through the outer horizon, down to the inner horizon. Compare to Figure 7.1 for a Schwarzschild black hole.

where $e_{\pm}$ are the radii of the ergospheres at latitude θ_J , equation (9.18),

$$e_{\pm} = M \pm \sqrt{M^2 - Q^2 - a^2 \cos^2 \theta_J} . \quad (23.35)$$

Define the angle ψ by

$$\frac{\psi}{\sqrt{\mathcal{K}}} \equiv - \int_{x_0} \frac{dx}{P_x} = \int_{y_0} \frac{dy}{P_y} , \quad (23.36)$$

which starts at $\psi = 0$ at the position x_0, y_0 . A null trajectory passing through an emitter or observer at radius $r_0 \equiv a \cot(ax_0)$ and latitude $y_0 \equiv -\cos \theta_0$ is given parametrically in terms of the angle ψ by

$$r = M - \frac{1}{2}(e_+ - e_-) \cos(\psi + \psi_r) = e_- + (e_+ - e_-) \sin^2 \left(\frac{\psi + \psi_r}{2} \right) , \quad y = -y_J \cos(\psi + \psi_y) , \quad (23.37)$$

where

$$\psi_r \equiv \text{acos} \left(-\frac{2(r_0 - M)}{e_+ - e_-} \right), \quad \psi_y \equiv \text{acos} \left(-\frac{y_0}{y_J} \right). \quad (23.38)$$

The radius r and polar height y vary periodically with the angle ψ , with period 2π .

Figure 23.2 illustrates trajectories in radius r and latitude θ (azimuth ϕ being projected to 0) for an infaller who reaches the inner horizon at latitude $\theta_0 = \pi/4$, equivalent to inclination angle $\pi/2 - \theta_0$ above the equator.

Compared to the case of a nonrotating (Schwarzschild) black hole shown in Figure 7.1, visibility inside the black hole is complicated by the fact that the rotating black hole drags frames around it. The Boyer-Linquist time and azimuthal coordinates t and ϕ along a geodesic, equations (23.12), are not the best choice at or inside the horizon, because they diverge at the horizon thanks to the $1/\Delta_x$ factor. A better choice of time and azimuthal coordinates are coordinates t_{pn} and ϕ_{pn} that corotate with the black hole along the ingoing or outgoing principal null congruences, equations (23.207),

$$t_{\text{pn}} \equiv -X^E = t \pm \int \frac{dx}{\Delta_x}, \quad \phi_{\text{pn}} \equiv X^L = \phi \pm \int \frac{\omega_x dx}{\Delta_x}, \quad (23.39)$$

where the upper (+) sign is outgoing, the lower (−) sign ingoing. The Boyer-Lindquist time coordinate t , which is spacelike between horizons, decreases along ingoing geodesics inside the outer horizon, per the Penrose diagram 8.7. The corotating time coordinate t_{pn} similarly decreases along ingoing geodesics; indeed it decreases not just between the horizons, but between the outer and inner ergospheres. Along outgoing geodesics, the ingoing corotating coordinate t_{pn} decreases between horizons, diverges at horizons, and increases outside horizons. The corotating azimuthal coordinate ϕ_{pn} decreases or increases in tandem with t_{pn} .

Between outer and inner horizons, the Hamilton-Jacobi parameter P_t can be either positive (outgoing) or negative (ingoing). The t component $p_t = P_t/(\rho\sqrt{\Delta_x})$ of the tetrad-frame momentum, equation (23.1), diverges at horizons, where $\Delta_x = 0$. At horizons, outer or inner, ingoing and outgoing frames stream through each other at the speed of light; ingoing and outgoing frames see each other infinitely blueshifted. The divergence occurs only at horizons; ingoing and outgoing frames that lie strictly between horizons stream through each other at less than the speed of light. In a real black hole formed from gravitational collapse, infallers do not encounter an actual divergence at the outer horizon. All objects that fall through the outer horizon are necessarily ingoing, Table 23.1. Ingoing infallers with sufficiently positive angular momentum L (prograde) can turn around between the horizons and become outgoing at the inner horizon; and ingoing objects can emit outgoing light inside the outer horizon. But light, whether ingoing or outgoing, tends to move away from the outer horizon, so light does not collect there. By contrast, ingoing and outgoing light just above the inner horizon moves towards the inner horizon. Even the tiniest streams of ingoing and outgoing light at the inner horizon ignite the mass inflation instability there.

For definiteness, consider an infaller who is ingoing when they reach the inner horizon. The ingoing infaller sees ingoing light to have finite blue/redshift. Ingoing light with infinite $\mathcal{K}$ necessarily has negative azimuthal angular momentum J (and outgoing light has positive J). Ingoing corotating coordinates t_{pn} and ϕ_{pn} along

a geodesic are, from equations (23.12),

$$t_{\text{pn}} = \int_{x_0} \left(\frac{P_t}{P_x} - 1 \right) \frac{dx}{\Delta_x} + \int_{y_0} \frac{P_\phi}{P_y} \frac{\omega_y dy}{\Delta_y} , \quad \phi_{\text{pn}} = \int_{x_0} \left(\frac{P_t}{P_x} - 1 \right) \frac{\omega_x dx}{\Delta_x} + \int_{y_0} \frac{P_\phi}{P_y} \frac{dy}{\Delta_y} . \quad (23.40)$$

If outgoing instead of ingoing frames were considered, then outgoing coordinates t_{pn} and ϕ_{pn} should be used, with a + instead of − sign in equations (23.39).

Abbreviate

$$r_\diamond \equiv \frac{1}{2}(r_+ - r_-) = \sqrt{M^2 - Q^2 - a^2} , \quad e_\diamond \equiv \frac{1}{2}(e_+ - e_-) = \sqrt{M^2 - Q^2 - a^2 y_J^2} , \quad (23.41)$$

and

$$c_r \equiv \cos(\psi + \psi_r) , \quad s_r \equiv \sin(\psi + \psi_r) , \quad (23.42a)$$

$$c_y \equiv \cos(\psi + \psi_y) , \quad s_y \equiv \sin(\psi + \psi_y) . \quad (23.42b)$$

The polar contribution to t_{pn} , equation (23.40), can be incorporated into the radial contribution by using $-dx/P_x = dy/P_y$ to recast it as

$$\int_{y_0} \frac{P_\phi \omega_y}{\Delta_y} \frac{dy}{P_y} = - \int_{x_0} \frac{P_\phi \omega_y}{\Delta_y} \frac{dx}{P_x} = - \int_{x_0} aJ \frac{dx}{P_x / \sqrt{\mathcal{K}}} , \quad (23.43)$$

valid in the limit of infinite $\mathcal{K}$, when $P_\phi = L = \sqrt{\mathcal{K}}J$. The time and azimuthal angles t_{pn} and ϕ_{pn} along null geodesics with $\mathcal{K} = \infty$ integrate to

$$t_{\text{pn}} = -(\pm\xi - \xi_0 + 2\xi_\diamond i_r) , \quad (23.44a)$$

$$\phi_{\text{pn}} = -(\pm\alpha - \alpha_0 + 2\alpha_\diamond i_r + \beta - \beta_0) , \quad (23.44b)$$

in which overall − signs have been introduced because, as described following equations (23.39), t_{pn} and ϕ_{pn} decrease along ingoing geodesics between the outer and inner ergospheres; for outgoing geodesics with respect to outgoing t_{pn} and ϕ_{pn} , the overall signs in equations (23.44) would be +. The r -dependent functions ξ , α in equations (23.44) vary periodically with $\psi + \psi_r$, with period 2π . The $\pm$ signs in $\pm\xi$ and $\pm\alpha$ in equations (23.44) are the sign of $\sin(\psi + \psi_r)$, and the integer $i_r \equiv [(\psi + \psi_r)/\pi]$ is the π phase of $\psi + \psi_r$, which increments by 1 at the same time as the $\pm$ sign of $\pm\xi$ or $\pm\alpha$ flips sign, ensuring that t_{pn} and ϕ_{pn} vary smoothly as ξ and α pass through their radial extrema. The functions ξ , α , and β in equations (23.44) are

$$\xi \equiv \text{sign}(c_r) \left(\frac{2M^2 - Q^2}{2r_\diamond} \ln \left(\frac{e_\diamond + a|J||s_r| - r_\diamond|c_r|}{e_\diamond + a|J||s_r| + r_\diamond|c_r|} \right) + e_\diamond|c_r| \right) - 2M \ln \left(1 + \frac{e_\diamond|s_r|}{a|J|} \right) , \quad (23.45a)$$

$$\xi_\diamond \equiv \frac{2M^2 - Q^2}{r_\diamond} \ln \left(\frac{e_\diamond + r_\diamond}{a|J|} \right) - e_\diamond , \quad (23.45b)$$

$$\alpha \equiv \text{sign}(c_r) \frac{a}{r_\diamond} \ln \left(\frac{r_\diamond|s_r| + a|J||c_r|}{r_\diamond + e_\diamond|c_r|} \right) , \quad (23.45c)$$

$$\alpha_\diamond \equiv \frac{a}{r_\diamond} \ln \left(\frac{r_\diamond + e_\diamond}{a|J|} \right) , \quad (23.45d)$$

$$\beta \equiv \text{acot}(|J|c_y, s_y) . \quad (23.45e)$$

The r -dependent functions ξ and α have minima respectively $\xi = -\xi_\diamond$ and $\alpha = -\alpha_\diamond$ at $\psi + \psi_r = 0$ where $r = e_-$, pass through 0 at $\psi + \psi_r = \pi/2$ where $r = M$, and reach maxima $\xi = \xi_\diamond$ and $\alpha = \alpha_\diamond$ at $\psi + \psi_r = \pi$ where $r = e_+$. At outer or inner horizons, $s_r = \pm a|J|/e_\diamond$ and $c_r = \pm\sqrt{1-s_r^2}$, with signs set by the phase of $\psi + \psi_r$. The sequence as $\psi + \psi_r$ increases from 0 to 2π is, inner horizon at $\psi + \psi_r \in [0, \pi/2]$, outer horizon at $\psi + \psi_r \in [\pi/2, \pi]$, outer horizon at $\psi + \psi_r \in [\pi, 3\pi/2]$, inner horizon at $\psi + \psi_r \in [3\pi/2, 2\pi]$. The values of ξ and α at outer and inner horizons are

$$\xi_\pm \equiv \pm \left(\frac{2M^2 - Q^2}{r_\diamond} \ln \left(\frac{e_\diamond}{a|J|} \right) - r_\diamond \right) - 2M \ln 2, \quad (23.46a)$$

$$\alpha_\pm \equiv \pm \frac{a}{r_\diamond} \ln \left(\frac{e_\diamond}{a|J|} \right), \quad (23.46b)$$

where the upper and lower signs are at the outer and inner horizons respectively.

An observer at r_0 watching an emitter at r sees the emitter as they used to be a time t_{pn} ago. During that time the corotating coordinate ϕ_{pn} at r has changed by $\omega_x t_{\text{pn}}$, so the observer sees the emitter at azimuthal coordinate

$$\phi_{\text{pn}} - \omega_x t_{\text{pn}}. \quad (23.47)$$

In particular, an ingoing observer at the inner horizon r_- watching an ingoing light ray with infinite $\mathcal{K}$ falling through the outer horizon r_+ sees the emitter at azimuthal coordinate

$$\begin{aligned} \phi_{\text{pn}} - \omega_x t_{\text{pn}} &= - \left((\alpha_+ - \alpha_-) + (\beta_+ - \beta_-) - \frac{a}{r_+^2 + a^2} (\xi_+ - \xi_-) \right) \\ &= \frac{4a}{r_+^2 + a^2} \left(M \ln \left(\frac{e_\diamond}{a|J|} \right) - r_\diamond \right) - (\beta_+ - \beta_-) \end{aligned} \quad (23.48)$$

relative to their own azimuthal coordinate ϕ_{pn} . The relative azimuthal coordinate (23.48) is not coordinate gauge-invariant, but the difference in the perceived azimuthal coordinates of different emitters at the same outer horizon radius is coordinate gauge-invariant. A takeaway from equation (23.48) is that the perceived azimuthal coordinate $\phi_{\text{pn}} - \omega_x t_{\text{pn}}$ diverges logarithmically as $|J| \rightarrow 0$, indicating that an observer at the inner horizon sees geodesics with tiny J (and infinite $\mathcal{K}$) circulate the black hole many times, dragged around by the black hole's rotation. Numerical experiment confirms that the boundary of the (in)visible regime is determined by geodesics with the smallest azimuthal angular momentum, $J \sim 0$. The geodesics hug the $J = 0$ surface in radius r and latitude y , spiralling around in azimuth ϕ_{pn} , all the way from outer to inner horizon. The resulting (in)visible region is illustrated in Figure 23.2.

Exercise 23.2 Near the Kerr-Newman singularity. This exercise reveals that among ideal black holes, the Schwarzschild geometry is exceptional, not typical, in having a gravitationally attractive singularity. Explore the behaviour of trajectories of test particles in the vicinity of the Kerr-Newman singularity, where $\rho \rightarrow 0$ (that is, where $r = 0$ and $ay = 0$). Under what conditions does a test particle reach the singularity?

1. Argue that for a particle to reach the singularity at $y = 0$, positivity of P_y^2 requires that

$$\mathcal{Q} \geq 0, \quad (23.49)$$

where $\mathcal{Q}$ is the Carter integral defined by equation (23.31).

2. Argue that for a particle to reach the singularity at $r = 0$, positivity of P_x^2 requires that

$$Q_\bullet^2(L - aE)^2 + (Q_\bullet^2 + a^2)\mathcal{Q} \leq 0. \quad (23.50)$$

3. Schwarzschild case: show that if $Q_\bullet = 0$ and $a = 0$, then a particle reaches the singularity provided that the mass of the black hole is positive, $M_\bullet > 0$.
4. Reissner-Nordström case: show that if $Q_\bullet^2 > 0$ and $a = 0$, then a particle can reach the singularity only if it has zero angular momentum, $\mathcal{Q} = L = 0$, and if the particle's charge exceeds its mass,

$$|q| \geq |m|. \quad (23.51)$$

In particular, a neutral particle reaches the singularity only if it has zero angular momentum and is massless.

5. Kerr case: show that if $Q_\bullet = 0$ but $a^2 > 0$, then a particle can reach the singularity only if it is moving in the equatorial plane ($y = 0$ and $\mathcal{Q} = 0$), and provided that the mass of the black hole is positive, $M_\bullet > 0$. [Hint: Show that if the particle is not already in the equatorial plane at $y = 0$, then the equation of motion for dy/dx shows that the particle never reaches $y = 0$.]
6. Kerr-Newman case: show that if $Q_\bullet^2 > 0$ and $a^2 > 0$, then a particle can reach the singularity only if $L = aE$ and it is moving in the equatorial plane, and if the particle's charge-to-mass is large enough,

$$|q| \geq |m| \sqrt{\frac{Q_\bullet^2 + a^2}{Q_\bullet^2}}, \quad (23.52)$$

which generalizes the Reissner-Nordström condition (23.51).

Solution. Equation (23.49) comes from

$$\mathcal{K} = \frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho_y^2 \geq \frac{P_\phi^2}{\Delta_y}, \quad (23.53)$$

and taking the limit $y \rightarrow 0$. Equation (23.50) comes from

$$R^4 \{-P_t^2 + [\mathcal{Q} + (L - aE)^2 + m^2 \rho_x^2] \Delta_x\} = -R^4 P_x^2 \leq 0, \quad (23.54)$$

and taking the limit $r \rightarrow 0$.

Exercise 23.3 When must t and ϕ progress forwards on a geodesic? Under what circumstances must the time coordinate t or azimuthal angle ϕ progress forwards along a geodesic?

1. Show that, in regions where $\Delta_x \geq 0$,

$$\frac{P_\phi^2}{\Delta_y} \leq \mathcal{K} \leq \frac{P_t^2}{\Delta_x}. \quad (23.55)$$

Hence show that in the Universe, Wormhole, and Antiverse regions outside the horizons, where $P_t < 0$,

$$\frac{dt}{d\lambda} \geq \frac{(1 - \omega_x \omega_y)^2 \sqrt{\mathcal{K}}}{\rho^4 \sqrt{\Delta_x \Delta_y} (\omega_y \sqrt{\Delta_x} + \sqrt{\Delta_y})} g_{\phi\phi} = -\frac{\sqrt{\mathcal{K} \Delta_x \Delta_y}}{\omega_y \sqrt{\Delta_x} + \sqrt{\Delta_y}} g^{tt}, \quad (23.56)$$

and

$$\frac{d\phi}{d\lambda} \geq \frac{(1 - \omega_x \omega_y)^2 \sqrt{\mathcal{K}}}{\rho^4 \sqrt{\Delta_x \Delta_y} (\sqrt{\Delta_x} + \omega_x \sqrt{\Delta_y})} g_{tt} = -\frac{\sqrt{\mathcal{K} \Delta_x \Delta_y}}{\sqrt{\Delta_x} + \omega_x \sqrt{\Delta_y}} g^{\phi\phi}. \quad (23.57)$$

Conclude that, in the $P_t < 0$ regions outside the outer and inner horizons, the time coordinate t must progress forwards if $g_{\phi\phi} \geq 0$, which is true outside the sisytube, while the azimuthal angle ϕ must progress forwards if $g_{tt} \geq 0$, which is true between the outer and inner ergospheres.

2. Argue that in the Parallel Universe, Parallel Wormhole, and Parallel Antiverse regions outside the horizons, where $P_t > 0$, the inequalities (23.56) and (23.57) hold with the left hand sides replaced by $d(-t)/d\lambda$ and $d(-\phi)/d\lambda$. Hence conclude that the time coordinate t must progress backwards outside the sisytube, while the azimuthal coordinate ϕ must progress backwards between the ergospheres.

Exercise 23.4 Inside the sisytube. The sisytube, §9.10, is the region where $g_{tt} < 0$ and $g_{\phi\phi} \leq 0$. The condition $g_{tt} < 0$ ensures that t is a timelike coordinate.

1. Consider a massive particle moving along a circular path at constant radius and latitude ($dx = dy = 0$), with tetrad-frame 4-velocity $u^m = \gamma\{1, 0, 0, v\}$, where γ is a Lorentz γ -factor and v the corresponding 3-velocity. A closed timelike curve (CTC) occurs when the time coordinate t is constant along the curve, $dt/d\tau = 0$. What is the critical velocity v_c for a closed timelike curve? Is the closed timelike curve prograde or retrograde? What is the condition on the velocity v of the particle for it to go backwards in time t ?
2. Can the circular path be a geodesic?

Solution.

1. The coordinate 4-velocity u^μ in terms of the tetrad-frame 4-velocity $u^m = \gamma\{1, 0, 0, v\}$ is

$$u^\mu \equiv \frac{dx^\mu}{d\tau} = e_m{}^\mu u^m = \frac{\gamma}{\rho} \left\{ \frac{1}{\sqrt{\Delta_x}} + \frac{v\omega_y}{\sqrt{\Delta_y}}, 0, 0, \frac{\omega_x}{\sqrt{\Delta_x}} + \frac{v}{\sqrt{\Delta_y}} \right\}. \quad (23.58)$$

The particle proceeds forwards or backwards in time t according to the sign of u^t . The particle follows a closed timelike curve if $u^t = 0$, which happens when its tetrad-frame velocity v takes the critical value

$$v_c = -\frac{\sqrt{\Delta_y}}{\omega_y \sqrt{\Delta_x}}. \quad (23.59)$$

The sisytube condition $g_{\phi\phi} \leq 0$ ensures that $|v_c| \leq 1$. The critical velocity equals the speed of light, $|v_c| = 1$, at the boundary $g_{\phi\phi} = 0$ of the sisytube. The critical velocity v_c , equation (23.59) is negative (retrograde) if ω_y is positive, and positive (prograde) if ω_y is negative. In Λ -Kerr-Newman ω_y is always positive, so closed timelike curves in the sisytube are retrograde. The situation with a finite NUT parameter $N_\bullet$ has been commented on in §22.7.2. At the critical velocity (23.59), the particle's azimuthal coordinate velocity u^ϕ is

$$u^\phi = -\frac{\gamma_c(1 - \omega_x \omega_y)}{\rho \omega_y \sqrt{\Delta_x}} = \frac{\gamma_c v_c (1 - \omega_x \omega_y)}{\rho \sqrt{\Delta_y}}, \quad (23.60)$$

whose sign is the same as that of the critical velocity v_c . The particle goes backwards in time if the absolute value of its velocity exceeds the critical value (23.59),

$$|v| > |v_c| . \quad (23.61)$$

2. No.

Exercise 23.5 Gödel's Universe. Gödel's Universe has a separable line-element of the form (22.1) with $\rho = 1$, $\omega_x = 0$, and $\Delta_x = 1$, thus

$$ds^2 = - (dt - \omega_y d\phi)^2 + dx^2 + \frac{dy^2}{\Delta_y} + \Delta_y d\phi^2 . \quad (23.62)$$

Find the components of the tetrad-frame energy-momentum tensor. Show that the tetrad-frame energy-momentum tensor is diagonal provided that ω_y is linear in y . Show that the energy-momentum is constant everywhere provided that Δ_y is quadratic in y . Recast the angular variable y in terms of the conventional polar angle θ defined by $r d\theta \equiv dy/\sqrt{\Delta_y}$, for some constant radius r . Show that there is a sisytube ($g_{tt} < 0$ and $g_{\phi\phi} \leq 0$) when y , or equivalently θ , exceeds some critical value. Is Gödel's Universe self-consistent in the sense that the rest frame of the fluid is everywhere geodesic? Explore Gödel's Universe.

Solution. The nonvanishing components of the tetrad-frame energy-momentum tensor are

$$G_{00} = \frac{3}{4} \left(\frac{d\omega_y}{dy} \right)^2 - \frac{1}{2} \frac{d\Delta_y}{dy} , \quad (23.63a)$$

$$G_{03} = \frac{\sqrt{\Delta_y}}{2} \frac{d^2\omega_y}{dy^2} , \quad (23.63b)$$

$$G_{11} = -\frac{1}{4} \left(\frac{d\omega_y}{dy} \right)^2 + \frac{1}{2} \frac{d\Delta_y}{dy} , \quad (23.63c)$$

$$G_{22} = G_{33} = \frac{1}{4} \left(\frac{d\omega_y}{dy} \right)^2 . \quad (23.63d)$$

The condition for the metric to be diagonal is $G_{03} = 0$, which is true provided that ω_y is linear in y . The transverse components $G_{22} = G_{33} = 8\pi p_\perp$ of the pressure are then automatically constant, and ω_y is

$$\omega_y = 2\sqrt{8\pi p_\perp}(y + r^2) , \quad (23.64)$$

where the constant of integration has been chosen a posteriori so that ω_y vanishes at the origin $\theta = 0$ of the cylindrical geometry, as will become evident in any of equations (23.68b), (23.70b), or (23.71b). The result of this choice is that the Gödel line-element (23.62) is asymptotically Minkowski near the origin. Notice that it is not possible to adjust the constant by a coordinate gauge transformation $t \rightarrow t + a\phi$ for some a , since the time t must be unchanged under $\phi \rightarrow \phi + 2\pi$. Notice also that the transverse pressure is the square of something, so must be positive, $p_\perp \geq 0$. The time and radial components $G_{00} = 8\pi\rho$ and $G_{11} = 8\pi p$ of the energy-momentum are constant provided that Δ_y is quadratic in y ,

$$\Delta_y = r^2 - (y/r)^2 , \quad (23.65)$$

where the possible term proportional to y has been set to zero by a coordinate gauge transformation of y , and r is a constant radius defined by

$$r \equiv \frac{1}{\sqrt{8\pi(\rho - 3p_{\perp})}} . \quad (23.66)$$

The radial pressure p is related to the density ρ and transverse pressure $p_{\perp}$ by

$$p = -\rho + 2p_{\perp} . \quad (23.67)$$

The fluid is perfect (equal radial and transverse pressures) in the particular case of an ultrahard fluid, with $\rho = p = p_{\perp}$. The angular coordinate y can be recast in terms of the conventional polar angle θ , giving

$$r d\theta \equiv \frac{dy}{\sqrt{\Delta_y}} , \quad y = -r^2 \cos \theta , \quad (23.68a)$$

$$\omega_y = 4\sqrt{8\pi p_{\perp}} r^2 \sin^2(\theta/2) , \quad \Delta_y = r^2 \sin^2 \theta . \quad (23.68b)$$

The resulting line-element is

$$ds^2 = - (dt - \omega_y d\phi)^2 + dx^2 + r^2(d\theta^2 + \sin^2 \theta d\phi^2) . \quad (23.69)$$

For $\rho - 3p_{\perp} > 0$, the radius r , equation (23.66), is positive, and the spatial geometry looks like the product of the real line with a 2-sphere of radius r . For $\rho - 3p_{\perp} < 0$, the radius r is imaginary, and the spatial geometry looks like the product of the real line with a 2-hypersphere of negative curvature $\kappa = -1/r^2$. In this case equations (23.68) are replaced by

$$r d\theta \equiv \frac{dy}{\sqrt{\Delta_y}} , \quad y = r^2 \cosh \theta , \quad (23.70a)$$

$$\omega_y = 2\sqrt{8\pi p_{\perp}}(y - r^2) = 4\sqrt{8\pi p_{\perp}} r^2 \sinh^2(\theta/2) , \quad \Delta_y = (y/r)^2 - r^2 = r^2 \sinh^2 \theta . \quad (23.70b)$$

In the limiting case of $\rho - 3p_{\perp} = 0$, the angular geometry becomes flat, and

$$r d\theta \equiv \frac{dy}{\sqrt{\Delta_y}} , \quad y = \frac{1}{2} r^2 \theta^2 , \quad (23.71a)$$

$$\omega_y = 2\sqrt{8\pi p_{\perp}} y = \sqrt{8\pi p_{\perp}} r^2 \theta^2 , \quad \Delta_y = r^2 \theta^2 . \quad (23.71b)$$

The metric component $g_{\phi\phi}$, whose vanishing defines the boundary of a sisytube, is

$$g_{\phi\phi} = \Delta_y - \omega_y^2 , \quad (23.72)$$

which is negative for

$$\theta > \begin{cases} 2 \operatorname{atan} \sqrt{\frac{\rho - 3p_{\perp}}{4p_{\perp}}} & \rho > 3p_{\perp} \\ 0 & \rho = 3p_{\perp} \\ 2 \operatorname{atanh} \sqrt{\frac{3p_{\perp} - \rho}{4p_{\perp}}} & \rho < 3p_{\perp} \end{cases} . \quad (23.73)$$

Yes, the solution is self-consistent. The rest frame of the fluid is the same as the rest frame of the tetrad, since the energy-momentum is diagonal in the tetrad rest frame. The split between ρ_x and ρ_y in $\rho_x^2 + \rho_y^2 = \rho^2 = 1$ can be taken to be $\rho_x = 1$ and $\rho_y = 0$. Rest geodesics satisfy $\pi_t = -m$, $\pi_\phi = m\omega_y$, and $\mathcal{K} = 0$, yielding $P_x = P_y = P_\phi = 0$ and $P_t = -m$, whence $p^k = \{m, 0, 0, 0\}$.

23.8 Penrose process

As first pointed out by Penrose, trajectories in the Kerr-Newman geometry can have negative energy E outside the horizon. In Newtonian gravity, gravitational energy is negative. If the gravitational binding energy of a particle more than cancels the kinetic energy of the particle, then the particle is in a bound orbit. In general relativity, the binding energy of a particle can be so great that in effect it cancels not only the kinetic energy, but also the rest mass energy of the particle. Such particles have negative energy.

It is possible to reduce the mass $M_\bullet$ of the black hole by dropping negative energy particles into the black hole. This process of extracting mass-energy from the black hole is called the **Penrose process**.

Exercise 23.6 Negative energy trajectories outside the horizon. Under what conditions can a test particle have negative energy, $E < 0$, outside the outer horizon of a Kerr-Newman black hole?

1. Argue that the negativity of P_t outside the outer horizon implies that $aL + qQr$ must be negative for the energy E to be negative. Show that, more stringently, negative E requires that

$$aL + qQr \leq -R^2 \sqrt{\left(\frac{L^2}{\Delta_y} + m^2 \rho^2\right) \Delta_x} . \quad (23.74)$$

2. Argue that for an uncharged particle, $q = 0$, negative energy trajectories exist only inside the ergosphere.
3. Do negative energy trajectories exist outside the ergosphere for a charged particle?
4. For the Penrose process to work, the negative energy particle must fall through the outer horizon, where $\Delta_x = 0$. Can this happen? Must it happen?

Solution. See the end of §23.17.

23.9 Constant latitude trajectories in the Kerr-Newman geometry

For simplicity, the next several sections, up to and including §23.20, are restricted to Kerr-Newman black holes with zero magnetic charge, $Q_\bullet = 0$, and zero cosmological constant, $\Lambda = 0$.

A trajectory is at constant latitude if it is at constant polar angle θ , or equivalently at constant $y \equiv -\cos \theta$,

$$y = \text{constant} . \quad (23.75)$$

Constant latitude orbits occur where the angular potential P_y^2 , equation (23.10b), not only vanishes, but is

an extremum,

$$P_y^2 = \frac{dP_y^2}{dy} = 0 , \quad (23.76)$$

the derivative being taken with the constants of motion E , L , and $\mathcal{K}$ of the orbit being held fixed. The condition $P_y^2 = 0$ sets the value of the Carter integral $\mathcal{K}$. Solving $dP_y^2/dy = 0$ yields the condition between energy E and angular momentum L

$$E = \pm \sqrt{m^2 + \frac{L^2}{a^2 \sin^4 \theta}} . \quad (23.77)$$

Solutions at any polar angle θ and any angular momentum L exist, ranging from $E = \pm m$ at $L = 0$, to $E = \pm L/(a \sin^2 \theta)$ at $L \rightarrow \pm \infty$. The solutions with $L = 0$ are those of the freely falling observers that define the Doran coordinate system, §9.18. The solutions with $L \rightarrow \infty$ define the principal null congruences discussed in §23.6.

23.10 Circular orbits in the Kerr-Newman geometry

For simplicity, this section 23.10 is restricted to Kerr-Newman black holes with zero magnetic charge and cosmological constant,

$$Q_\bullet = \Lambda = 0 . \quad (23.78)$$

For brevity, the black hole subscripts will be dropped from the black hole mass and electric charge $M_\bullet$ and $Q_\bullet$,

$$M_\bullet = M , \quad Q_\bullet = Q . \quad (23.79)$$

23.10.1 Condition for a circular orbit

An orbit can be termed circular if it is at constant radius r ,

$$r = \text{constant} . \quad (23.80)$$

It is convenient to call such an orbit circular even if the orbit is at finite inclination (not confined to the equatorial plane) about a rotating black hole, and therefore follows the surface of a spheroid (in Boyer-Lindquist coordinates).

Orbits turn around in r , reaching periapsis or apoapsis, where the radial potential P_x^2 , equation (23.10a), vanishes. Circular orbits occur where the radial potential P_x^2 not only vanishes, but is an extremum,

$$P_x^2 = \frac{dP_x^2}{dr} = 0 , \quad (23.81)$$

the derivative being taken with the constants of motion E , L , and $\mathcal{Q}$ of the orbit being held fixed. Circular

orbits may be either stable or unstable. The stability of a circular orbit is determined by the sign of the second derivative of the potential

$$\frac{d^2 P_x^2}{dr^2}, \quad (23.82)$$

with $-$ for stable, $+$ for unstable circular orbits. Marginally stable orbits occur where $d^2 P_x^2 / dr^2 = 0$.

Circular orbits occur not only in the equatorial plane, but at general inclinations. The inclination of an orbit can be characterized by the maximum latitude $y_{\max}$, or equivalently the minimum polar angle $\theta_{\min}$, that the orbit reaches. An astronomer would call $\arcsin(y_{\max}) = \pi/2 - \theta_{\min}$ the inclination angle of the orbit. It is convenient to define an inclination parameter α by

$$\alpha \equiv y_{\max}^2 = \cos^2 \theta_{\min}, \quad (23.83)$$

which lies in the interval $[0, 1]$. Equatorial orbits, at $y = 0$, correspond to $\alpha = 0$, while polar orbits, those that go over the poles at $y = \pm 1$, correspond to $\alpha = 1$.

The maximum latitude $y_{\max}$ reached by an orbit occurs at the turnaround point $P_y = 0$. Inserting this condition into equation (23.10b) allows the Carter constant $\mathcal{K}$, or equivalently the Carter integral $\mathcal{Q}$, equation (23.31), to be eliminated in favour of the inclination parameter α , equation (23.83)

$$\mathcal{Q} = \mathcal{K} - (L - aE)^2 = \alpha \left[a^2(m^2 - E^2) + \frac{L^2}{1 - \alpha} \right]. \quad (23.84)$$

Equation (23.84) is a quadratic equation in α , so has two roots for α at fixed E , L , and $\mathcal{Q}$. The quadratic is $\mathcal{Q}(1 - \alpha) + \alpha(1 - \alpha)a^2(E^2 - m^2) - \alpha L^2$, which equals $\mathcal{Q}$ at $\alpha = 0$, and $-L^2$ at $\alpha = 1$. Therefore there is one root in $\alpha \in [0, 1]$ if $\mathcal{Q} > 0$, and two roots if $\mathcal{Q} < 0$ (given that, for an orbit to exist, at least one root must lie in $\alpha \in [0, 1]$),

$$\begin{aligned} \mathcal{Q} > 0 & \quad 1 \text{ root in } \alpha \in [0, 1], \\ \mathcal{Q} < 0 & \quad 2 \text{ roots in } \alpha \in [0, 1]. \end{aligned} \quad (23.85)$$

For one root in $\alpha \in [0, 1]$, the orbit has only a maximum latitude; for two roots, the orbit has a minimum as well as a maximum latitude. All the equations in what follows hold true for α the inclination parameter at an extremum, whether maximum or minimum.

The energy per unit mass of a particle at infinity must exceed its rest mass, $|E/m| \geq 1$ (E is positive in the Universe, negative in the Parallel Universe). A particle with energy less than its rest mass, $|E/m| < 1$, cannot go to infinity, and is said to be bound. Equation (23.84) implies that the Carter integral $\mathcal{Q}$ is positive for bound orbits, $\mathcal{Q} \geq 0$ (with $\mathcal{Q} = 0$ for equatorial orbits, $\alpha = 0$). Therefore all bound orbits have only a maximum latitude; they all pass through the equator.

Exercise 23.7 Circular geodesics at constant latitude? Are there circular geodesics at constant latitude? Are there circular geodesics in the equatorial plane.

Solution. Inserting the constant latitude conditions on $\mathcal{K}$ and E from §23.9 into the quadratic equation (23.84) for the inclination parameter α shows that a necessary and sufficient condition for a circular

orbit at constant latitude is $\mathcal{Q} = 0$. Teodesics in the equatorial plane have inclination parameter $\alpha = 0$, and all such geodesics have $\mathcal{Q} = 0$.

23.11 General solution for circular orbits

The general solution for circular orbits of a test particle of arbitrary electric charge q in the Kerr-Newman geometry is as follows. For vanishing electric charge, see §23.12.

The rest mass m of the test particle can be set equal to unity, $m = 1$, without loss of generality. Circular orbits of particles with zero rest mass, $m = 0$, discussed in §23.13 below, occur in cases where the circular orbits for massive particles attain infinite energy and angular momentum.

In the radial potential P_x^2 , equation (23.10a), eliminate the Carter integral $\mathcal{K}$ in favour of the inclination parameter α using equation (23.84). Furthermore, eliminate the energy $E \equiv -\pi_t$ in favour of P_t , equation (23.5a). The radial derivatives $d^n P_x^2 / dr^n$ must be taken *before* E is replaced by P_t , since E is a constant of motion, whereas P_t varies with r . For Kerr-Newman, the expression (23.5a) for P_t is

$$P_t = -E + \frac{aL}{R^2} + \frac{qQr}{R^2}. \quad (23.86)$$

In accordance with Table 23.1, solutions with negative P_t correspond to orbits in the Universe, Wormhole, or Antiverse parts of the Kerr-Newman geometry in the Penrose diagram of Figure 9.6, while solutions with positive P_t correspond to orbits in their Parallel counterparts. If only the Universe region is considered, then P_t is necessarily negative. By contrast, the energy E can be either positive or negative in the same region of the Kerr-Newman geometry (the energy E is negative for orbits of sufficiently large negative angular momentum L inside the ergosphere of the Universe). Circular orbits cannot occur between the outer and inner horizons (why not?).

The condition $P_x^2 = 0$, equation (23.81), is a quadratic equation in the azimuthal angular momentum $L \equiv \pi_\phi$, whose solutions are

$$\frac{L}{\sqrt{1-\alpha}} = \frac{R^2}{r^2 + a^2\alpha} \left[a\sqrt{1-\alpha} \left(-P_t + \frac{qQr}{R^2} \right) \pm \sqrt{\frac{P_t^2}{\Delta_x} - (r^2 + a^2\alpha)} \right]. \quad (23.87)$$

Numerically, it is better to characterize an orbit by $L/\sqrt{1-\alpha}$ rather than by L itself, since the former remains finite as $\alpha \rightarrow 1$, whereas L and $1-\alpha$ both tend to zero at $\alpha \rightarrow 1$. Substituting the two ($\pm$) expressions (23.87) for L into dP_x^2/dr , and setting the product of the resulting two expressions for dP_x^2/dr equal to zero, equation (23.81), yields a quartic equation

$$\boxed{p_0 + p_1 P + p_2 P^2 + p_3 P^3 + p_4 P^4 = 0}, \quad (23.88)$$

for the dimensionless quantity P (not to be confused with P_t or P_x) defined by

$$P \equiv -\frac{P_t}{R^2 \Delta_x}. \quad (23.89)$$

The minus sign is introduced so as to make P positive in the region of usual interest, which is the Universe region of the Kerr-Newman geometry, where P_t is negative (see Table 23.1). The sign of P is always opposite to that of P_t , since circular orbits exist only where $\Delta_x \geq 0$, outside horizons. The coefficients p_i of the quartic (23.88) are

$$p_0 \equiv r^2(r^2 + a^2\alpha)^2, \quad (23.90a)$$

$$p_1 \equiv -2qQr(r^2 - a^2\alpha)(r^2 + a^2\alpha), \quad (23.90b)$$

$$p_2 \equiv -2r^2(r^2 + a^2\alpha)(r^2 - 3Mr + 2Q^2 + a^2\alpha + a^2\alpha M/r) + q^2Q^2(r^2 - a^2\alpha)^2, \quad (23.90c)$$

$$p_3 \equiv 2qQr(r^2 - a^2\alpha)(r^2 - 3Mr + 2Q^2 + 2a^2 - a^2\alpha + a^2\alpha M/r), \quad (23.90d)$$

$$p_4 \equiv [r^6 - 6Mr^5 + (9M^2 + 4Q^2 + 2a^2\alpha)r^4 - 4M(3Q^2 + a^2)r^3 + (4Q^4 - 6a^2\alpha M + 4a^2Q^2 + a^4\alpha^2)r^2 + 2a^2\alpha(2Q^2 + 2a^2 - a^2\alpha)Mr + a^4\alpha^2M^2]. \quad (23.90e)$$

The quartic (23.88) is the condition for an orbit at radius r to be circular. Physical solutions P must be real. Barring degenerate cases, the quartic (23.88) has either zero, two, or four real solutions at any one radius r . Numerically, it is better to solve the quartic (23.88) for the reciprocal $1/P$ rather than P , since the vanishing of $1/P$ defines the location of circular orbits of massless particles, §23.13. Roots of the quartic (23.88) as a function of radius are illustrated in Figure 23.3 for a charged particle in Kerr-Newman black hole, with illustrative values of black hole and particle parameters.

The azimuthal angular momentum $L/\sqrt{1-\alpha}$, energy E , and stability $d^2P_x^2/dr^2$ of a circular orbit are, in terms of a solution P of the quartic (23.88),

$$\begin{aligned} \frac{L}{\sqrt{1-\alpha}} &= \frac{1}{2a\sqrt{1-\alpha}} [R^2P^{-1} - qQ(r^2 - a^2)/r - (R^2 - 3Mr + 2Q^2 + a^2M/r)P] \\ &= \pm \frac{1}{r^2 + a^2\alpha} \sqrt{l_{-1}P^{-1} + l_0 + l_1P + l_2P^2}, \end{aligned} \quad (23.91a)$$

$$\begin{aligned} E &= \frac{1}{2} [P^{-1} + qQ/r + (1 - M/r)P] \\ &= \pm \frac{1}{r^2 + a^2\alpha} \sqrt{e_{-1}P^{-1} + e_0 + e_1P + e_2P^2}, \end{aligned} \quad (23.91b)$$

$$\frac{d^2P_x^2}{dr^2} = \frac{2}{(r^2 + a^2\alpha)^2} (q_{-1}P^{-1} + q_0 + q_1P + q_2P^2), \quad (23.91c)$$

where the coefficients l_i , e_i , and q_i are

$$l_{-1} \equiv qQrR^2(r^2 + a^2\alpha), \quad (23.92a)$$

$$l_0 \equiv -R^2(r^2 + a^2\alpha)(2Mr - Q^2) - q^2Q^2(r^4 - a^4\alpha), \quad (23.92b)$$

$$\begin{aligned} l_1 &\equiv -\frac{qQ}{r} [2r^6 - 5Mr^5 + 3(Q^2 + a^2)r^4 - a^2(1+\alpha)Mr^3 + a^2(Q^2 + \alpha Q^2 + a^2 - a^2\alpha)r^2 \\ &\quad + 3a^4\alpha Mr - a^4\alpha(Q^2 + a^2)], \end{aligned} \quad (23.92c)$$

$$l_2 \equiv [3Mr^3 - 2Q^2r^2 + a^2(1+\alpha)Mr - a^2(1+\alpha)Q^2 - a^4\alpha M/r] R^4 \Delta_x, \quad (23.92d)$$

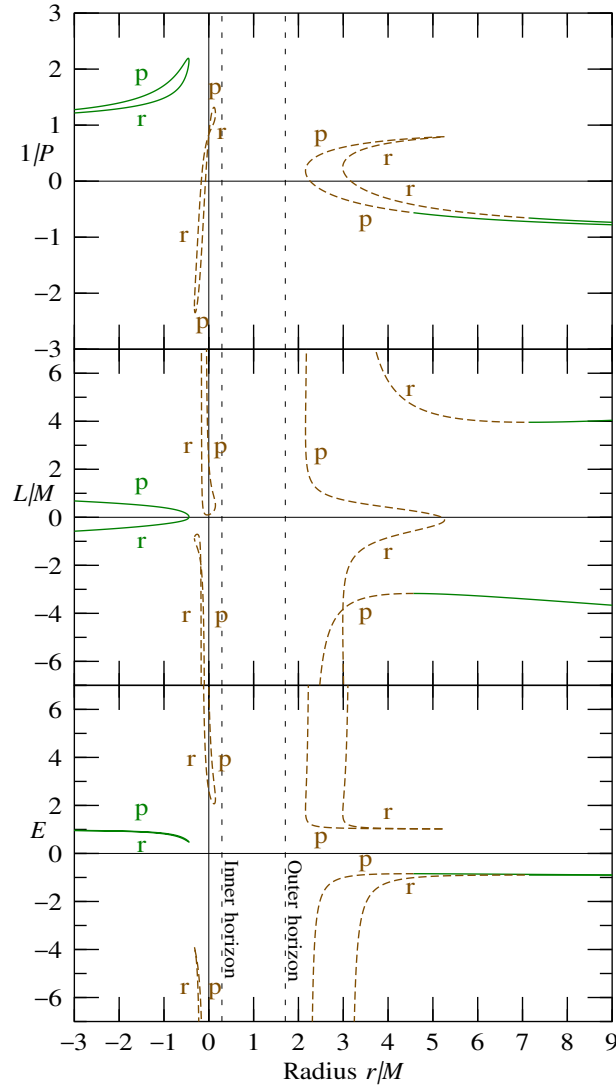


Figure 23.3 Values of $1/P$, equation (23.89), angular momentum L , and energy E , for circular orbits at radius r of a charged particle about a Kerr-Newman black hole. The parameters are illustrative: the black hole has spin parameter $a/M = 0.5$ and charge $Q/M = 0.5$, and the particle has charge-to-mass $q/m = 2.4$ (so $qQ/(mM) = 1.2$) on an orbit of inclination parameter $\alpha = 0.5$. The values $1/P$ are real roots of the quartic (23.88); generically there are either zero, two, or four real roots at any one radius. Solid (green) lines indicate stable orbits; dashed (brown) lines indicate unstable orbits. Positive $1/P$ orbits occur in Universe, Wormhole, and Antiverse regions; negative $1/P$ orbits occur in their Parallel counterparts; zero $1/P$ orbits are null. The fact that the particle is charged breaks the symmetry between positive and negative $1/P$. If the charge of the particle were flipped, $q/m = -2.4$, then the diagrams would be reflected about the horizontal axes (the signs of $1/P$, E , and L would flip). Orbits are marked p for prograde, r for retrograde. In the Universe ($r > r_+$), a positive charge q is repelled by the positive charge Q of the black hole; with $qQ \geq mM$, as here, the electrical repulsion exceeds the gravitational attraction, and there are no circular orbits at large r . Conversely, a negative charge q is attracted by the positive charge Q of the black hole, and there are circular orbits at large r . In the Antiverse ($r < 0$), the situation is symmetrically equivalent to one in which the radius is positive and the mass and charge are flipped, transformation (23.98); the positive charge q effectively sees a black hole with negative mass $-M$ and negative charge $-Q$, and is therefore attracted by the charged black hole. Thus in the Antiverse, there are circular orbits at large negative r for $qQ \geq mM$, as here, but not for $qQ < mM$.

$$e_{-1} \equiv qQr(r^2 + a^2\alpha) , \quad (23.93a)$$

$$e_0 \equiv (r^2 + a^2\alpha)(r^2 - 2Mr + Q^2 + a^2\alpha) + q^2Q^2a^2\alpha , \quad (23.93b)$$

$$e_1 \equiv \frac{qQ}{r} [Mr^3 - (Q^2 + a^2 - 2a^2\alpha)r^2 - 3a^2\alpha Mr + a^2\alpha(Q^2 + a^2)] , \quad (23.93c)$$

$$e_2 \equiv (Mr - Q^2 - a^2\alpha M/r)R^4\Delta_x , \quad (23.93d)$$

and

$$q_{-1} \equiv 2qQr(r^2 - a^2\alpha)(r^2 + a^2\alpha) , \quad (23.94a)$$

$$q_0 \equiv -4(r^2 + a^2\alpha)(Mr^3 - Q^2r^2 - a^2\alpha Mr) - q^2Q^2(r^2 - a^2\alpha)^2 , \quad (23.94b)$$

$$q_1 \equiv -\frac{qQ}{r} [r^6 - 4Mr^5 + 3(Q^2 + a^2 - 2a^2\alpha)r^4 + 12a^2\alpha Mr^3 - a^2\alpha(6Q^2 + 6a^2 - a^2\alpha)r^2 - a^4\alpha^2(Q^2 + a^2)] , \quad (23.94c)$$

$$q_2 \equiv (3Mr^3 - 4Q^2r^2 - 6a^2\alpha Mr - a^4\alpha^2 M/r) R^4\Delta_x . \quad (23.94d)$$

Equations (23.91) determine the values of L , E , and $d^2P_x^2/dr^2$ uniquely for any given root P of the quartic (23.88). The expressions on the second lines of equations (23.91a) for L and equations (23.91b) for E are equivalent to the expressions on the first lines, the sign of the second expressions being chosen to agree with those of the first expressions. For L , the first expression has the virtue of being unambiguous in sign, while the second expression has the virtue of remaining well-behaved in the limit $a \rightarrow 0$ or $1 - \alpha \rightarrow 0$. The two expressions (23.91a) for L are moreover equivalent to the expression (23.87) with one of the two choices of sign in the latter.

For nonzero a , the reality of a solution P of the quartic (23.88) is a necessary and sufficient condition for a corresponding circular orbit to exist. In particular, the argument of the square root in the expression (23.91a) for L is guaranteed to be positive. For zero a , however, the quartic (23.88), which reduces in this case to the square of a quadratic, §23.20, admits real solutions that do not correspond to a circular orbit. For these invalid solutions, the argument of the square root in the second-line expression (23.91a) for L is negative. Thus for zero a , a necessary and sufficient condition for a circular orbit to exist is that the solutions for both P and L be real.

Equation (23.91b) shows immediately that circular orbits of neutral ($q = 0$) particles necessarily have positive energy E in the Universe region outside the horizon, where $P \geq 0$ and $r \geq M$. It is true, but not so obvious, that circular orbits of charged particles ($q \neq 0$) must also have positive energy E in the Universe region outside the horizon. As discussed in §23.17, equation (23.122), the circular orbits with the smallest possible energy are equatorial orbits at the horizon of an extremal uncharged black hole.

Also of interest is the derivative $dP_y^2/d\alpha$ of the angular potential at turnaround, where $P_y = 0$. Orbits at constant latitude occur where $dP_y^2/d\alpha$ vanishes at turnaround. In terms of a solution P of the quartic (23.88), the derivative $dP_y^2/d\alpha$ is

$$\frac{dP_y^2}{d\alpha} = \frac{1}{r^2 + a^2\alpha} (k_{-1}P^{-1} + k_0 + k_1P + k_2P^2) , \quad (23.95)$$

where the coefficients k_i are

$$k_{-1} \equiv -qQr(r^2 + a^2\alpha) , \quad (23.96a)$$

$$k_0 \equiv (r^2 + a^2\alpha)(2Mr - Q^2) + q^2Q^2(r^2 - a^2\alpha) , \quad (23.96b)$$

$$k_1 \equiv \frac{qQ}{r} [(2r^4 - 5Mr^3 + 3(Q^2 + a^2 - 2a^2\alpha)r^2 + a^2\alpha(3Mr - Q^2 - a^2))] , \quad (23.96c)$$

$$k_2 \equiv -(3Mr - 2Q^2 - a^2\alpha M/r) R^4 \Delta_x , \quad (23.96d)$$

23.11.1 Discrete symmetries of the orbital structure

The orbital structure in the Kerr-Newman geometry has two discrete symmetry transformations, parallel and radial flips. The parallel flip, which arises from time reversal symmetry $t \leftrightarrow -t$ of the Kerr-Newman geometry, exchanges Universes, Wormholes, and Antiverses with their Parallel counterparts,

$$P \leftrightarrow -P , \quad Q \leftrightarrow -Q , \quad L \leftrightarrow -L , \quad E \leftrightarrow -E . \quad (23.97)$$

The radial flip exchanges Universes and Antiverses,

$$r \leftrightarrow -r , \quad M \leftrightarrow -M , \quad Q \leftrightarrow -Q . \quad (23.98)$$

23.11.2 Prograde and retrograde orbits

At zero spin, $a = 0$, the quartic (23.88) reduces to the square of a quadratic (this is the Reissner-Nordström case considered in §23.20). Each real root P in this case is doubly degenerate. The two roots have opposite signs of the angular momentum $L/\sqrt{1-\alpha}$. As the spin a is increased away from zero, the two roots for P are rotationally split. The root with the more positive angular momentum L (the direction of the axis of the black hole being taken so that a is positive) is called prograde, while the root with the more negative angular momentum is called retrograde (this is in the Universe, Wormhole, and Antiverse parts of the geometry, where P is positive; in their parallel counterparts, the prograde orbit has more negative L , consistent with the symmetry transformation (23.97); in all, the prograde orbit is the one with the more positive $PaL/\sqrt{1-\alpha}$).

Every transition between prograde and retrograde occurs at a double root P of the quartic; but not every double root has such a transition. For example, in the charged particle case illustrated in Figure 23.3, in the Universe part of the geometry ($P > 0$, $r > r_+$), there are two prograde orbits at the same radius at and just inside the prograde null circular orbit ($1/P \rightarrow +0$); and similarly there are two retrograde orbits at the same radius at and just inside the retrograde null circular orbit.

23.12 Circular geodesics (orbits for particles with zero electric charge)

Geodesics are trajectories for freely falling neutral particles, whose motion is influenced only by gravity. For a particle with zero electric charge, $q = 0$, the odd coefficients p_i vanish in the quartic condition (23.88) for a

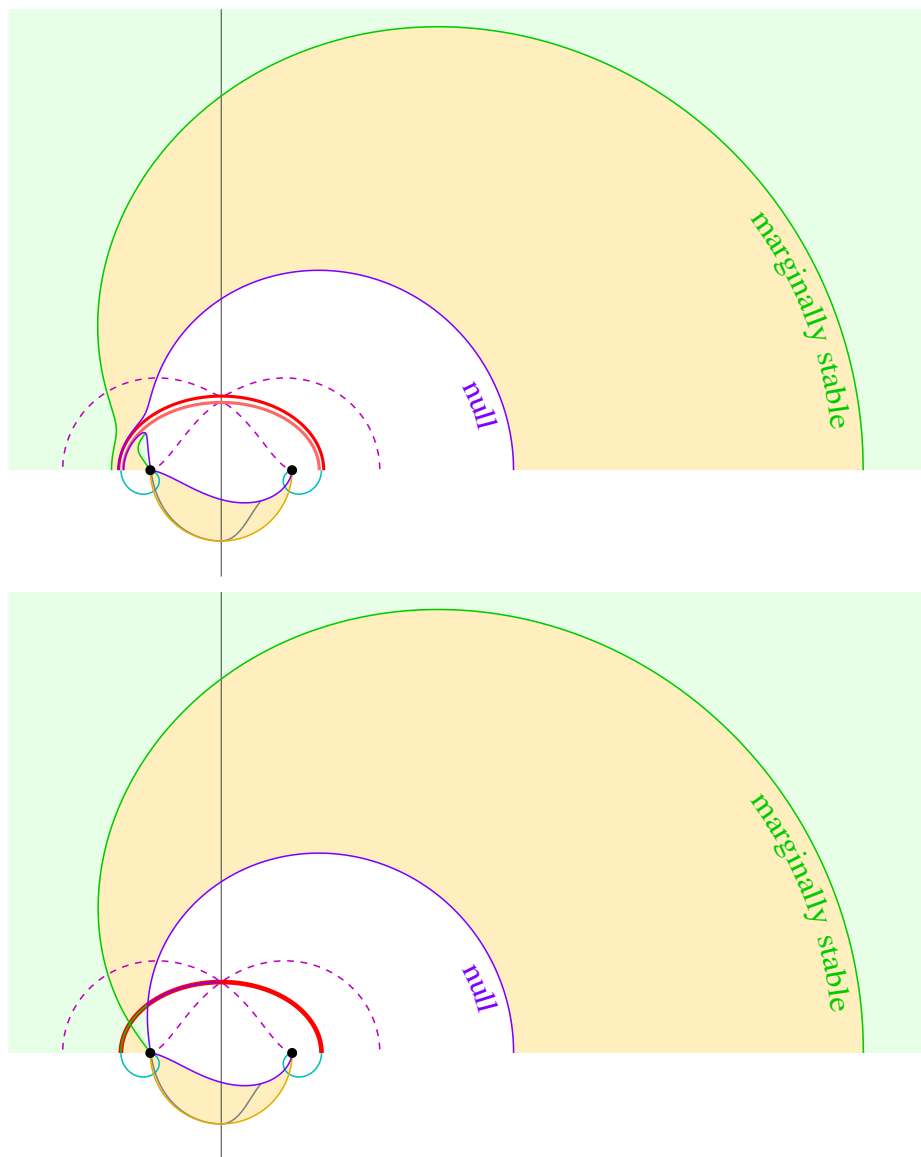


Figure 23.4 Location of stable (shaded green) and unstable (shaded amber) circular orbits in a Kerr black hole with spin (top) slightly sub-extremal ($a = 0.999M$), and (bottom) extremal ($a = M$). The plotted latitude of each circular orbit is its inclination, the maximum latitude reached by the orbit. Null (violet), marginally stable (green), and constant-latitude (grey; inside the Antiverse, at $r < 0$) circular orbits are marked. Regions where circular orbits exist are bounded by the two conditions (23.101) (brown and violet violet respectively). Prograde orbits are drawn to the left of the vertical axis, retrograde orbits to the right. Outer and inner ergospheres (dashed, purple), outer and inner horizons (red), sisy tubes (cyan), and singularities (black) are shown as in Figures 9.1 and 9.3.

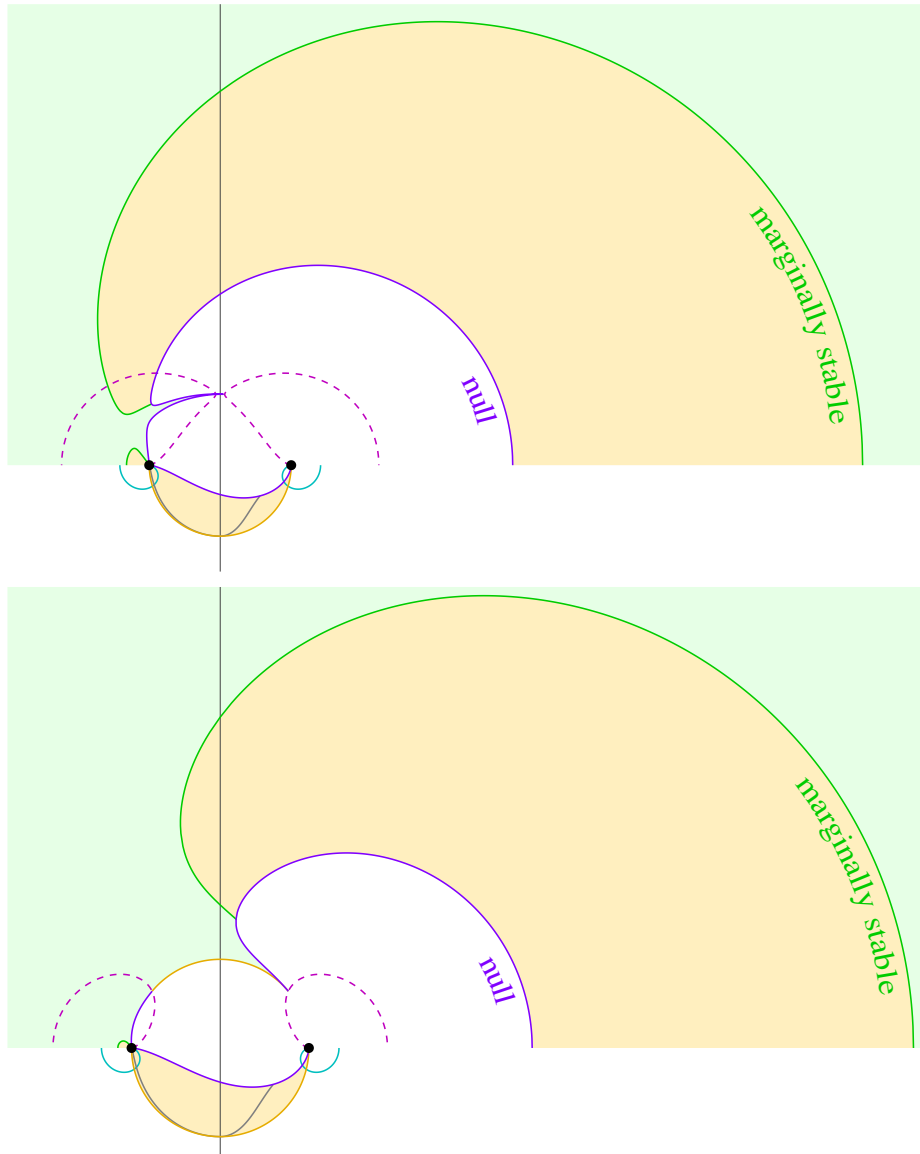


Figure 23.5 As Figure 23.4, but for a Kerr black hole with spin (top) slightly super-extremal ($a = 1.001M$), and (bottom) super-extremal ($a = 1.25M$). Ergospheres, sisy tubes, and singularities are shown as in Figure 9.4.

circular orbit vanish, and the quartic reduces to a quadratic in P^2 . Solving the quadratic yields two possible

solutions

$$1/P^2 = \frac{F_{\pm}}{r^2 + a^2\alpha} , \quad (23.99)$$

where $F_{\pm}$ are

$$F_{\pm} \equiv r^2 - 3Mr + 2Q^2 + a^2\alpha(1 + M/r) \pm 2a\sqrt{(1-\alpha)(Mr - Q^2 - a^2\alpha M/r)} . \quad (23.100)$$

with $+$ and $-$ defining respectively prograde and retrograde orbits. By flipping the direction of the rotation axis, the spin parameter a can always be chosen to be positive, $a \geq 0$. For nonzero spin $a \neq 0$, the necessary and sufficient condition for the existence of a circular orbit is that P be real, which requires that $F_{\pm}$ be real and positive, that is,

$$Mr - Q^2 - a^2\alpha M/r \geq 0 \quad \text{and} \quad F_{\pm} \geq 0 . \quad (23.101)$$

The conditions (23.101) remain necessary and sufficient in the limit $a = 0$ of zero spin (where P is real even without the first of the two conditions (23.101)). For zero electric charge q , the expressions (23.91) for the angular momentum L , energy E , and stability $d^2P_x^2/dr^2$ of a circular orbit, and the expression (23.95) for the angular derivative $dP_y^2/d\alpha$ of the angular potential, simplify to

$$\begin{aligned} \frac{L}{\sqrt{1-\alpha}} &= \frac{1}{2a\sqrt{1-\alpha}} [R^2P^{-1} - (R^2 - 3Mr + 2Q^2 + a^2M/r)P] \\ &= \pm \frac{1}{r^2 + a^2\alpha} \sqrt{l_0 + l_2P^2} , \end{aligned} \quad (23.102a)$$

$$\begin{aligned} E &= \frac{1}{2} [P^{-1} + (1 - M/r)P] \\ &= \pm \frac{1}{r^2 + a^2\alpha} \sqrt{e_0 + e_2P^2} , \end{aligned} \quad (23.102b)$$

$$\frac{d^2P_x^2}{dr^2} = \frac{2}{(r^2 + a^2\alpha)^2} (q_0 + q_2P^2) , \quad (23.102c)$$

$$\frac{dP_y^2}{d\alpha} = \frac{1}{r^2 + a^2\alpha} (k_0 + k_2P^2) . \quad (23.102d)$$

The coefficients l_i , e_i , and q_i from equations (23.92), (23.93), and (23.94) reduce to

$$l_0 \equiv -R^2(r^2 + a^2\alpha)(2Mr - Q^2) , \quad (23.103a)$$

$$l_2 \equiv [3Mr^3 - 2Q^2r^2 + a^2(1+\alpha)Mr - a^2(1+\alpha)Q^2 - a^4\alpha M/r] R^4\Delta_x , \quad (23.103b)$$

$$e_0 \equiv (r^2 + a^2\alpha)(r^2 - 2Mr + Q^2 + a^2\alpha) , \quad (23.104a)$$

$$e_2 \equiv (Mr - Q^2 - a^2\alpha M/r)R^4\Delta_x , \quad (23.104b)$$

and

$$q_0 \equiv -4(r^2 + a^2\alpha)(Mr^3 - Q^2r^2 - a^2\alpha Mr) , \quad (23.105a)$$

$$q_2 \equiv (3Mr^3 - 4Q^2r^2 - 6a^2\alpha Mr - a^4\alpha^2 M/r) R^4\Delta_x , \quad (23.105b)$$

while the coefficients k_i from equations (23.96) reduce to

$$k_0 \equiv (r^2 + a^2\alpha)(2Mr - Q^2) , \quad (23.106a)$$

$$k_2 \equiv -(3Mr - 2Q^2 - a^2\alpha M/r)R^4\Delta_x . \quad (23.106b)$$

Figures 23.4 and 23.5 illustrate the location of stable and unstable circular orbits in the Kerr geometry ($Q = 0$) with sub-extremal and extremal spins (Figure 23.4), and super-extremal spin (Figure 23.5). The four spins shown, $a/M = 0.999, 1, 1.001$, and 1.25 , are chosen to bring out how the orbital structure changes from sub- to super-extremal.

The locations of circular orbits are bounded by the two conditions (23.101). The boundaries corresponding to the two conditions (23.101) are marked respectively by solid amber and violet lines in Figures 23.4 and 23.5. As discussed further in §23.13, the boundary of the second of the two conditions (23.101), $F_{\pm} = 0$, corresponds to null circular orbits.

All circular orbits at $r > 0$ have positive Carter integral, $\mathcal{Q} \geq 0$ (with $\mathcal{Q} = 0$ for equatorial orbits), and therefore pass through the equator according to condition (23.85). Conversely, all circular orbits at $r < 0$ have strictly negative Carter integral $\mathcal{Q} < 0$, and therefore do not pass through the equator: they have both a maximum and minimum latitude.

23.13 Null circular orbits

Null circular orbits define the **photon sphere**, marked by solid violet lines in Figures 23.4 and 23.5. Circular orbits for massless particles, $m = 0$, or null circular orbits, follow from the solutions for massive particles in the case where the energy and angular momentum on the circular orbit become infinite, which occurs when $P_t \rightarrow \pm\infty$. Except at horizons, where $\Delta_x = 0$, this occurs when a solution $P \equiv -P_t/(R^2\Delta_x)$ of the quartic (23.88) diverges, which happens when the ratio p_4/p_0 of the highest to lowest order coefficients vanishes. The ratio p_4/p_0 , equations (23.90), factors as

$$\frac{p_4}{p_0} = \frac{F_+ F_-}{(r^2 + a^2\alpha)^2} , \quad (23.107)$$

where $F_{\pm}$ are defined by equation (23.100). A null circular orbit thus occurs at a radius r such that

$$F_+ = 0 \quad \text{or} \quad F_- = 0 , \quad (23.108)$$

with $+$ for prograde ($aL > 0$) orbits, $-$ for retrograde ($aL < 0$) orbits. The location of null circular orbits are independent of the charge q of the particle, since $F_{\pm}$ are independent of charge q .

The condition (23.108) for a photon sphere is a quadratic equation for the inclination parameter α , yielding

$$\sqrt{1 - \alpha_p} = \frac{1}{a(r_p - M)} \left[\pm r_p \sqrt{R_p^2 - 2Mr_p + Q^2} \pm \sqrt{2Mr_p^3 - (3M^2 + Q^2)r_p^2 + 2MQ^2r_p + a^2M^2} \right] . \quad (23.109)$$

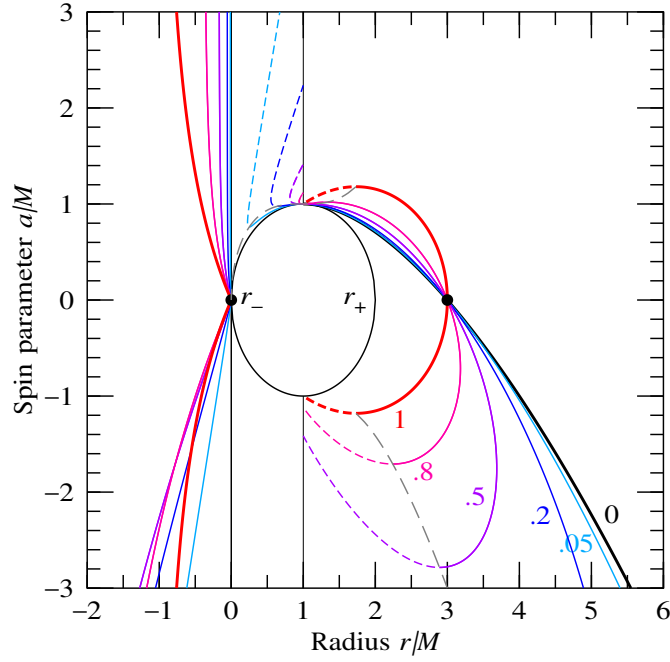


Figure 23.6 Radii of null circular orbits (generalization of the photon sphere) for a Kerr black hole with various spin parameters a , including super-extremal spin parameters, $|a/M| > 1$. Positive and negative a/M signify prograde ($F_+ = 0$) and retrograde ($F_- = 0$) orbits respectively. Lines are labelled with values of the inclination parameter α , varying from equatorial orbits ($\alpha = 0$) to polar orbits ($\alpha = 1$). Solid lines indicate unstable orbits; dashed lines indicate stable orbits; long dashed lines mark the transition between unstable and stable orbits. The radii r_- and r_+ of the inner and outer horizons are shown for reference.

The photon sphere radius r_p ranges over values such that $\alpha_p \in [0, 1]$. The azimuthal angular momentum $J_p \equiv L/E$ per unit energy on the photon sphere is, from equation (23.91a) in the limit $P_t \rightarrow \pm\infty$,

$$J_p = \frac{R_p^2 - 3Mr_p + 2Q^2 + a^2M/r_p}{a(M/r_p - 1)} . \quad (23.110)$$

The Carter constant $\mathcal{K}_p \equiv \mathcal{K}/E^2$ on the photon sphere is, equation (23.84),

$$\mathcal{K}_p = \frac{4r_p^2(R_p^2 - 2Mr_p + Q^2)}{(r_p - M)^2} . \quad (23.111)$$

The case of zero spin $a = 0$ is addressed in §23.20. For zero spin, the photon sphere is at $r_p^2 - 3Mr_p + 2Q^2 = 0$, equation (23.143), and the angular momentum and Carter constant reduce to

$$\frac{J_p^2}{1 - \alpha_p} = \mathcal{K}_p = \frac{2r_p^3}{r_p - M} . \quad (23.112)$$

In the limit of an extremal Kerr-Newman black hole, the angular momentum (23.110) and Carter constant (23.111) on the photon sphere reduce to

$$J_p = \frac{-r_p^2 + 2Mr_p + a^2}{a}, \quad \mathcal{K}_p = 4r_p^2. \quad (23.113)$$

In this case of an extremal Kerr-Newman black hole, there is an additional range of Carter constants $\mathcal{K}_p$ for the part of the photon sphere that is on the horizon, $r_p = M$,

$$J_p^2 \leq \mathcal{K}_p \leq 4r_p^2. \quad (23.114)$$

See §23.17 for more on orbits at the horizon of an extremal black hole.

Figure 23.6 illustrates the radii of null circular orbits for a Kerr (uncharged) black hole, for various spin and inclination parameters a and α , including super-extremal ($|a/M| > 1$) spins. At zero spin, $a = 0$, a Schwarzschild black hole, there is just a single null circular orbit, at $r = 3M$. For a spinning black hole with given positive a/M (a negative a/M can be made positive by flipping the direction of the north pole), there are, barring degenerate cases, 2, 4, or 6 distinct null circular orbits at each inclination, of which 2 or 4 are unstable. At any inclination there are always 2 unstable null circular orbits at negative radius, one prograde and one retrograde (in Figure 23.6, prograde and retrograde orbits are plotted with a/M respectively positive and negative). In the usual case of a sub-extremal ($|a/M| < 1$) Kerr black hole, there are also 2 unstable null circular orbits at positive radius, one prograde and one retrograde. If the black hole is sufficiently near extremal, then in addition there are 2 stable null circular orbits at positive radius. If the black hole is sub-extremal ($|a/M| < 1$), then the additional 2 stable null orbits exist at small inclinations, $\alpha < -3 + 2\sqrt{3} \approx 0.464$; the 2 orbits lie between $r = 0$ and the inner horizon $r = r_-$, and are both prograde. If the black hole is super-extremal ($|a/M| > 1$), then the additional 2 stable null orbits exist at large inclinations, $\alpha > -3 + 2\sqrt{3} \approx 0.464$; one orbit is prograde, the other retrograde.

23.14 The silhouette of a black hole

An isolated (non-accreting) black hole should appear as a black disk silhouetted against the starry background. The edge of the black disk is defined by unstable null circular orbits, the photon sphere, discussed in the previous section 23.13. Figure 23.1 illustrates the silhouette of a Kerr black hole for various spin parameters, as seen by a distant observer in the equatorial plane.

23.15 Marginally stable circular orbits

Figure 23.7 illustrates the radii of marginally stable orbits, those satisfying $d^2P_x^2/dr^2 = 0$, for a Kerr (uncharged) black hole for various spin and inclination parameters a and α . Marginally stable circular orbits are marked by solid green lines in Figures 23.4 and 23.5.

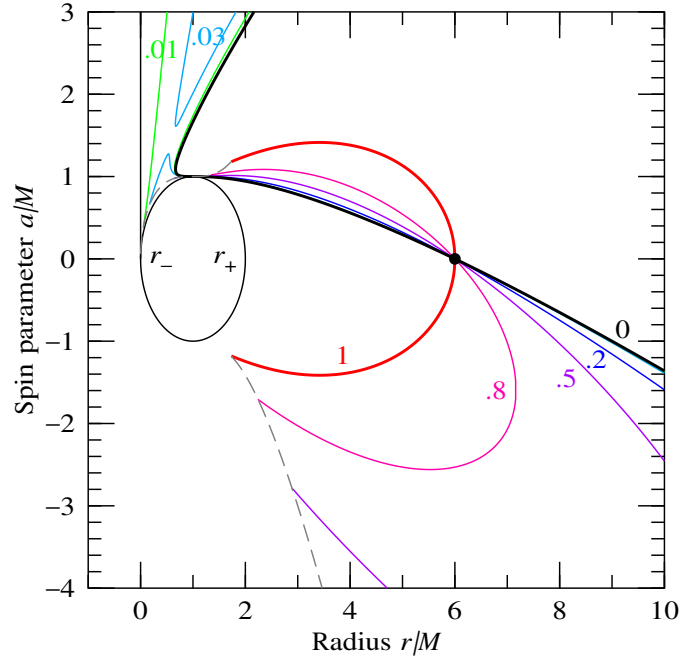


Figure 23.7 Radii of marginally stable circular orbits for a Kerr black hole with various spin parameters a , including super-extremal spin parameters, $|a/M| > 1$. As in Figure 23.6, positive and negative a/M signify prograde and retrograde orbits respectively. Lines are labelled with values of the inclination parameter α , varying from equatorial orbits ($\alpha = 0$) to polar orbits ($\alpha = 1$). Long dashed lines, which are the same as in Figure 23.6, mark where marginally stable orbits become null and terminate, examples of which are illustrated in Figures 23.4 and 23.5. The marginally stable equatorial circular orbit (thick black line) is commonly called the ISCO (innermost stable circular orbit) when the black hole is sub-extremal and the orbit is prograde, $0 \leq a/M \leq 1$. The radii r_- and r_+ of the inner and outer horizons are shown for reference.

23.16 Circular orbits at constant latitude in the Antiverse

In the Antiverse ($r < 0$), there are orbits that are not only circular but also at constant latitude, satisfying $dP_y^2/d\alpha = 0$. These orbits are marked by solid grey lines in Figures 23.4 and 23.5. None of these orbits lies inside the retrograde sisytube, so all of them progress forwards, not backwards, in Boyer-Lindquist time t .

As Figures 23.4 and 23.5 show, there are circular orbits that pass through the retrograde sisytube; but their back-and-forth motion in latitude takes them in and out of the sisytube. These orbits spend a part of their orbit going backwards, and a part going forwards, in Boyer-Lindquist time t .

In the more general situation of charged particles in spinning charged black holes, do there exist any constant-latitude circular orbits that go backwards in Boyer-Lindquist time t ? I have not been able to find any. Do there exist circular orbits of any kind (not necessarily constant latitude) that go backwards in time t ? I have not been able to find any. Nevertheless, if a particle is allowed to accelerate arbitrarily, there

are trajectories inside the retrograde sisytube that go backwards in Boyer-Lindquist time t , and others on which Boyer-Lindquist time t does not change. The latter trajectories, when the azimuthal coordinate ϕ has incremented by -2π , constitute Closed Timelike Curves.

23.17 Circular orbits at the horizon of an extremal black hole

Away from horizons, the vanishing of $F_{\pm}$ defines the location of null circular orbits, §23.13. The case where F_+ vanishes at a horizon (F_- never vanishes at a horizon) is special. This occurs when the black hole is extremal, $M^2 = Q^2 + a^2$. A circular orbit on the horizon, always prograde, is non-null: if $\Delta_x = 0$, as is true on the horizon, then the vanishing of $1/P \equiv -R^2\Delta_x/P_t$ no longer implies that P_t diverges.

A careful analysis shows that the limiting value of $P_t/\sqrt{\Delta_x}$ is finite for a circular orbit at the horizon of an extremal black hole, so in fact $P_t = 0$ for such an orbit. Specifically, let $\tilde{P}$ be the dimensionless quantity

$$\tilde{P} \equiv -\frac{P_t}{M\sqrt{\Delta_x}} . \quad (23.115)$$

For circular orbits on the horizon of an extremal black hole, where $r = M$ and $M^2 = Q^2 + a^2$, the quartic condition (23.88) reduces to a quadratic

$$\tilde{p}_2 + \tilde{p}_3\tilde{P} + \tilde{p}_4\tilde{P}^2 = 0 , \quad (23.116)$$

where the coefficients $\tilde{p}_i$ are

$$\tilde{p}_2 \equiv 4a^2(1-\alpha)(M^2 + a^2\alpha) + (qQ/M)^2(M^2 - a^2\alpha)^2 , \quad (23.117a)$$

$$\tilde{p}_3 \equiv -2(qQ/M)(M^2 - a^2\alpha)(M^2 + a^2\alpha) , \quad (23.117b)$$

$$\tilde{p}_4 \equiv (M^2 + a^2)^2 - a^2(1-\alpha)(6M^2 + a^2 + a^2\alpha) . \quad (23.117c)$$

The azimuthal angular momentum L , energy E , and stability $d^2P_x^2/dr^2$ of circular orbits on the horizon are

$$\frac{L}{\sqrt{1-\alpha}} = \frac{1}{2a} \left[(a^2 - M^2)(qQ/M) + (M^2 + a^2)\tilde{P} \right] = \pm \frac{1}{M^2 + a^2\alpha} \sqrt{\tilde{l}_0 + \tilde{l}_1\tilde{P} + \tilde{l}_2\tilde{P}^2} , \quad (23.118a)$$

$$E = \frac{1}{2} \left(\tilde{P} + qQ/M \right) , \quad (23.118b)$$

$$\frac{d^2P_x^2}{dr^2} = 0 , \quad (23.118c)$$

where the coefficients $\tilde{l}_i$ are

$$\tilde{l}_0 \equiv -(M^2 + a^2)^2(M^2 + a^2\alpha) - q^2Q^2(M^4 - a^4\alpha) , \quad (23.119a)$$

$$\tilde{l}_1 \equiv qQM(M^2 + a^2)(M^2 + a^2\alpha) , \quad (23.119b)$$

$$\tilde{l}_2 \equiv M^2(M^2 + a^2)^2 . \quad (23.119c)$$

Circular orbits on the horizon are always marginally stable, equation (23.118c). Any small perturbation to a marginally stable orbit starts it plunging into the unstable side of the orbit.

Circular orbits on the horizon occur only for small enough inclinations α . For neutral particles, $qQ = 0$, the coefficient $\tilde{p}_3$ vanishes, and $\tilde{p}_2$ is positive, so the quadratic (23.116) has a real root $\tilde{P}$ only as long as $\tilde{p}_4$ is negative. This imposes the condition that

$$\alpha \leq \frac{M(4a^2 - M^2)}{a^2 (3M + 2\sqrt{2M^2 + a^2})} . \quad (23.120)$$

For a Kerr (uncharged) black hole, where $a = M$, the inclination must be less than

$$\alpha \leq -3 + 2\sqrt{3} = 0.464 , \quad (23.121)$$

as illustrated in the bottom panel of Figure 23.4.

The orbital energy E remains finite for a circular orbit at the horizon of an extremal black hole. An interesting case is the circular orbit in the equatorial plane at the horizon of an uncharged extremal ($Q = 0$, $a = M$) black hole, since this orbit has the smallest possible energy per unit mass among all circular orbits in the Universe region (i.e. outside or at the outer horizon) of a Kerr-Newman black hole,

$$E = \frac{qQ}{3M} + \sqrt{\frac{1}{3} + \left(\frac{qQ}{3M}\right)^2} . \quad (23.122)$$

Won't qQ vanish if $Q = 0$? In reality, not necessarily. Real astronomical black holes are almost neutral in part because of the enormous charge-to-mass ratio of a proton, $e/m_p \approx 10^{18}$ in Planck units. (Concept question: Why?) But the same large charge-to-mass ratio means that qQ could be appreciable in spite of the smallness of the black hole charge Q . The smallest possible energy E of a circular orbit occurs as qQ diverges to $-\infty$,

$$E \rightarrow 0 \quad \text{as } qQ \rightarrow -\infty . \quad (23.123)$$

The smallest possible energy for a circular orbit for a neutral particle, $q = 0$, is

$$E = \frac{1}{\sqrt{3}} . \quad (23.124)$$

Of course, there are trajectories with negative energy E in the outer ergosphere, but these trajectories are not circular. The absence of circular orbits with negative energies outside or at the outer horizon implies that all trajectories with negative energy must fall inside the horizon.

Concept question 23.8 Are principal null geodesics circular orbits? Outgoing principal null geodesics hold steady on the outer horizon, remaining at constant $r = r_+$ as time t goes by. Are outgoing principal null geodesics therefore null circular orbits on the horizon? **Answer.** No. The resolution of the conundrum is that whereas no Boyer-Lindquist time t passes on a geodesic at the horizon, proper time does pass. An orbit is circular if it is so for a massive particle; and a circular orbit is null in the limit of a relativistic massive particle. If a massive particle is put on the outer horizon on a relativistic geodesic, then the massive particle necessarily falls off the horizon into the black hole in a finite proper time: it is impossible for the geodesic to hold steady on the horizon. The exception to circular orbits on the horizon

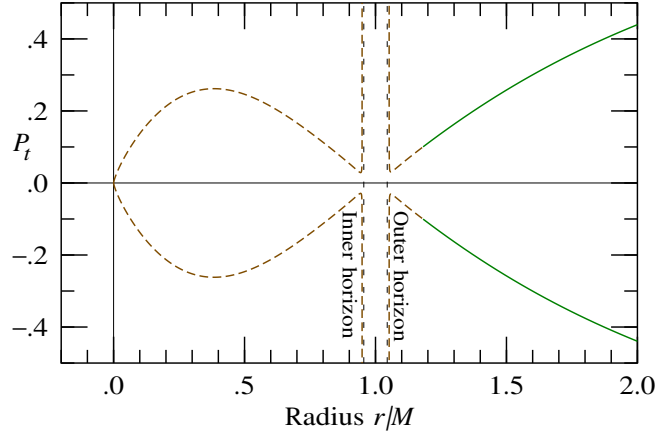


Figure 23.8 Values of the Hamilton-Jacobi parameter P_t for circular orbits at radius r in the equatorial plane of a near-extremal Kerr black hole, with black hole spin parameter $a = 0.999M$. The diagram illustrates that as the orbital radius r approaches the horizon, P_t first approaches zero, but then increases sharply to infinity, corresponding to null circular orbits. In the case of an exactly extremal black hole, P_t goes as to zero at the horizon, there is no increase of P_t to infinity, and no null circular orbit. Solid (green) lines indicate stable orbits; dashed (brown) lines indicate unstable orbits.

is that, as discussed §23.17, an extremal black hole may have circular orbits at its horizon; but these orbits have $P_t = 0$, and are not null.

23.18 Equatorial circular orbits in the Kerr geometry

The case of greatest practical interest to astrophysicists is that of circular orbits in the equatorial plane of an uncharged black hole, the Kerr geometry.

For circular orbits in the equatorial plane, $\alpha = 0$, of an uncharged black hole, $Q = 0$, the solution (23.99) for P simplifies to

$$1/P^2 = \frac{F_{\pm}}{r^2} \quad (23.125)$$

where $F_{\pm}$, equation (23.100), reduce to

$$F_{\pm} \equiv r^2 - 3Mr \pm 2a\sqrt{Mr}, \quad (23.126)$$

with $+$ for prograde ($aL > 0$) orbits, $-$ for retrograde ($aL < 0$) orbits.

As discussed in §23.13, null circular orbits occur where $F_{\pm} = 0$, except in the special case that the circular orbit is at the horizon, which occurs when the black hole is extremal. In the limit where the Kerr black hole is near but not exactly extremal, $a \rightarrow |M|$, null circular orbits occur at $r \rightarrow M$ (prograde) and $r \rightarrow 4M$

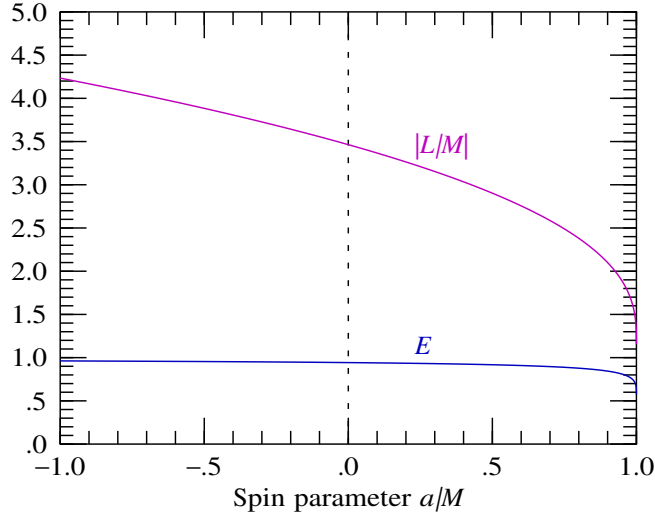


Figure 23.9 Energy E and azimuthal angular momentum $|L/M|$ of circular orbits on the ISCO of a Kerr black hole as a function of the spin parameter a/M . The angular momentum L is positive for $a > 0$ (prograde), negative for $a < 0$ (retrograde).

(retrograde). For an exactly extremal Kerr black hole, $a = |M|$, the (prograde) circular orbit at the horizon is no longer null. The situation of a near-extremal Kerr black hole is illustrated by Figure 23.8.

23.18.1 Innermost stable circular orbit (ISCO)

Astronomers generally argue that the inner edge of an accretion disk is likely to occur at the innermost stable equatorial circular orbit, commonly called the ISCO in the literature. An orbit at this point has marginal stability, $d^2P_x^2/dr^2 = 0$. Simplifying the stability $d^2P_x^2/dr^2$ from equation (23.102c) to the case of equatorial orbits, $\alpha = 0$, and zero black hole charge, $Q = 0$, yields the condition of marginal stability

$$r^2 - 6Mr - 3a^2 \pm 8a\sqrt{Mr} = 0. \quad (23.127)$$

The + (prograde) orbit has the smaller radius, and so defines the innermost stable circular orbit. For an extremal Kerr black hole, $a = |M|$, marginally stable circular equatorial orbits are at $r = M$ (prograde) and $r = 9M$ (retrograde).

The energy E and angular momentum L of a particle on a marginally stable circular equatorial orbit are

$$E = \sqrt{1 - \frac{2M}{3r}}, \quad (23.128a)$$

$$L = \pm \frac{2M}{3\sqrt{3}} \sqrt{\frac{12r}{M} - 7 + 4\sqrt{\frac{3r}{M} - 2}}, \quad (23.128b)$$

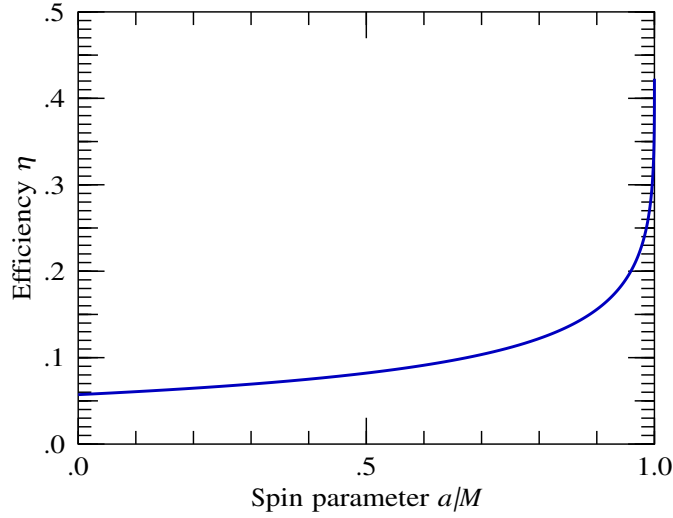


Figure 23.10 Efficiency of accretion on to a Kerr black hole, equation (23.129). The efficiency varies from $\eta = 0.06$ at $a = 0$ to $\eta = 0.42$ at $a = M$.

which are illustrated in Figure 23.9.

23.19 Thin disk accretion

There is a vast observational and theoretical literature on astrophysical accretion flows on to black holes, which is beyond the intended scope of this book (see Abramowicz and Fragile (2013) for a review).

The simplest model of accretion on to a spinning astronomical black hole consists of a thin pressureless disk with particles moving on nearly circular orbits in the equatorial plane (Bardeen, 1970). Viscous forces cause the particles to spiral slowly inward. Observed accretion rates are orders of magnitude larger than can be accounted for by particle viscosity. It is considered likely that the required viscosity arises from turbulence driven by the magneto-rotation instability (Balbus and Hawley, 1998; Balbus, 2003). In the simple model, upon reaching the ISCO (innermost stable circular orbit), particles fall dynamically on to the black hole without further dissipation.

To spiral inward from large radius, where its energy equals its rest mass, $E_\infty = 1$, down to the ISCO, where $E_{\text{ISCO}} = \sqrt{1 - 2M/(3r)}$, equation (23.128a), a particle must lose fractional energy

$$\eta \equiv \frac{E_\infty - E_{\text{ISCO}}}{E_\infty} = 1 - \sqrt{1 - \frac{2M}{3r}}. \quad (23.129)$$

In the simple thin-disk model, particles in the disk lose energy by emitting radiation, which astronomers can detect. The fractional energy η represents the efficiency with which rest mass energy is converted to

radiation. The efficiency η , illustrated in Figure 23.10, varies from $\eta = 1 - \sqrt{8/9} = 0.06$ for a non-spinning black hole ($a = 0$) to $\eta = 1 - \sqrt{1/3} = 0.42$ for a maximally spinning black hole ($a = M$). By comparison, nuclear fusion of hydrogen to helium-4 releases 0.007 of the rest mass, while fusion of hydrogen all the way to iron-56, the most tightly bound of all nuclei, releases 0.009 of the rest mass. Thus gravitational accretion on to a black hole releases energy more efficiently than fusion, by a factor of 10 or more. This explains why gravitational accretion on to black holes can power some of the most luminous objects observed in the Universe, such as quasars and gamma-ray bursts.

23.19.1 Thorne limit

Accretion from the ISCO increases the angular momentum a of the black hole by the angular-momentum to energy ratio L/E of particles on the ISCO. As seen in Figure 23.9, on the ISCO the angular momentum L is always greater than M , and the energy E is always less than 1, so the angular momentum L/E per unit energy on the ISCO always exceeds M . Therefore accretion from the ISCO tends to spin up a sub-extremal black hole towards extremality (Bardeen, 1970). As Thorne (1974) points out, this is problematic because an extremal black hole has zero Hawking temperature. Cooling a thermodynamic object to zero temperature should be difficult if not impossible.

For particles to reach the ISCO from far away, they must lose energy. Thorne (1974) remarked that if the lost energy is emitted as radiation from a thin equatorial disk, then some of that radiation will be absorbed by the black hole, and that radiation will tend to spin down the black hole. Thorne calculated that the maximum spin that a black hole accreting from a thin, radiating disk could achieve is $a = 0.998M$, the precise number depending slightly on the directionality of the radiation emitted from the disk (Thorne considered isotropic radiation, and electron-scattering dipole radiation).

Most of the processes that one can think of serve to reduce the angular momentum even further below extremality. For example, the gas that accretes on to a supermassive black hole may originate from various directions and therefore carry various amounts of azimuthal angular momentum. Although not a rigorous limit, the limit of $a = 0.998$ is often taken by astronomers as a plausible upper bound to the spin of an astronomical black hole, the **Thorne limit**.

Exercise 23.9 Icarus. In Brian Greene's story "Icarus," the boy Icarus goes on a space journey, arrives at a black hole, and goes into orbit around it. When he leaves the black hole, he finds that a large time has passed in the outside world. Is the story realistic?

Solution. Equation (23.4) with $m = 1$ and $q = 0$ implies that the rate $dt/d\tau$ at which time t elapses at infinity relative to the proper time τ experienced by Icarus is

$$\frac{dt}{d\tau} = \frac{1}{\rho^2} \left(-\frac{P_t}{\Delta_x} + \frac{\omega_y P_\phi}{\Delta_y} \right) = \frac{1}{r^2 + a^2 \cos^2 \theta} [R^2 P + a(L - aE \sin^2 \theta)] . \quad (23.130)$$

The first term on the right hand side can become large, with large P , for a circular orbit near the horizon of a near-extremal black hole. For such an orbit, the first term in equation (23.130) dominates. For large P ,

equation (23.102b) shows that

$$P \approx \frac{2E}{1 - M/r} . \quad (23.131)$$

A natural strategy is for Icarus to sail in on to the unstable circular orbit with $E = 1$, since he can manoeuvre into this orbit, and then out of it, without using much rocket energy. For $E = 1$,

$$P \approx \frac{2}{1 - M/r} . \quad (23.132)$$

Equation (23.132) shows that the closer Icarus can get to $r = M$, the more rapidly time passes in the outside world. Any black hole will not do. Icarus must find himself a rotating black hole that is very close to extremal. For a circular orbit in the equatorial plane ($\alpha = 0$, $\theta = \pi/2$) of a Kerr black hole ($Q = 0$), the time dilation factor (23.130) simplifies to

$$\frac{dt}{d\tau} = \frac{r \pm a\sqrt{M/r}}{\sqrt{F_{\pm}}} . \quad (23.133)$$

For $E = 1$, equation (23.133) becomes

$$\frac{dt}{d\tau} = 1 + \frac{2}{\sqrt{1 - a/M}(1 + \sqrt{1 - a/M})} \approx \frac{2}{\sqrt{1 - a/M}} . \quad (23.134)$$

At the Thorne limit $a = 0.998M$, the time dilation factor is

$$\frac{dt}{d\tau} = 44 . \quad (23.135)$$

Exercise 23.10 Interstellar. In the Hollywood movie “Interstellar,” for which Kip Thorne was an Executive Producer, the intrepid band of astronauts lands their spacecraft on planet Miller in orbit around the black hole Gargantua. For each hour the team spends on planet Miller, seven years pass on the outside. That’s a time dilation factor of 60,000. Is it plausible?

Solution. The situation differs from that in the “Icarus” story in that whereas Icarus can manoeuvre his rocket into an unstable circular orbit, a planet must be in a stable orbit. The largest time dilation occurs on the prograde innermost stable circular orbit in the equatorial plane. For a Kerr black hole ($Q = 0$), the time dilation factor (23.130) on the prograde equatorial ISCO is, to lowest order in $1 - a/M$,

$$\frac{dt}{d\tau} \approx \frac{2^{4/3}}{\sqrt{3}(1 - a/M)^{1/3}} . \quad (23.136)$$

To achieve the required time dilation factor requires, to lowest order,

$$1 - a/M \approx \frac{16}{3\sqrt{3}(dt/d\tau)^3} , \quad (23.137)$$

which for $dt/d\tau \approx 60,000$ is

$$1 - a/M \approx 10^{-14} , \quad (23.138)$$

or $a \approx 0.9999999999999999M$. This is much closer to extremality than the Thorne limit. At the Thorne limit $a = 0.998M$, the time dilation factor is

$$\frac{dt}{d\tau} = 11 . \quad (23.139)$$

23.20 Circular orbits in the Reissner-Nordström geometry

Circular orbits of particles in the Reissner-Nordström geometry follow from those in the Kerr-Newman geometry in the limit of a nonrotating black hole, $a = 0$. For a nonrotating black hole, an orbit can be taken without loss of generality to circulate right-handedly in the equatorial plane, $\theta = \pi/2$, so that $\alpha = 0$ and the azimuthal angular momentum L equals the positive total angular momentum L_{tot} . The Carter constant equals the total angular momentum squared, $\mathcal{K} = L_{\text{tot}}^2$. For nonequatorial orbits, the relation between azimuthal and total angular momentum is $L = \pm\sqrt{1 - \alpha} L_{\text{tot}}$.

For a nonrotating black hole, $a = 0$, the quartic condition (23.88) for a circular orbit of a particle of rest mass $m = 1$ and electric charge q reduces to the square of a quadratic,

$$r^2 - qQrP - (r^2 - 3Mr + 2Q^2)P^2 = 0 . \quad (23.140)$$

Solving the quadratic (23.140) yields two solutions

$$1/P = \frac{qQ}{2r} \pm \sqrt{1 - \frac{3M}{r} + \frac{2Q^2}{r^2} + \frac{q^2Q^2}{4r^2}} . \quad (23.141)$$

The sign of P , equation (23.89), is positive in the Universe, Wormhole, and Antiverse regions of the Reissner-Nordström geometry in the Penrose diagram of Figure 8.6, negative in their Parallel counterparts. The angular momentum L , energy E , and stability $d^2P_x^2/dr^2$ of a circular orbit are, in terms of a solution (23.141) P of the quadratic,

$$L = \sqrt{P^2 R^4 \Delta_x - r^2} , \quad (23.142a)$$

$$E = \frac{PR^4 \Delta_x}{r^2} + \frac{qQ}{r} , \quad (23.142b)$$

$$\frac{d^2P_x^2}{dr^2} = 2(r^2 - 6Mr + 5Q^2 + q^2Q^2) - 2\left(1 - \frac{6M}{r} + \frac{6Q^2}{r^2}\right)P^2R^2\Delta_x . \quad (23.142c)$$

For massless particles, circular orbits occur where the solution (23.141) for $1/P$ vanishes, which occurs when

$$r^2 - 3Mr + 2Q^2 = 0 , \quad (23.143)$$

independent of the charge q of the particle. The condition (23.143) is consistent with the Kerr-Newman condition for a null circular orbit, the vanishing of $F_{\pm}$ given by equation (23.100). However, for Kerr-Newman, the argument of the square root on the right hand side of equation (23.100) for $F_{\pm}$ must be positive, even

in the limit of infinitesimal a . In the limit of small a , this requires that $Mr - Q^2 \geq 0$. If the charge Q of the Reissner-Nordström black hole lies in the standard range $0 \leq Q^2 \leq M^2$, then one of the solutions of the quadratic (23.143) lies outside the outer horizon, while the other lies between the outer and inner horizons. As one might hope, the additional condition $Mr - Q^2 \geq 0$ eliminates the undesirable solution between the horizons, leaving only the solution outside the horizon, which is

$$r = \frac{3M}{2} \left(1 + \sqrt{1 - \frac{8Q^2}{9}} \right) \quad \text{for } 0 \leq Q^2 \leq M^2 . \quad (23.144)$$

In (unphysical) cases $Q^2 < 0$ or $M^2 < Q^2 \leq (9/8)M^2$, both solutions of equation (23.143) are valid.

23.21 Ray-tracing

This section 23.21 offers guidance on how to do ray-tracing throughout the analytically extended Kerr-Newman geometry. The cosmological constant will be taken to vanish. For vanishing cosmological constant, the Hamilton-Jacobi potentials P_x and P_y are square roots of quartic polynomials in respectively r and y , so that ray-tracing integrals reduce to elliptic integrals. There are fast numerical algorithms to evaluate elliptic integrals.

Ray-tracing involves following null geodesics of light between emitter and observer. A standard trick is to start from the position of the observer at $\{r_{\text{obs}}, \theta_{\text{obs}}, \phi_{\text{obs}}\}$, and to ray-trace backwards to the position of the emitter at $\{r_{\text{em}}, \theta_{\text{em}}, \phi_{\text{em}}\}$.

In the Boyer-Lindquist tetrad frame, the 4-momentum p^m of an object is related to the Hamilton-Jacobi parameters P_m by equation (23.1). It is convenient to denote by k^m the tetrad-frame 4-momentum of a lightray from an emitter to an observer, reserving p^m for the momentum of the emitter or observer. The Hamilton-Jacobi parameters P_m of the lightray are given by equations (23.5) with vanishing charge q . The conserved energy $-\pi_t$ on a null geodesic can be scaled without loss of generality to 1, and the angular momentum π_ϕ per unit energy can be denoted $\pi_\phi = J$, equations (23.6),

$$\pi_t = -1 , \quad \pi_\phi = J . \quad (23.145)$$

The Hamilton-Jacobi parameters P_t and P_ϕ , equations (23.5), on a null geodesic are

$$P_t = -1 + J\omega_x , \quad P_\phi = J - \omega_y . \quad (23.146)$$

For zero rest mass, $m = 0$, the separated Hamilton-Jacobi equation (23.9) simplifies to

$$\frac{P_t^2 - P_x^2}{\Delta_x} = \frac{P_y^2 + P_\phi^2}{\Delta_y} = \mathcal{K} , \quad (23.147)$$

with $\mathcal{K}$ the Carter constant, which the second equality of equations (23.147) implies must be greater than

or equal to 0. The separated Hamilton-Jacobi equations (23.147) imply that

$$P_x = \pm \sqrt{P_t^2 - \mathcal{K} \Delta_x} , \quad (23.148a)$$

$$P_y = \pm \sqrt{-P_\phi^2 + \mathcal{K} \Delta_y} . \quad (23.148b)$$

The angular potential P_y^2 is a quadratic in y^2 . As previously, it is often convenient to replace the Carter constant $\mathcal{K}$, or equivalently the Carter integral $\mathcal{Q}$, equation (23.31), by the inclination parameter $\alpha \equiv y_{\max}^2$, equation (23.83), defined by the maximum latitude $y_{\max}$ reached by the orbit, equation (23.84). The angular potential P_y^2 is

$$P_y^2 = (\alpha - y^2)(a^2 y^2 + J_a^2 - a^2) \quad (23.149)$$

where

$$J_a \equiv \frac{J}{\sqrt{1 - \alpha}} . \quad (23.150)$$

Numerically, it is better to characterize an orbit by J_a rather than by J itself, since J_a remains finite as $\alpha \rightarrow 1$, whereas J and $1 - \alpha$ both tend to zero at $\alpha \rightarrow 1$. If $|J_a| > |a|$, then P_y^2 has 2 roots, and the geodesic oscillates between $\pm y_{\max}$. If $|J_a| = |a|$, then the orbit asymptotes to an unstable orbit at the equator, $y = 0$. If $|J_a| < |a|$, then P_y^2 has a second pair of roots at $\pm y_{\min}$, with

$$y_{\min}^2 = 1 - J_a^2/a^2 , \quad (23.151)$$

and the geodesic oscillates between minimum and maximum latitudes $y_{\min}$ and $y_{\max}$, or between their negatives $-y_{\min}$ and $-y_{\max}$; in these cases the geodesic never passes through the equator at $y = 0$.

23.21.1 Perceived position of an emitter

For an observer not between the horizons, the coordinate t at the position of the observer is timelike, the observed Boyer-Lindquist tetrad-frame momentum k^m of the lightray satisfies $(k^t)^2 = (k_x)^2 + (k_y)^2 + (k_z)^2$, and the perceived unit direction n_m from observer to emitter in the Boyer-Lindquist tetrad frame is

$$\{n_x, n_y, n_\phi\} = \frac{k_m}{k_t} , \quad (23.152)$$

which points in the direction *opposite* to the momentum of the lightray; note there are two cancelling minus signs in equation (23.152), one from n_m pointing in the opposite direction to k_m , the other from $k_t = -k^t$. The covariant time component k_t is negative in the Universe, Wormhole, and Antiverse parts of the geometry, positive in their Parallel counterparts, Table 23.1. Positive n_x points in the direction of increasing radius r , positive n_y points northward, in the direction of higher latitude, and positive n_ϕ points prograde. The spatial tetrad directions $\{\gamma_x, \gamma_y, \gamma_\phi\}$ form a right-handed system.

For an observer between the horizons, the coordinate x at the position of the observer is timelike, the

observed tetrad-frame momentum of the lightray satisfies $(k^x)^2 = (k_t)^2 + (k_y)^2 + (k_z)^2$, and the perceived unit direction n_m from observer to emitter is

$$\{n_x, n_y, n_\phi\} = \frac{k_m}{k_x}, \quad (23.153)$$

where it is convenient to continue to use the subscript x on the radial direction $n_x = k_t/k_x$ to underline the fact that the direction n_m is continuous as the observer crosses horizons, §23.21.4. The direction n_m from observer to emitter again points in the direction *opposite* to the momentum of the lightray. The covariant time component k_x of the momentum is negative in the Black Hole parts of the geometry, positive in the White Hole parts, Table 23.1.

In terms of the perceived unit direction n_m to the emitter in the Boyer-Linquist frame at position of the observer, the Carter constant $\mathcal{K}$ and angular momentum J on the null geodesic respectively outside and between horizons are

$$\mathcal{K} = \frac{(1 - n_x^2)(1 - \omega_x \omega_y)^2}{(\sqrt{\Delta_x} - n_\phi \omega_x \sqrt{\Delta_y})^2} \Big|_{\text{obs}}, \quad J = \frac{\omega_y \sqrt{\Delta_x} - n_\phi \sqrt{\Delta_y}}{\sqrt{\Delta_x} - n_\phi \omega_x \sqrt{\Delta_y}} \Big|_{\text{obs}} \quad \text{outside horizons}, \quad (23.154a)$$

$$\mathcal{K} = \frac{(1 - n_x^2)(1 - \omega_x \omega_y)^2}{(n_x \sqrt{-\Delta_x} - n_\phi \omega_x \sqrt{\Delta_y})^2} \Big|_{\text{obs}}, \quad J = \frac{\omega_y \sqrt{-\Delta_x} - n_\phi \sqrt{\Delta_y}}{n_x \sqrt{-\Delta_x} - n_\phi \omega_x \sqrt{\Delta_y}} \Big|_{\text{obs}} \quad \text{between horizons}, \quad (23.154b)$$

where all quantities are evaluated at the position of the observer.

23.21.2 Perceived blueshift of an emitter

The blueshift is the ratio of the perceived to emitted energies of a lightray.

If a lightray has 4-momentum k^m in the Boyer-Linquist tetrad frame, and an observer or emitter has 4-velocity u^m relative to the Boyer-Lindquist frame, then the perceived energy of the observed or emitted lightray is $-u_m k^m$ relative to the Boyer-Lindquist frame. The 4-momenta k^m of the lightray are given by equation (23.1) at the respective positions of the observer and emitter.

23.21.3 Scenes in 3D

Modern action movies are often made in 3D, which requires a measure of distance to an emitter.

Concept question 1.21 argues that the distance to an emitter perceived by an observer in special relativity (i.e. in Minkowski space) is the affine distance, which is the affine parameter λ scaled to measure proper distance in the locally inertial frame of the observer. Hamilton and Polhemus (2010) argue that the affine distance continues to offer the best measure of perceived distance to an emitter in general relativity. The affine parameter λ is obtained by integrating equation (23.13) along the lightray, in the same way that the coordinates $x^\mu = \{t, x, y, \phi\}$ of the emitter are obtained by integrating equations (23.11) and (23.12) along the lightray. The affine distance is scaled correctly by multiplying the affine parameter λ by the observed energy k^t of the lightray relative to its energy emitted at rest at infinity.

Binocular vision would no longer suffice for a unique perception of distance, because curved spacetime



Figure 23.11 A three-eyed inhabitant of a highly curved spacetime. Drawing by Wildrose Hamilton.

would distort wavefronts of light away from spherical, confusing your binocular perception. The conflicting visual cues might make you feel queasy. But the failure of binocular vision is merely a limitation of beings who have evolved in flat spacetime. Trinocular vision would work fine. The three-eyed ape in Figure 23.11, drawn by my daughter Wildrose, would have no problem leaping from tree to tree in a highly curved spacetime.

23.21.4 Ray-tracing through the horizon

The Boyer-Lindquist tetrad frame used in this Chapter is that derived in §22.3, which is adapted to the separation of the Hamilton-Jacobi equation. Outside the horizons, an observer instantaneously at rest in the Boyer-Lindquist tetrad frame, $p_x = p_y = p_\phi = 0$, has Hamilton-Jacobi parameters $P_x = P_y = P_\phi = 0$ according to equation (23.1), and they move along a trajectory satisfying $dx = dy = d\phi - \omega_x dt = 0$. At the horizon, an observer at rest in the tetrad frame is corotating with the black hole, but is moving at the speed of light through more realistic observers who free-fall through the horizon. The unrealistic observer at rest sees the outside Universe infinitely blueshifted, and aberrated to a point.

An observer of rest mass m at rest in the Boyer-Lindquist tetrad frame outside the horizons has tetrad-frame momentum

$$p_m = \{-m, 0, 0, 0\} \quad (23.155)$$

(the sign $-m$ of p_t is for the Universe, Wormhole, Antiverse parts of the geometry; the sign is flipped in

their Parallel counterparts, per Table 23.1). The time Hamilton-Jacobi parameter is then $P_t = -m\rho\sqrt{\Delta_x}$ from equation (23.1), which vanishes at horizons, where $\Delta_x = 0$. Realistic freely falling observers have finite Hamilton-Jacobi parameters P_t and P_x , so their tetrad-frame momenta p_t and p_x diverge at the horizon, because of the factors of $1/\sqrt{\Delta_x}$ between momenta and Hamilton-Jacobi parameters. The signs of P_t and P_x are given in Table 23.1.

Ray-tracing smoothly to an observer who falls through the horizon requires a tetrad frame different from the Boyer-Lindquist frame. Different tetrad frames are related by a Lorentz transformation. In general a Lorentz boost suffices; the general form of a Lorentz boost between a rest frame and a moving frame is given by equation (14.80).

Suppose that the tetrad frame is boosted radially by velocity β relative to the Boyer-Lindquist frame. If the perceived unit direction to the emitter is n_m in the boosted frame, then the unit direction n_m^{BL} in the Boyer-Lindquist frame is

$$n_x^{\text{BL}} = \frac{n_x + \beta}{1 + n_x\beta}, \quad n_y^{\text{BL}} = \frac{n_y\sqrt{1-\beta^2}}{1 + n_x\beta}, \quad n_\phi^{\text{BL}} = \frac{n_\phi\sqrt{1-\beta^2}}{1 + n_x\beta}. \quad (23.156)$$

The Carter constant $\mathcal{K}$ and azimuthal angular momentum J on the null geodesic then follow from equations (23.154), and are

$$\mathcal{K} = \frac{(1 - n_x^2)(1 - \omega_x\omega_y)^2}{\left((1 + n_x\beta)\sqrt{\frac{\Delta_x}{1-\beta^2}} - n_\phi\omega_x\sqrt{\Delta_y}\right)^2} \bigg|_{\text{obs}}, \quad J = \frac{(1 + n_x\beta)\omega_y\sqrt{\frac{\Delta_x}{1-\beta^2}} - n_\phi\sqrt{\Delta_y}}{(1 + n_x\beta)\sqrt{\frac{\Delta_x}{1-\beta^2}} - n_\phi\omega_x\sqrt{\Delta_y}} \bigg|_{\text{obs}} \quad \text{outside horizons,} \quad (23.157a)$$

$$\mathcal{K} = \frac{(1 - n_x^2)(1 - \omega_x\omega_y)^2}{\left((n_x + \beta)\sqrt{\frac{-\Delta_x}{1-\beta^2}} - n_\phi\omega_x\sqrt{\Delta_y}\right)^2} \bigg|_{\text{obs}}, \quad J = \frac{(n_x + \beta)\omega_y\sqrt{\frac{-\Delta_x}{1-\beta^2}} - n_\phi\sqrt{\Delta_y}}{(n_x + \beta)\sqrt{\frac{-\Delta_x}{1-\beta^2}} - n_\phi\omega_x\sqrt{\Delta_y}} \bigg|_{\text{obs}} \quad \text{between horizons,} \quad (23.157b)$$

where all quantities are evaluated at the position of the observer. Replacing the boost velocity by its reciprocal, $\beta \rightarrow 1/\beta$, transforms the expressions for $\mathcal{K}$ and J outside horizons to those between horizons. The Carter constant and angular momentum are continuous through the horizon at $\Delta_x = 0$ provided that the boost velocity there is the speed of light, $\beta = \pm 1$.

One choice of boost velocity is the Doran choice $\beta = \pm\sqrt{1 - R^2\Delta_x}$, equation (9.35). The Lorentz boost from the Boyer-Lindquist to Doran tetrad frames is given by equation (19.45), and involves an azimuthal boost as well as a radial boost. The Boyer-Lindquist tetrad-frame momentum of a freely falling Doran observer is

$$p_m = \begin{cases} \frac{m}{\rho} \left\{ \mp \frac{1}{\sqrt{\Delta_x}}, \mp \frac{\beta}{\sqrt{\Delta_x}}, \mp a \cos \theta, \mp a \sin \theta \right\} & \text{outside horizons,} \\ \frac{m}{\rho} \left\{ \mp \frac{\beta}{\sqrt{\Delta_x}}, \mp \frac{1}{\sqrt{\Delta_x}}, \mp a \cos \theta, \mp a \sin \theta \right\} & \text{between horizons,} \end{cases} \quad (23.158)$$

where the various signs must accord with Table 23.1.

23.21.5 Ray-tracing to an emitter at ellipsoidal radius r_{em}

Consider the problem of ray-tracing to an emitter at ellipsoidal radius r_{em} , filling out all angular positions at that radius. For example, the outer and inner horizons are at ellipsoidal radii r_+ and r_- , equation (22.15). The distant sky may be treated as situated at some large but finite radius r . The entrance to the Antiverse is at $r = 0$.

From the perspective of an observer, the perceived boundary of an emitter at radius r_{em} is determined by either of two possible conditions. The first possible condition is that the boundary is where the geodesic of the lightray asymptotes to an unstable null circular orbit; for example, this condition defines the silhouette of the black hole, §23.14. The second possible condition is that the geodesic of the lightray turns around at the emitter radius r_{em} , which can happen only if r_{em} lies outside horizons (outside the outer horizon or inside the inner horizon). In general, the perceived boundary of the emitter may consist of multiple disjoint sections coming from either of the two conditions. There is also the possibility that a section of the boundary may consist of two abutting segments, one from each of the two conditions, which happens if the emitter radius r_{em} intersects the range of radii of unstable null circular orbits. The two conditions are considered in turn below.

The conditions for null circular orbits are as follows. For a black hole with horizons, the unstable null circular boundary perceived by an observer ranges between the prograde (+) and retrograde (−) zeros of $F_{\pm}$ given by equation (23.100) with inclination parameter $\alpha = \cos^2 \theta_{\text{obs}}$. The ranges are defined by the positive regime of the Hamilton-Jacobi potential P_y^2 , equation (23.148b), at the null circular values J_p and $\mathcal{K}_p$ of the azimuthal angular momentum and Carter constant given by equations (23.110) and (23.111),

$$P_y^2 = -\frac{1}{a^2(r_p - M)^2}(F_+ F_-)|_{r=r_p, \alpha=\cos^2 \theta_{\text{obs}}} . \quad (23.159)$$

The product $(F_+ F_-)|_{r=r_p, \alpha=\cos^2 \theta_{\text{obs}}}$ is a sextic in r_p , and the boundaries r_p where P_y^2 vanishes are roots of the sextic. In general there are two unstable null circular boundaries, one outside the outer horizon, the other at negative radius, as illustrated in Figure 23.4.

The black hole may support null circular orbits even if it is super-extremal, $M^2 - Q^2 - a^2 < 0$, and therefore has no horizons, as illustrated in Figure 23.5. Like a black hole with horizons, the unstable null circular orbit boundary may supply part of the perceived boundary of an emitter at radius r_{em} . Like a black hole with horizons, the Hamilton-Jacobi potential P_y^2 along the null circular boundary, equation (23.159), must be positive. But for a super-extremal black hole, additional constraints are imposed by the positivity of the Hamilton-Jacobi potential P_x^2 , equation (23.148a), along the null circular boundary. The Hamilton-Jacobi potential P_x^2 along the radius r of the null geodesic at the null circular values J_p and $\mathcal{K}_p$ of the azimuthal angular momentum and Carter constant given by equations (23.110) and (23.111) is

$$P_x^2 = \frac{(r - r_p)^2}{R^4(r_p - M)^2} \left((r_p - M)^2(r + r_p)^2 - 4Mr_p(r_p - r_-)(r_p - r_+) \right) . \quad (23.160)$$

The final factor on the right hand side of equation (23.160) is a quartic in r_p , whose zeros at the radius $r = r_{\text{obs}}$ of the observer define the boundaries of ranges of positive P_x^2 . Both P_y^2 and P_x^2 , equations (23.159)

and (23.160), must be positive along the null circular orbit boundary. The null boundary may comprise multiple distinct segments.

The Hamilton-Jacobi potential P_x^2 given by equation (23.160) vanishes at $r = r_p$, and its first radial derivative also vanishes there, as it should for a circular orbit, equation (23.81). The condition that the null circular orbit be unstable is that the second derivative $d^2 P_x^2 / dr^2$ is positive. The second derivative of P_x^2 on the null circular orbit at $r = r_p$ is

$$\left. \frac{d^2 P_x^2}{dr^2} \right|_{r=r_p} = \frac{8r_p}{(r_p^2 + a^2)^2 (r_p - M)^2} ((r_p - M)^3 + M(M^2 - Q^2 - a^2)) . \quad (23.161)$$

Equation (23.161) has real roots at $r_p = 0$ and at a critical value r_c ,

$$r_c = M + \left(-M(M^2 - Q^2 - a^2) \right)^{1/3} , \quad (23.162)$$

where the real cube root of ± 1 is $(\pm 1)^{1/3} = \pm 1$. The second derivative (23.161) is positive, and the null circular orbit at r_p is therefore unstable, if

$$r_p(r_p - r_c) > 0 . \quad (23.163)$$

For positive mass M (as can be assumed without loss of generality, given the discrete symmetry (23.98)), the critical radius r_c is positive if $Q^2 + a^2$ is positive, the usual case. As noted in §8.17, it is formally possible to consider imaginary charge, $Q^2 < 0$, which is fun but completely unphysical. For the usual case of positive r_c , the boundary of an emitter at r_{em} can include an unstable null circular component only if $r_{\text{em}} \geq r_c$ or $r_{\text{em}} \leq 0$. Not coincidentally, the azimuthal angular momentum J_p along the null circular geodesic, equation (23.110) passes through an extremum at the same real root $r_p = r_c$,

$$\frac{dJ_p}{dr_p} = \frac{2}{a(r_p - M)^2} ((r_p - M)^3 + M(M^2 - Q^2 - a^2)) . \quad (23.164)$$

The second derivative of J_p at the extremum is

$$\left. \frac{d^2 J_p}{dr_p^2} \right|_{r_p=r_c} = -\frac{6}{a} , \quad (23.165)$$

which is negative if the spin a is taken without loss of generality to be positive, so the extremum of J_p is a maximum.

If the black hole has horizons, then the unstable null boundary perceived by an observer, which defines the silhouette of the black hole, is closed. But if the black hole does not have horizons, then the null boundary at positive radius is present only for $r_p \geq r_c$. If the range of positivity of P_y^2 and P_x^2 , equations (23.159) and (23.160), encompasses r_c , then the null boundary terminates at $r_p = r_c$. Since J_p is a maximum at that point, J_p along the null boundary is constrained to be less than the maximum value.

The above constraints on the unstable null circular boundary determine allowed ranges of P_x^2 and P_y^2 , but the perceived directions n_x and n_y depend also on the signs of P_x and P_y . The angular direction n_y runs over both signs. In the usual case, the radial direction n_x is positive or negative as the radius r_{obs} of the observer

is less than or greater than the direction-dependent radius r_p of the unstable null circular orbit. However, in the analytically extended Kerr-Newman geometry, there are unstable null circular orbits at positive radius such that, if a lightray falls inward off the orbit, it can turn around, and return back to the same unstable null circular orbit at the same positive radius (if the black hole has horizons, the sequence is, start in a Universe, fall into the Black Hole, turn around inside the Inner Horizon, and exit through a White Hole to a New Universe). In that case, an observer (in the New Universe) can see an unstable null circular boundary from the previous Universe, and the radial direction n_x is negative regardless of whether r_{obs} is greater or less than r_p .

What an observer *actually* sees depends on the boundary conditions imposed. In the realistic case of a black hole formed from gravitational collapse, an observer outside the black hole cannot see inside its outer horizon, and an observer who falls inside the outer horizon can see emitters only at larger radius, $r_{\text{em}} \geq r_{\text{obs}}$, within their past lightcone. The infaller's journey necessarily comes to an end at the inner horizon, where the mass inflation instability does its thing. However, this section is considering the more general, unrealistic, case of an analytically extended black hole, in which case the past lightcone of an observer offers a richer menagerie of solutions.

The second possible condition on the perceived boundary of an emitter at radius r_{em} is that the geodesic of the lightray turns around at r_{em} . This can happen only if r_{em} lies outside horizons (outside the outer horizon or inside the inner horizon). The geodesic turns around when $P_x = 0$ at r_{em} , which occurs when the Carter constant $\mathcal{K}$ satisfies

$$\mathcal{K} = \left. \frac{P_t^2}{\Delta_x} \right|_{\text{em}} = \frac{(-1 + \omega_{x,\text{em}} J)^2}{\Delta_{x,\text{em}}} . \quad (23.166)$$

For the emitter to be visible to the observer, the radial and angular potentials P_x^2 and P_y^2 must be positive at all intermediate radii along the null geodesic from emitter to observer.

Given the Carter constant (23.166), the radial and angular potentials P_x^2 and P_y^2 , equations (23.148), are quadratics in the azimuthal angular momentum J ,

$$P_x^2 = A_x (J - J_{x,\text{min}})(J - J_{x,\text{max}}) \quad (23.167a)$$

$$P_y^2 = A_y (J - J_{y,\text{min}})(J - J_{y,\text{max}}) , \quad (23.167b)$$

where the coefficients A_x and A_y are

$$A_x = -\frac{\omega_x^2 \omega_{x,\text{em}}^2}{\Delta_{x,\text{em}}} \left(\frac{\Delta_x}{\omega_x^2} - \frac{\Delta_{x,\text{em}}}{\omega_{x,\text{em}}^2} \right) , \quad A_y = -\frac{\omega_{x,\text{em}}^2}{\Delta_{x,\text{em}}} \left(\frac{\Delta_{x,\text{em}}}{\omega_{x,\text{em}}^2} - \Delta_y \right) , \quad (23.168)$$

and the roots $J_{x,\text{min}}$, $J_{x,\text{max}}$, $J_{y,\text{min}}$, and $J_{y,\text{max}}$ are

$$J_{x,\text{min}} = \frac{\sqrt{\Delta_x} - \sqrt{\Delta_{x,\text{em}}}}{\omega_{x,\text{em}} \sqrt{\Delta_x} - \omega_x \sqrt{\Delta_{x,\text{em}}}} , \quad J_{x,\text{max}} = \frac{\sqrt{\Delta_x} + \sqrt{\Delta_{x,\text{em}}}}{\omega_{x,\text{em}} \sqrt{\Delta_x} + \omega_x \sqrt{\Delta_{x,\text{em}}}} , \quad (23.169a)$$

$$J_{y,\text{min}} = \frac{\sqrt{\Delta_y} - \omega_y \sqrt{\Delta_{x,\text{em}}}}{\omega_{x,\text{em}} \sqrt{\Delta_y} - \sqrt{\Delta_{x,\text{em}}}} , \quad J_{y,\text{max}} = \frac{\sqrt{\Delta_y} + \omega_y \sqrt{\Delta_{x,\text{em}}}}{\omega_{x,\text{em}} \sqrt{\Delta_y} + \sqrt{\Delta_{x,\text{em}}}} . \quad (23.169b)$$

If the observer is outside horizons, then all roots are real. If the observer is between horizons, then the roots

$J_{x,\min}$, $J_{x,\max}$ are complex, in which case they place no constraint on the allowed range of the azimuthal angular momentum J .

If the observer is outside horizons, then the roots $J_{x,\min}$ and $J_{x,\max}$ are real, and the potential P_x^2 is positive between the roots $J_{x,\min}$ and $J_{x,\max}$ if the coefficient A_x , equation (23.168), is negative; or else outside the roots $J_{x,\min}$ and $J_{x,\max}$ if A_x is positive. The coefficient A_x is negative provided that $\Delta_x/\omega_x^2 > \Delta_{x,\text{em}}/\omega_{x,\text{em}}^2$, which is true provided that $|r - M| > |r_{\text{em}} - M|$. The ordering of roots is $J_{x,\min} < J_{x,\max}$, except in a narrow regime $0 < r + r_{\text{em}} < 2M$ (the ordering of roots depends on the ordering of Δ_x/ω_x^2 versus $\Delta_{x,\text{em}}/\omega_{x,\text{em}}^2$, and also on the ordering of ω_x versus $\omega_{x,\text{em}}$, thus on the ordering of $|r - M|$ versus $|r_{\text{em}} - M|$ and of $|r|$ versus $|r_{\text{em}}|$). In the limit of zero spin, $a = 0$, the roots of the quadratic expression (23.167a) for P_x^2 go to $\pm\infty$, and J are all allowed or all excluded as $|r - M|$ is $\geq$ or $< |r_{\text{em}} - M|$.

Given that the emitter must be outside horizons for the geodesic to turn around at the emitter, the roots $J_{y,\min}$ and $J_{y,\max}$, equations (23.169b), are always real. The potential P_y^2 is positive between/outside roots $J_{y,\min}$ and $J_{y,\max}$ as the coefficient A_y is negative/positive. The coefficient A_y is negative provided that $\Delta_{x,\text{em}}/\omega_{x,\text{em}}^2 > \Delta_y$. The boundary of the ergosphere is at $\Delta_x/\omega_x^2 = \Delta_y$, equation (9.18), so A_y is negative provided that the emitter lies outside the radial position of the ergosphere at the latitude of the observer. The ordering of roots is $J_{y,\min} < J_{y,\max}$ if A_y is negative, $J_{y,\min} > J_{y,\max}$ if A_y is positive.

The sign of the derivative of P_x^2 at turnaround at r_{em} determines in which radial direction P_x^2 increases from zero, thus whether the emitted lightray is outfalling or infalling in the immediate vicinity of the point of emission,

$$\left. \frac{dP_x^2}{dr} \right|_{r=r_{\text{em}}} = -\frac{2a^2(r-M)}{R^8\Delta_x}(J - \omega_x^{-1})(J - J_p) \Big|_{r=r_{\text{em}}}. \quad (23.170)$$

The derivative vanishes at $J = J_p$, as expected for a null circular orbit at $r_{\text{em}} = r_p$. The derivative also vanishes at the boundary $J = \omega_x^{-1}$ between prograde and retrograde, and at $r_{\text{em}} = M$, which can be a turnaround point if the black hole has no horizons.

If the emitter radius r_{em} intersects the range of null circular orbits, then the perceived boundary of the emitter may contain a segment with two abutting parts, one from the unstable null circular boundary condition, the other from the turnaround condition. The point of abutment occurs where $r_{\text{em}} = r_p$, at which point $J = J_p$. The angular momentum J_p along the null boundary decreases away from the terminal point at $r_p = r_c$, equation (23.164), so $J \leq J_p$ along the unstable null circular boundary. Along the turnaround part of the boundary, the Hamilton-Jacobi potential P_x^2 given by equation (23.170) must increase away from point of abutment at $r_{\text{em}} = r_p$. In the usual case, the radius r increases or decreases away from r_p as the observer's radius r_{obs} is greater or less than r_p , and then equation (23.170) determines whether J increases or decreases along the turnaround boundary away from the point of abutment at J_p . However, there is also the unusual case remarked earlier, where a lightray falls inward off an unstable null circular orbit, turns around at smaller radius (inside the Inner Horizon, if the black hole has horizons), before being observed. In that case the radius r decreases away from r_p .

23.21.6 Ray-tracing to other surfaces

The direction n_m from observer to emitter, equation (23.152) or (23.153) for observers respectively outside or between the horizons, is invertibly related to the Carter constant $\mathcal{K}$ and azimuthal angular momentum J of the observed lightray. The coordinates x^μ of points along the null geodesic can be calculated by integrating equations (23.11) and (23.12) along the geodesic. For the Kerr-Newman geometry, the integrals are elliptic integrals, for which fast algorithms are available. However, the inverse problem of determining the observed direction n_m given an emitter at a particular coordinate x^μ is not so straightforward, requiring a numerical search in the $\mathcal{K}$ - J plane.

Surfaces, as opposed to points, are more tractable, because they cover a region on the sky, and it may be possible to determine the boundaries of the region in the $\mathcal{K}$ - J plane. The previous subsection 23.21.5 covered an example of this, namely an emitting surface of radius r_{em} . Suppose that an emitting surface is determined by the constancy of some function f ,

$$f = \text{constant} . \quad (23.171)$$

For example, the surface could be one of constant radius r_{em} , as in §23.21.5, or of constant latitude, $y_{\text{em}} = \text{constant}$, which in the case of an equatorial surface, $y_{\text{em}} = 0$, would yield the iconic images of a black hole seen in the movie *Interstellar*. Other examples are the ergosphere and sisytube, defined by respectively $g_{tt} = 0$ and $g_{\phi\phi} = 0$, equations (9.18) and (9.23). The perceived boundary of the surface is determined (leaving aside for the moment the issue of unstable null circular boundaries) by the requirement that the null geodesic turns around at the surface, which requires

$$\frac{\partial f}{\partial x^\mu} \frac{dx^\mu}{d\lambda} = 0 . \quad (23.172)$$

If the surface condition f depends on the radial and polar coordinates x and y , but is independent of time and azimuthal angle t and ϕ (i.e. has time and azimuthal symmetry), as are all the surfaces mentioned, then the condition (23.172) reduces to, with dx^μ given by equation (23.4),

$$\frac{\partial f}{\partial x} \frac{dx}{d\lambda} + \frac{\partial f}{\partial y} \frac{dy}{d\lambda} = \frac{1}{\rho^2} \left(-\frac{\partial f}{\partial x} P_x + \frac{\partial f}{\partial y} P_y \right) = 0 . \quad (23.173)$$

Given the possibility of either sign of P_y (and possibly of P_x), the condition (23.173) becomes

$$\left(\frac{\partial f}{\partial x} \right)^2 P_x^2 - \left(\frac{\partial f}{\partial y} \right)^2 P_y^2 = 0 . \quad (23.174)$$

Equation (23.174) is linear in the Carter constant $\mathcal{K}$ and quadratic in the azimuthal angular momentum J , so yields a value for $\mathcal{K}$ that is quadratic in J , analogous to the case (23.166) for constant r_{em} . Inserting this value of $\mathcal{K}$ into equations (23.148) for the Hamilton-Jacobi potentials P_x and P_y yields expressions for P_x^2 and P_y^2 that are quadratic in J , again analogous to the case (23.169) for constant r_{em} . The perceived boundary of the emitter follows from varying J over the regions where P_x^2 and P_y^2 are both positive.

As in the previous subsection 23.21.5, unstable null circular orbit boundaries may supply some or all of the perceived boundary of the surface. If the surface intersects an unstable null circular orbit, then the perceived

boundary of the emitter may contain a segment with two abutting parts, one from the unstable null circular boundary condition, the other from the turnaround condition.

Exercise 23.11 **What would you like to ray-trace?** What do you think would be most interesting to ray-trace in the Kerr-Newman geometry?

23.22 Hypersurface-orthogonal congruences

The Hamilton-Jacobi separated solution makes it possible to construct congruences (§18.1) of timelike or null geodesics in the Λ -Kerr-Newman geometry, or more generally in any stationary, axisymmetric, separable geometry. Of particular interest are hypersurface-orthogonal congruences, which were discussed in the context of singularity theorems in §§18.7 and 18.8.

It should be remarked from the outset that the principal null congruences of the Λ -Kerr-Newman geometry are *not* hypersurface-orthogonal, Exercise 23.12, except in the special case of spherical symmetry.

23.22.1 Hypersurface-orthogonality condition

As discussed in §18.7, a timelike hypersurface-orthogonal congruence is constructed by picking an arbitrary spacelike 3-dimensional hypersurface on which the action is taken to be constant, and projecting geodesics along the direction orthogonal to the hypersurface at each point. The timelike congruence is orthogonal to hypersurfaces of constant action. Similarly, as discussed in §18.8, a null hypersurface-orthogonal congruence is constructed by foliating an initial 3-dimensional hypersurface into 2-dimensional spatial surfaces of constant action, and projecting pairs of outgoing and ingoing null geodesics orthogonally from the 2-surfaces.

The starting point for constructing timelike or null congruences of geodesics in the Λ -Kerr-Newman geometry is the separated expression (22.22) for the action S of a single particle, with generalized momenta π_x and π_y coming from equations (23.5b) and (23.5c),

$$S = \int \left(-E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy \right). \quad (23.175)$$

Equation (23.175) holds for charged as well as uncharged particles, since for Kerr-Newman the components $\mathcal{A}_x$ and $\mathcal{A}_y$ of the electromagnetic potential vanish, equation (23.3). However, for the remainder of this Chapter, the particle will be taken to be uncharged. The Hamilton-Jacobi parameters P_x and P_y , equations (23.10), depend on the particle mass m and on the constants of motion $C_\alpha \equiv \{E, L, \mathcal{K}\}$. The mass m may be either positive or zero. The integrand on the right hand side of equation (23.175) is manifestly integrable, being a sum of 4 terms each depending on only one of each of the 4 coordinates t, ϕ, x, y . The action (23.175), which is that of a single particle with fixed constants of motion C_α , can be extended to a congruence of geodesics as long as the integral is understood to be taken along geodesics. The constants of motion C_α are by definition constant along each geodesic, but may vary (smoothly) from one geodesic to another. The particle mass m

can be scaled to a global constant without loss of generality, positive for a timelike congruence, zero for a null congruence. Since the integral (23.175) is along geodesics, and the constants of motion are constant along geodesics, the action integrates to the separated expression

$$S - S_i = -E(t - t_i) + L(\phi - \phi_i) - \int_{x_i} \frac{P_x}{\Delta_x} dx + \int_{y_i} \frac{P_y}{\Delta_y} dy, \quad (23.176)$$

in which the constants of motion C_α are held constant in the integrals over x and y even when those constants vary across geodesics. The constants $x_i^\mu \equiv \{t_i, x_i, y_i, \phi_i\}$ are the values of the coordinates x^μ on some arbitrarily chosen initial 3-dimensional hypersurface from which the geodesics are projected. For a timelike congruence the value S_i of the action on the initial hypersurface is constant and can be set to zero, $S_i = 0$, but for a null congruence the initial action S_i must vary over the hypersurface.

Derivatives of the action S (23.176) with respect to the constants of motion C_α yield comoving spatial coordinates $X^\alpha \equiv \{X^E, X^K, X^L\}$ defined by

$$X^E - X_i^E \equiv \frac{\partial S}{\partial E} = \int \left(-dt + \frac{P_t dx}{P_x \Delta_x} + \frac{\omega_y P_\phi dy}{P_y \Delta_y} \right) = -(t - t_i) + \int_{x_i} \frac{P_t dx}{P_x \Delta_x} + \int_{y_i} \frac{\omega_y P_\phi dy}{P_y \Delta_y}, \quad (23.177a)$$

$$X^K - X_i^K \equiv \frac{\partial S}{\partial K} = \int \left(\frac{dx}{2P_x} + \frac{dy}{2P_y} \right) = \int_{x_i} \frac{dx}{2P_x} + \int_{y_i} \frac{dy}{2P_y}, \quad (23.177b)$$

$$X^L - X_i^L \equiv \frac{\partial S}{\partial L} = \int \left(d\phi - \frac{\omega_x P_t dx}{P_x \Delta_x} - \frac{P_\phi dy}{P_y \Delta_y} \right) = \phi - \phi_i - \int_{x_i} \frac{\omega_x P_t dx}{P_x \Delta_x} - \int_{y_i} \frac{P_\phi dy}{P_y \Delta_y}, \quad (23.177c)$$

where X_i^α are the (arbitrary) values of the comoving coordinates on the arbitrarily chosen initial 3-dimensional hypersurface. As in the action (23.176), the integrals in the definitions (23.177) are to be understood as being taken along geodesics. And as in the action (23.176), because the constants of motion are constant along geodesics, the coordinates X^α integrate to the separated expressions on the rightmost sides of equations (23.177) with the constants of motion held constant even when those constants vary across geodesics. As is evident from equations (23.11) and (23.12), the comoving coordinates X^α are constant along geodesics,

$$dX^\alpha = 0, \quad (23.178)$$

justifying their designation as comoving coordinates. The total derivative of the action (23.176) is

$$dS = -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy + dS_i + (X^\alpha - X_i^\alpha) dC_\alpha, \quad (23.179)$$

in which the penultimate term dS_i vanishes for a timelike congruence (where S_i is constant), but is nonvanishing for a null congruence, and the last term $(X^\alpha - X_i^\alpha) dC_\alpha$ takes into account the possible variation of the constants of motion C_α across geodesics.

Timelike geodesics are orthogonal to hypersurfaces of constant action if, equation (18.40),

$$p_\mu = \frac{\partial S}{\partial x^\mu}. \quad (23.180)$$

Equation (23.180) is equivalent to the condition that the total derivative (23.179) of the action is

$$dS = -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy . \quad (23.181)$$

Comparing equations (23.179) and (23.181) shows that a timelike congruence is hypersurface-orthogonal if and only if

$$(X^\alpha - X_i^\alpha) dC_\alpha = 0 . \quad (23.182)$$

The comoving coordinates defined by equations (23.177) are constant along geodesics, $X^\alpha = X_i^\alpha$. The hypersurface-orthogonal condition (23.182) is then satisfied regardless of whether the constants of motion C_α vary across geodesics.

The condition (23.181) for hypersurface-orthogonality can continue to be imposed in the massless limit, where the congruence becomes null. However, for a null congruence the condition (23.181) need not be equivalent to the condition (23.180). As discussed in §18.8, in the massless limit the momentum is not only orthogonal but also tangent to the limiting null hypersurface, and equation (23.180) need be imposed only over each 3-dimensional null hypersurface projected from 2-dimensional surfaces of constant action S_i on the initial 3-dimensional hypersurface, not over the entire 4-dimensional spacetime. By definition, the initial action S_i is constant for each null hypersurface, so $dS_i = 0$ over each null hypersurface. Comparing equations (23.179) and (23.181) shows that a null congruence is hypersurface-orthogonal if and only if once again the condition (23.182) holds, the same condition as for a timelike congruence. And once again, since the comoving coordinates are constant along geodesics, $X^\alpha = X_i^\alpha$, the hypersurface-orthogonal condition (23.182) is satisfied regardless of whether the constants of motion C_α vary across geodesics.

The action S is proportional to the proper time τ along a congruence, $S = -m\tau$, equation (18.39). For a timelike congruence, the action S and 3 comoving coordinates X^α can be used, if desired, as the 4 coordinates along the congruence. But for a null congruence the action S does not progress along worldlines, and the action degenerates to a linear combination of the comoving coordinates X^α . Thus for a null congruence S and X^α are not 4 independent coordinates. But the difference between the action and the linear combination of comoving coordinates, divided by m^2 , remains finite in the limit $m \rightarrow 0$ of zero mass, and defines the coordinate X^S ,

$$X^S \equiv -\frac{1}{m^2} [S - S_i - E(X^E - X_i^E) - 2\mathcal{K}(X^\mathcal{K} - X_i^\mathcal{K}) - L(X^L - X_i^L)] = -\int_{x_i} \frac{\rho_x^2 dx}{P_x} + \int_{y_i} \frac{\rho_y^2 dy}{P_y} . \quad (23.183)$$

As in the action (23.176) and comoving coordinates (23.177), the integrals on the rightmost side of equation (23.183) are to be understood as being taken along geodesics. The variation dX^S of the coordinate X^S equals the variation $d\lambda$ of the affine parameter along geodesics, equation (23.13),

$$dX^S \Big|_{X^E, X^\mathcal{K}, X^L} = -\frac{dS}{m^2} \Big|_{X^E, X^\mathcal{K}, X^L} = d\lambda . \quad (23.184)$$

If desired, the coordinate X^S can be used (in place of S) for timelike as well as null congruences.

23.22.2 Stationary and axisymmetric congruences

In principle the constants of motion C_α can be chosen arbitrarily across geodesics. But it is natural to consider congruences that are stationary and axisymmetric, which requires that the constants C_α be independent of time t and azimuthal angle ϕ (but C_α may depend on the radial and latitude coordinates x and y). A stationary and axisymmetric congruence can be constructed by starting on an arbitrary 1-dimensional line in the x - y plane, and projecting geodesics orthogonally from that 1-dimensional line. The initial action S_i on the 1-dimensional line is constant for a timelike congruence, but varies for a null congruence. The congruence is extended to a full congruence in 4 dimensions by translating and rotating it symmetrically in time t and azimuth ϕ .

For a stationary and axisymmetric congruence, the comoving coordinates X^E and X^L are Killing coordinates at fixed x , and y , as follows from

$$dX^E|_{x,y,X^L} = -dt|_{x,y,\phi} \ , \quad dX^L|_{x,y,X^E} = d\phi|_{t,x,y} \ . \quad (23.185)$$

If X^E and X^L are to be preserved as Killing coordinates, then in place of x and y it is possible to choose any other pair of independent coordinates that depend only on x and y . A possible choice is X^K and X^S , equations (23.177b) and (23.183).

23.22.3 Hypersurface-orthogonal line-element

One way to construct a hypersurface-orthogonal line-element is to use coordinates consisting of the action S and its partial derivatives with respect to the constants of motion, the comoving coordinates $X^\alpha \equiv \partial S / \partial C^\alpha$, equations (23.177). The inverse vierbein in terms of these action coordinates $\{S, \partial S / \partial E, \partial S / \partial K, \partial S / \partial L\}$ is

$$e_m{}^\mu = \frac{1}{\rho} \begin{pmatrix} \frac{P_t}{\sqrt{\Delta_x}} & -\frac{1}{\sqrt{\Delta_x}} & 0 & \frac{\omega_x}{\sqrt{\Delta_x}} \\ \frac{P_x}{\sqrt{\Delta_x}} & -\frac{P_t}{P_x \sqrt{\Delta_x}} & -\frac{\sqrt{\Delta_x}}{2P_x} & \frac{\omega_x P_t}{P_x \sqrt{\Delta_x}} \\ \frac{P_y}{\sqrt{\Delta_y}} & \frac{\omega_y P_\phi}{P_y \sqrt{\Delta_y}} & \frac{\sqrt{\Delta_y}}{2P_y} & -\frac{P_\phi}{P_y \sqrt{\Delta_y}} \\ \frac{P_\phi}{\sqrt{\Delta_y}} & -\frac{\omega_y}{\sqrt{\Delta_y}} & 0 & \frac{1}{\sqrt{\Delta_y}} \end{pmatrix} . \quad (23.186)$$

The vierbein is

$$e^m_{\mu} = \frac{1}{m^2 \rho} \begin{pmatrix} \frac{P_t}{\sqrt{\Delta_x}} & \frac{\pi_t P_t}{\sqrt{\Delta_x}} - \frac{m^2 \rho^2 \sqrt{\Delta_x}}{1 - \omega_x \omega_y} & -\frac{2(\mathcal{K} - m^2 \rho_y^2) P_t}{\sqrt{\Delta_x}} & -\frac{\pi_\phi P_t}{\sqrt{\Delta_x}} - \frac{m^2 \rho^2 \omega_y \sqrt{\Delta_x}}{1 - \omega_x \omega_y} \\ -\frac{P_x}{\sqrt{\Delta_x}} & -\frac{\pi_t P_x}{\sqrt{\Delta_x}} & \frac{2(\mathcal{K} - m^2 \rho_y^2) P_x}{\sqrt{\Delta_x}} & \frac{\pi_\phi P_x}{\sqrt{\Delta_x}} \\ -\frac{P_y}{\sqrt{\Delta_y}} & -\frac{\pi_t P_y}{\sqrt{\Delta_y}} & \frac{2(\mathcal{K} + m^2 \rho_x^2) P_y}{\sqrt{\Delta_y}} & \frac{\pi_\phi P_y}{\sqrt{\Delta_y}} \\ -\frac{P_\phi}{\sqrt{\Delta_y}} & -\frac{\pi_t P_\phi}{\sqrt{\Delta_y}} + \frac{m^2 \rho^2 \omega_x \sqrt{\Delta_y}}{1 - \omega_x \omega_y} & \frac{2(\mathcal{K} + m^2 \rho_x^2) P_\phi}{\sqrt{\Delta_y}} & \frac{\pi_\phi P_\phi}{\sqrt{\Delta_y}} + \frac{m^2 \rho^2 \sqrt{\Delta_y}}{1 - \omega_x \omega_y} \end{pmatrix}. \quad (23.187)$$

23.22.4 Hypersurface-orthogonal timelike outgoing and ingoing congruences

Given the symmetries of the Λ -Kerr-Newman geometry, it is possible to choose timelike or null congruences in symmetrically related outgoing (+) and ingoing (−) partners, the actions $S_\pm$ for which provide coordinates for a line-element (23.193) that describes hypersurface-orthogonal outgoing and ingoing congruences. If the congruence lies outside the horizons, and the action (23.176) describes an outgoing congruence, which is true if the Hamilton-Jacobi parameter P_x is positive (in the Universe part of the geometry), then a corresponding ingoing congruence can be defined by flipping the signs of both P_x and P_y .

Define a time coordinate T and a spatial coordinate Z by

$$T \equiv E(t - t_i) - L(\phi - \phi_i), \quad (23.188a)$$

$$Z \equiv -\int_{x_i} \frac{P_x}{\Delta_x} dx + \int_{y_i} \frac{P_y}{\Delta_y} dy, \quad (23.188b)$$

which are constructed so that the actions $S_\pm$ for the outgoing and ingoing congruences are

$$S_\pm = -T \pm Z. \quad (23.189)$$

The initial coordinates x_i^μ are the same for both outgoing and ingoing actions. The spatial coordinate Z increases outwards if P_x is positive (recall that the radial coordinate x increases inwards).

The flip in the signs of P_x and P_y implies that the comoving coordinate $X^\mathcal{K}$ defined by equation (23.177b) differs by a sign flip along outgoing and ingoing geodesics. Consequently the coordinate $X^\mathcal{K}$ is simultaneously constant along both outgoing and ingoing congruences, allowing the condition $X^\mathcal{K} = 0$ to be imposed simultaneously on both outgoing and ingoing congruences. By contrast, as long as x_i^μ are the same for both outgoing and ingoing actions, as required by the definitions (23.188) of the coordinates T and Z , neither X^E nor X^L can be set simultaneously to zero along both outgoing and ingoing congruences. Therefore the hypersurface-orthogonality condition (23.182) can be satisfied simultaneously by both outgoing and ingoing timelike congruences only if E and L are constant across geodesics.

As long as both outgoing and ingoing congruences are hypersurface-orthogonal, which requires that E and L be constant, the outgoing and ingoing actions $S_\pm$, or equivalently T and Z , can be used as coordinates of

a line-element. For hypersurface-orthogonal congruences, the total derivatives of both outgoing and ingoing actions take the form (23.181), and the total derivatives of T and Z are

$$dT \equiv E dt - L d\phi, \quad (23.190a)$$

$$dZ \equiv -\frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy. \quad (23.190b)$$

The other two coordinates in the line-element of the hypersurface-orthogonal congruences can be taken to be ϕ and either $X^\mathcal{K}$ if $\mathcal{K}$ is constant, or $\mathcal{K}$ if $\mathcal{K}$ varies. Specifically,

$$\frac{dx}{2P_x} + \frac{dy}{2P_y} = dX^\mathcal{K} - \frac{\partial X^\mathcal{K}}{\partial \mathcal{K}} d\mathcal{K}, \quad (23.191)$$

where

$$-\frac{\partial X^\mathcal{K}}{\partial \mathcal{K}} = \int_{x_i} \frac{\Delta_x dx}{4P_x^3} + \int_{y_i} \frac{\Delta_y dy}{4P_y^3}. \quad (23.192)$$

The right hand side of equation (23.191) reduces to $dX^\mathcal{K}$ if $\mathcal{K}$ is constant, or to $-(\partial X^\mathcal{K}/\partial \mathcal{K})d\mathcal{K}$ if $\mathcal{K}$ varies. The line-element of the hypersurface-orthogonal timelike congruences in terms of coordinates T, Z, ϕ , and either $X^\mathcal{K}$ or $\mathcal{K}$, is

$$ds^2 = \frac{\rho^2}{P_x^2 \Delta_y + P_y^2 \Delta_x} \left\{ \Delta_x \Delta_y (-C^2 dT^2 + dZ^2) + 4P_x^2 P_y^2 \left\{ \begin{array}{ll} dX^\mathcal{K} & \text{if } \mathcal{K} \text{ constant} \\ -\frac{\partial X^\mathcal{K}}{\partial \mathcal{K}} d\mathcal{K} & \text{if } \mathcal{K} \text{ varies} \end{array} \right\}^2 \right. \\ \left. + \frac{C^2}{E^2(1 - \omega_x \omega_y)^2} [(P_t^2 \Delta_y - P_\phi^2 \Delta_x) d\phi + (\omega_x P_t \Delta_y - P_\phi \Delta_x) dT]^2 \right\}, \quad (23.193)$$

where the coefficient C is

$$C \equiv \left(\frac{P_x^2 \Delta_y + P_y^2 \Delta_x}{P_t^2 \Delta_y - P_\phi^2 \Delta_x} \right)^{1/2} = \left(1 - \frac{m^2 \rho^2 \Delta_x \Delta_y}{P_t^2 \Delta_y - P_\phi^2 \Delta_x} \right)^{1/2} = \left(1 + \frac{m^2 \rho^2 \Delta_x \Delta_y}{P_x^2 \Delta_y + P_y^2 \Delta_x} \right)^{-1/2}, \quad (23.194)$$

which is always positive. For $m \neq 0$, the coefficient C is less than 1 outside the horizon ($\Delta_x > 0$), equal to 1 at the horizon ($\Delta_x = 0$), and greater than 1 inside the horizon ($\Delta_x < 0$).

The line-element (23.193) is in ADM form (17.8). The comoving coordinate $X^\mathcal{K}$, the constant of motion $\mathcal{K}$, and the one-form in brackets on the second line of the line-element (23.193) all vanish along both outgoing and ingoing geodesics. Thus the only part of the line-element (23.193) that varies along geodesics is the part proportional to $-C^2 dT^2 + dZ^2$. The proper times τ along the timelike geodesics of the outgoing and ingoing congruences satisfy $m\tau = T \mp Z$.

23.22.5 Double-null hypersurface-orthogonal congruences

The line-element (23.193) for hypersurface-orthogonal congruences remains well-defined in the limit of zero particle mass, $m = 0$. In the massless limit, the coefficient C , equation (23.194), is unity,

$$C = 1. \quad (23.195)$$

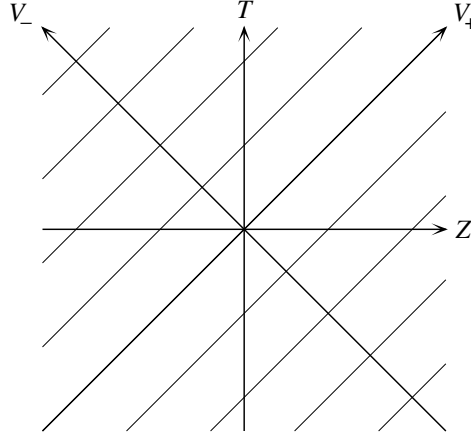


Figure 23.12 Null coordinates $V_{\pm}$, equation (23.197), on a spacetime diagram in T and Z . The diagonal grid of outgoing null lines, which increase in the outgoing null V_+ direction and are lines of constant ingoing null coordinate V_- , are lines of constant phase for an outgoing null wave in the geometric optics (high frequency) limit.

Moreover, for massless particles the energy E can be scaled to ± 1 without loss of generality,

$$|E| = 1. \quad (23.196)$$

Define outgoing (+) and ingoing (−) null coordinates $V_{\pm}$ by

$$V_{\pm} \equiv T \pm Z = -S_{\mp}, \quad (23.197)$$

which equal minus the action along the opposing null geodesic direction. The $V_{\pm}$ null coordinates transform into each other under a flip of the signs of the Hamilton-Jacobi parameters P_x and P_y . If P_x is positive, then V_+ is an outgoing null coordinate that increases along the outgoing null congruence, while V_- is an ingoing null coordinate that increases along the ingoing null congruence. Since the action vanishes along a null geodesic, the outgoing null coordinate is constant along the ingoing congruence, while the ingoing null coordinate is constant along the outgoing congruence, as illustrated in Figure 23.12.

For massless particles, the line-element (23.193) in terms of the coordinates V_+ , V_- , ϕ , and either $X^{\mathcal{K}}$ or $\mathcal{K}$, takes the double-null form

$$ds^2 = \frac{\rho^2}{P_x^2 \Delta_y + P_y^2 \Delta_x} \left\{ -\Delta_x \Delta_y dV_- dV_+ + 4P_x^2 P_y^2 \left\{ \begin{array}{ll} dX^{\mathcal{K}} & \text{if } \mathcal{K} \text{ constant} \\ -\frac{\partial X^{\mathcal{K}}}{\partial \mathcal{K}} d\mathcal{K} & \text{if } \mathcal{K} \text{ varies} \end{array} \right\}^2 \right. \\ \left. + \frac{1}{(1 - \omega_x \omega_y)^2} \left[(P_t^2 \Delta_y - P_{\phi}^2 \Delta_x) d\phi + \frac{1}{2} (P_t \Delta_y - P_{\phi} \Delta_x) (dV_+ + dV_-) \right]^2 \right\}. \quad (23.198)$$

As in the massive case, $dX^{\mathcal{K}}$, $d\mathcal{K}$, and the 1-form in brackets on the second line of the line-element (23.198) all vanish along both outgoing and ingoing null geodesics. The affine parameter $\lambda_{\pm}$ along outgoing (+) or

ingoing (−) geodesics satisfies

$$d\lambda_{\pm} = \frac{\rho^2 \Delta_x \Delta_y}{2(P_x^2 \Delta_y + P_y^2 \Delta_x)} dV_{\pm} . \quad (23.199)$$

Taking the massless limit of the line-element (23.193) does not preserve the condition (23.180) that the momenta along geodesics are orthogonal to hypersurfaces of constant action throughout the 4-dimensional spacetime. Rather, the massless limit of the condition (23.180) imposes the weaker condition that the momenta are orthogonal to hypersurfaces of constant action only within those 3-dimensional hypersurface. This is precisely the definition of hypersurface-orthogonality for null congruences discussed in §18.8.

23.23 The Doran congruence

Congruences in which geodesics follow lines of constant latitude y are of special interest. Constant latitude geodesics must satisfy the two conditions $P_y^2 = dP_y^2/dy = 0$, equations (23.76). These conditions translate into two relations between the three constants E , L , and $\mathcal{K}$ of motion, which may be expressed for example as

$$E = \frac{\partial \sqrt{(\mathcal{K} - m^2 \rho_y^2) \Delta_y} / \partial y}{\partial \omega_y / \partial y} , \quad L = - \frac{\partial \left[\sqrt{(\mathcal{K} - m^2 \rho_y^2) \Delta_y} / \omega_y \right] / \partial y}{\partial (1/\omega_y) / \partial y} , \quad (23.200)$$

the partial derivatives being taken with $\mathcal{K}$ held fixed. For Kerr-Newman without a cosmological constant, the conditions (23.200) imply the relation (23.77) between E and L . Generically, the two conditions allow at most one combination of E , L , and $\mathcal{K}$ to be held constant over spacetime.

However, as discussed in §23.22.4, congruences that are hypersurface-orthogonal simultaneously in both outgoing and ingoing directions can be constructed only if E and L are both constant. For Kerr-Newman without a cosmological constant, the relation (23.77) between E and L for constant latitude geodesics admits just one solution with both E and L constant, the Doran conditions

$$|E| = m , \quad L = 0 , \quad \mathcal{K} = m^2 a^2 . \quad (23.201)$$

For congruences of constant latitude geodesics, where P_y vanishes identically, the comoving coordinates (23.177) can be evaluated by replacing $dy/P_y \rightarrow -dx/P_x$ in the expressions for X^E and X^L ,

$$X^E = -(t - t_i) + \int_{x_i} \left(\frac{P_t}{\Delta_x} - \frac{\omega_y P_\phi}{\Delta_y} \right) \frac{dx}{P_x} , \quad (23.202a)$$

$$X^K = \int_{y_i} \frac{dy}{2P_y} , \quad (23.202b)$$

$$X^L = \phi - \phi_i - \int_{x_i} \left(\frac{\omega_x P_t}{\Delta_x} - \frac{P_\phi}{\Delta_y} \right) \frac{dx}{P_x} . \quad (23.202c)$$

With a suitable choice of boundary conditions, the comoving L coordinate X^L with P_x taken negative (ingoing congruence) coincides with the angular coordinate ϕ_{ff} of the usual Doran metric, equation (9.33),

$X^L = \phi_{\text{ff}}$. The expression (23.202b) for the comoving coordinate X^K appears to diverge, but it appears in the hypersurface-orthogonal line-element (23.193) as $2P_y dX^K \rightarrow dy$, so the end result is well behaved.

For the Doran congruence, the time and spatial coordinates T and Z defined by equation (23.188) are

$$T \equiv mt, \quad (23.203a)$$

$$Z \equiv - \int_{x_i} \frac{P_x}{\Delta_x} dx. \quad (23.203b)$$

The outgoing and ingoing actions $S_{\pm}$ are

$$S_{\pm} = -T \pm Z = -m \left(t \pm \int \frac{\beta dr}{1 - \beta^2} \right), \quad (23.204)$$

where $\beta = P_x/m$ is given by equation (9.35) (with a + sign). As expected, the actions $S_{\pm}$ equal $-m$ times the proper times along the outgoing and ingoing congruences.

The line-element (23.193) of the Doran congruences in hypersurface-orthogonal form is

$$ds^2 = \rho^2 \left[\frac{\Delta_x}{m^2 \beta^2} (-C^2 dT^2 + dZ^2) + \frac{dy^2}{\Delta_y} + \frac{\Delta_y - \omega_y^2 \Delta_x}{(1 - \omega_x \omega_y)^2} \left(d\phi - \frac{\omega_x \Delta_y - \omega_y \Delta_x}{\Delta_y - \omega_y^2 \Delta_x} \frac{dT}{m} \right)^2 \right], \quad (23.205)$$

where the coefficient C is

$$C \equiv \left(\frac{\beta^2 \Delta_y}{\Delta_y - \omega_y^2 \Delta_x} \right)^{1/2} = \left(1 - \frac{\rho^2 \Delta_x \Delta_y}{\Delta_y - \omega_y^2 \Delta_x} \right)^{1/2} = \left(1 + \frac{\rho^2 \Delta_x}{\beta^2} \right)^{-1/2}. \quad (23.206)$$

23.24 Principal null congruences

The principal null congruences are defined by the Carter constant taking its smallest possible value, zero, $\mathcal{K} = 0$, which requires the mass m , and the angular Hamilton-Jacobi parameters P_y and P_{ϕ} , all to vanish identically. For massless particles the energy E can be scaled to ± 1 without loss of generality. The condition $P_{\phi} = 0$ requires that $L = E\omega_y$, so L cannot be constant. The hypersurface-orthogonality condition (23.182) then holds provided that the comoving coordinate X^L is arranged to vanish everywhere. The comoving coordinates X^E and X^L on the principal null congruences are, equations (23.202),

$$X^E = -(t - t_i) \mp \int_{x_i} \frac{dx}{\Delta_x}, \quad X^L = \phi - \phi_i \pm \int_{x_i} \frac{\omega_x dx}{\Delta_x}, \quad (23.207)$$

where the upper sign is for the outgoing congruence, the lower for the ingoing congruence. While X^L can be arranged to vanish on one or other of the outgoing or ingoing null congruences, it cannot be made to vanish simultaneously on both. Thus although the principal null congruences are geodesic, they cannot be described by a line-element that is hypersurface-orthogonal simultaneously on both outgoing and ingoing congruences. According to the theorem proved in §18.1, this implies that there is no vorticity-free tetrad that aligns with the principal null congruences. Exercise 23.12 explores the vorticity ϖ and other components of the extrinsic curvature along the principal null congruences of the Λ -Kerr-Newman geometry.

Exercise 23.12 Expansion, vorticity, and shear along the principal null congruences of the Λ -Kerr-Newman geometry. The separable line-element (22.1) defines a tetrad aligned with the principal null frame, that is, the tetrad-frame Weyl tensor has only a spin 0 part. The outgoing (v) and ingoing (u) null directions lie along the basis elements γ_v and γ_u of the corresponding Newman-Penrose tetrad, equations (41.1).

1. Show that the expansion ϑ , vorticity ϖ , and shear σ along the outgoing (upper sign) and ingoing (lower sign) principal null congruences are

$$\vartheta + i\varpi = s \frac{R^2 \sqrt{|\Delta_x|} (\pm \rho_x + i\rho_y)}{\sqrt{2} \rho^3}, \quad \sigma = 0, \quad (23.208)$$

where the overall sign s is $+$, except that s is $-$ along the outgoing congruence inside the horizon ($\Delta_x < 0$).

2. Define

$$\lambda_{\pm} \equiv (\rho_y \pm i\rho_x) \sqrt{\Delta_y} \frac{\partial \ln \rho^2 / \partial y}{d\omega_y / dy}, \quad \nu \equiv \ln \left(\frac{\rho \sqrt{\Delta_x}}{1 - \omega_x \omega_y} \right), \quad \mu \equiv \text{atan} \left(\frac{\rho_y}{\rho_x} \right). \quad (23.209)$$

Show that the following Lorentz transformation of the tetrad

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda^2 & 1 & \lambda_- & \lambda_+ \\ \lambda_+ & 0 & 1 & 0 \\ \lambda_- & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-\nu} & 0 & 0 & 0 \\ 0 & e^{\nu} & 0 & 0 \\ 0 & 0 & e^{-i\mu} & 0 \\ 0 & 0 & 0 & e^{i\mu} \end{pmatrix} \begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \quad (23.210)$$

brings the tetrad to a form parallel-transported along the outgoing principal null direction γ_v , with vanishing acceleration and precession

$$\Gamma_{kmv} = 0 \quad \text{for all } km, \quad (23.211)$$

and similarly that the Lorentz transformation

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \lambda^2 & -\lambda_+ & -\lambda_- \\ 0 & 1 & 0 & 0 \\ 0 & -\lambda_- & 1 & 0 \\ 0 & -\lambda_+ & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{\nu} & 0 & 0 & 0 \\ 0 & e^{-\nu} & 0 & 0 \\ 0 & 0 & e^{i\mu} & 0 \\ 0 & 0 & 0 & e^{-i\mu} \end{pmatrix} \begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \quad (23.212)$$

brings the tetrad to a form parallel-transported along the ingoing principal null direction γ_u , with vanishing acceleration and precession

$$\Gamma_{kmu} = 0 \quad \text{for all } km. \quad (23.213)$$

The rightmost of the two Lorentz transformations in equations (23.210) and (23.212) boosts and rotates about the radial direction, leaving the directions of all the null tetrad axes $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$ unchanged, while the leftmost of the two Lorentz transformations boost-rotates in such a fashion as to leave just the

outgoing γ_v (respectively ingoing γ_u) axis unchanged, transforming the remaining axes. The Lorentz-transformed frames are no longer principal null. In the outgoing (respectively ingoing) transformed frame, the nonvanishing components of the Weyl tensor are its spin 0, -1 , and -2 (respectively 0, $+1$, and $+2$) components.

23.25 Pretorius-Israel double-null congruence

Generically, congruences cover only part of the spacetime, and geodesics in the congruence cross. The best congruences are those that cover the maximum amount of spacetime, and nowhere cross. The Doran congruence covers all of spacetime down to the inner horizon and beyond, and crosses nowhere, so provides a satisfactory example for massive particles. For massless particles, Pretorius and Israel (1998) pointed out a double-null hypersurface-orthogonal congruence whose geodesics fill all of Kerr spacetime down to the Antiverse ($r = 0$) without crossing.

As found in §23.23, there is no double-null hypersurface-orthogonal congruence with $P_y = 0$. As discussed in §23.22.4, the hypersurface-orthogonality condition (23.182) can be accomplished simultaneously for outgoing and ingoing congruences only if the angular momentum L is constant across geodesics, while the Carter constant $\mathcal{K}$ may vary across geodesics. For massless particles, the energy E can be scaled without loss of generality to ± 1 , with $E = +1$ in the Universe part of the geometry.

In the Λ -Kerr-Newman geometry, motion in latitude extends to the south and north poles only if

$$L = 0 . \quad (23.214)$$

In addition, in order to avoid the geodesics turning around in latitude and therefore crossing, the Carter constant must satisfy

$$\mathcal{K} \geq \frac{\omega_y^2}{\Delta_y} \quad (23.215)$$

at all latitudes y on a geodesic. To fill all of the polar region of spacetime, a geodesic that starts at a pole must remain on the pole, so it must be that $\mathcal{K} = 0$ at the poles (where $\omega_y = 0$). Therefore $\mathcal{K}$ must vary across geodesics in order to satisfy the condition (23.215). Requiring that radial geodesics fall through the outer horizon, and consequently also the inner horizon, places an upper limit on $\mathcal{K}$. The deepest penetration inside the black hole is attained when $\mathcal{K}$ is as small as possible. This leads to the Pretorius and Israel (1998) proposal to set $\mathcal{K}$ to the smallest value consistent with the condition (23.215). This is achieved by choosing $\mathcal{K}$ such that P_y vanishes at infinite radius, which imposes

$$\mathcal{K} = \left. \frac{\omega_y^2}{\Delta_y} \right|_{\infty} . \quad (23.216)$$

For Kerr-Newman without a cosmological constant, this is

$$\mathcal{K} = a^2 \sin^2 \theta_{\infty} , \quad (23.217)$$

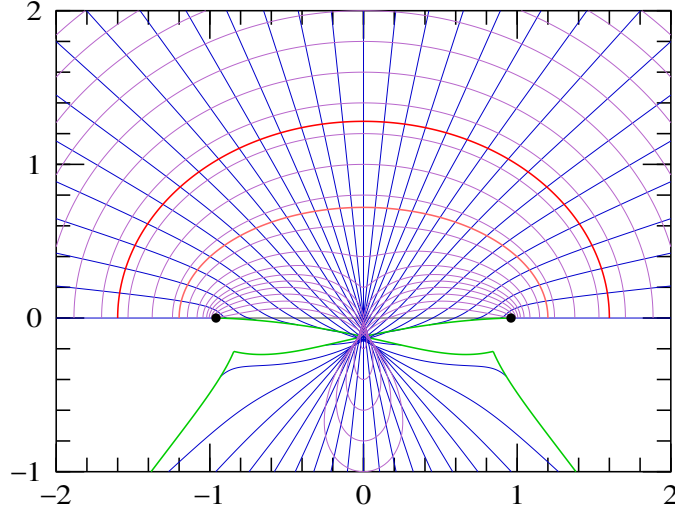


Figure 23.13 Null geodesics (blue lines) and surfaces of constant outgoing and ingoing action (or phase) (purple lines) in the Pretorius and Israel (1998) double-null hypersurface-orthogonal congruence, for a Kerr black hole with $a = 0.96M$. The coordinates are Boyer-Lindquist, and the units are geometric ($c = G = M = 1$). The congruence covers the entire spacetime at $r > 0$ without crossing. The first geodesic crossing occurs on the polar axis at $r = 0$. High latitude geodesics cross when they pass through the pole, turning around in latitude; the continuation of these geodesics through the pole is shown here to illustrate their future progression. Geodesics at low latitude turn around in radius at $r < 0$. The locus of turnaround points is marked by a thick (green) line, where the geodesics shown here are terminated (to avoid cluttering the diagram). Mid-latitude geodesics, after passing through the pole, turn around in latitude for a second time. The locus of turnaround points is marked by a continuation of the thick (green) line, where the geodesics shown here are terminated (again to avoid cluttering the diagram). Thick (reddish) lines mark the outer and inner horizons, and filled circles mark the ring singularity.

where θ_∞ is the polar angle of the geodesic at infinite radius. Null geodesics with $L = 0$ and $\mathcal{K}$ given by equation (23.216) vary in latitude from a minimum latitude that exhausts the condition (23.215) at infinite radius. The Pretorius-Israel double null line-element is equation (23.198) with $L = 0$ and nonconstant $\mathcal{K}$ given by equation (23.216). For $E = 1$ and $L = 0$, the time and azimuth Hamilton-Jacobi parameters are $P_t = -1$ and $P_\phi = -\omega_y$.

Figure 23.13 illustrates the Pretorius-Israel congruence in a Kerr black hole of spin parameter $a = 0.96M$. The outgoing and ingoing congruences lie in, and are orthogonal to, 3-dimensional null hypersurfaces of constant action $S_\pm = -T \pm Z$, where the coordinates T and Z are given by equations (23.188). The initial 3-dimensional hypersurface from which the null hypersurfaces project is a spheroid of constant radius r at infinity for the ingoing congruence, and a spheroid of constant radius r at the outer horizon for the outgoing congruence. As discussed at the end of §23.22.1, the parameters x_i , y_i , and ϕ_i are their values on the initial 3-dimensional hypersurface, and the time parameter t_i is zero. For $E = 1$ and $L = 0$, the time coordinate

T is just the Boyer-Lindquist time coordinate, $T = t$, and the action on the initial hypersurface is $S_i = -t$. The hypersurfaces of constant action, or phase, shown in Figure 23.13 are lines of constant Z at fixed t .

23.25.1 Application of the null singularity theorem to a Λ -Kerr-Newman black hole

Penrose's (1965) original singularity theorem considered hypersurface-orthogonal null congruences. The Pretorius and Israel (1998) double-null congruence provides an example of such a congruence in the Λ -Kerr-Newman geometry.

The surfaces of constant action in Figure 23.13 mark the positions of 2-surfaces from which outgoing and ingoing geodesics project orthogonally. These 2-surfaces are trapped inside the outer horizon, with negative expansion ϑ along both outgoing and ingoing congruences. If the dog-leg proposition (§18.10.1) held, then the future would terminate at the caustic surface of first crossing marked in Figure 23.13 by the upper thick (green) line inside the Antiverse ($r < 0$). However, as discussed in Concept question 18.3, the Kerr-Newman geometry does not satisfy the dog-leg proposition (the same holds if $\Lambda \neq 0$), so the future extends past the caustic crossing.

The failure of the dog-leg proposition in the Λ -Kerr-Newman geometry is associated with the fact that geodesics can emerge without causal precedent from the ring singularity, leading to a breakdown of predictability inside the inner horizon. One should not be surprised that the inner horizon of a real astronomical black hole is subject to an instability, the Poisson and Israel (1990) inflationary instability, that inevitably and profoundly changes the geometry from just above the inner horizon inward.

The interiors of rotating black holes

THIS CHAPTER IS SCARCELY BEGUN

When a black hole first forms by stellar collapse, or when two black holes merge, the resulting object wobbles about, radiating gravitational waves, settling asymptotically to the Kerr geometry, which cannot radiate gravitational waves. After several black hole crossing times, the black hole is already well-approximated by the Kerr geometry.

This picture holds outside the outer horizon, and down to the inner horizon, but it fails dramatically at (just above) the inner horizon of the black hole. The inner horizon is subject to the inflationary instability discovered by Poisson and Israel (1990). Extended to rotating black holes Barrabès, Israel, and Poisson (1990)

There are also spacetimes in which geodesics of massless particles, but not massive particles, are Hamilton-Jacobi separable. Such line-elements are called **conformally separable**.

24.1 Nonlinear evolution

Choose tetrad frame such that the null directions are the geodesic continuations of the outgoing and ingoing principal null geodesics, that the blueshift and the rotation of the outgoing and ingoing principal null geodesics appears the same in the tetrad frame,

$$K_{+vv} = K_{-vv} = K_{+uu} = K_{-uu} = \Gamma_{uvx} = \Gamma_{+-x} = 0 . \quad (24.1)$$

24.2 Focussing along principal null directions

24.3 Conformally separable geometries

24.3.1 Conformally separable line-element

As remarked in §rotbh-chap, the Kerr-Newman line-element has the remarkable mathematical property that the equations of motion of test particles in it, massive or massless, neutral or charged, are Hamilton-Jacobi separable. A weaker condition on the spacetime is that the equations of motion of massless particles are Hamilton-Jacobi separable.

Among the remarkable mathematical properties of is the fact that, as first shown by Carter (1968), the equations of motion of test particles, massive or massless, neutral or charged, are Hamilton-Jacobi separable.

The Kerr geometry is stationary, axisymmetric, and separable.

Choose coordinates $x^\mu \equiv \{t, x, y, \phi\}$ in which t is the time with respect to which the spacetime is stationary, ϕ is the azimuthal angle with respect to which the spacetime is axisymmetric, and x and y are radial and angular coordinates. In §22.3 it is shown that the line-element may be taken to be

$$ds^2 = \rho^2 \left[-\frac{\Delta_x}{(1 - \omega_x \omega_y)^2} (dt - \omega_y d\phi)^2 + \frac{dx^2}{\Delta_x} + \frac{dy^2}{\Delta_y} + \frac{\Delta_y}{(1 - \omega_x \omega_y)^2} (d\phi - \omega_x dt)^2 \right], \quad (24.2)$$

24.4 Conditions from conformal Hamilton-Jacobi separability

24.5 Tetrad-frame connections

Extrinsic curvatures Γ_{azb} along the radial directions $z = t$ and x , and Γ_{yaz} along the angular directions $a = y$ and ϕ . Expansions

$$\Gamma_{202} = \Gamma_{303} = \vartheta_0 = 0, \quad (24.3a)$$

$$\Gamma_{212} = \Gamma_{313} = \vartheta_1 = \partial_1 \ln \rho, \quad (24.3b)$$

$$-\Gamma_{020} = \Gamma_{121} = \vartheta_2 = \partial_2 \ln \rho, \quad (24.3c)$$

$$-\Gamma_{030} = \Gamma_{131} = \vartheta_3 = 0, \quad (24.3d)$$

Twists

$$\Gamma_{320} = -\Gamma_{203} = \Gamma_{302} = \varpi_0 = \frac{\sqrt{\Delta_x}}{2\rho(1 - \omega_x \omega_y)} \frac{d\omega_y}{dy}, \quad (24.4a)$$

$$\Gamma_{321} = -\Gamma_{213} = \Gamma_{312} = \varpi_x = 0, \quad (24.4b)$$

$$\Gamma_{012} = -\Gamma_{120} = \Gamma_{021} = \varpi_2 = 0, \quad (24.4c)$$

$$\Gamma_{tx\phi} = -\Gamma_{x\phi t} = \Gamma_{t\phi x} = \varpi_\phi = \frac{\sqrt{\Delta_y}}{2\rho(1 - \omega_x \omega_y)} \frac{d\omega_x}{dx}. \quad (24.4d)$$

Shear vanishes

$$\Gamma_{202} = \Gamma_{303} = \Gamma_{212} = \Gamma_{313} = 0 , \quad (24.5a)$$

$$\Gamma_{020} = \Gamma_{121} = \Gamma_{030} = \Gamma_{131} = 0 . \quad (24.5b)$$

$$\Gamma_{azz} = 0 , \quad (24.6a)$$

$$\Gamma_{zaa} = 0 , \quad (24.6b)$$

$$\Gamma_{100} = \partial_0 \ln \nu , \quad (24.6c)$$

$$\Gamma_{trr} = \partial_1 \ln \nu , \quad (24.6d)$$

$$\Gamma_{100} = \partial_1 \ln \nu , \quad (24.6e)$$

$$\Gamma_{trr} = \partial_0 \ln \nu , \quad (24.6f)$$

$$\nu \equiv \ln \left(\frac{1 - \omega_x \omega_y}{\rho} \right) , \quad \mu \equiv \ln \left(\frac{1 - \omega_x \omega_y}{\rho} \right) \quad (24.7)$$

24.6 Inevitability of mass inflation

Mass inflation requires the simultaneous presence of both outgoing and ingoing streams near the inner horizon. Will that happen in real black holes? Any real black hole will of course accrete matter from its surroundings, so certainly there will be a stream of one kind or another (outgoing or ingoing) inside the black hole. But is it guaranteed that there will also be a stream of the other kind? The answer is probably.

One of the remarkable features of the mass inflation instability is that, as long as outgoing and ingoing streams are both present, the smaller the perturbation the more violent the instability. That is, if say the outgoing stream is reduced to a tiny trickle compared to the ingoing stream (or vice versa), then the length scale (and time scale) over which mass inflation occurs gets shorter. During mass inflation, as the counter-streaming streams drop through an interval Δr of circumferential radius, the interior mass $M(r)$ increases exponentially with length scale l

$$M(r) \propto e^{\Delta r/l} . \quad (24.8)$$

It turns out that the inflationary length scale l is proportional to the accretion rate

$$l \propto \dot{M} , \quad (24.9)$$

so that smaller accretion rates produce more violent inflation. Physically, the smaller accretion rate, the closer the streams must approach the inner horizon before the pressure of their counter-streaming begins to dominate the gravitational force. The distance between the inner horizon and where mass inflation begins effectively sets the length scale l of inflation.

Given this feature of mass inflation, that the tinier the perturbation the more rapid the growth, it seems

almost inevitable that mass inflation must occur inside real black holes. Even the tiniest piece of stuff going the wrong way is apparently enough to trigger the mass inflation instability.

One way to avoid mass inflation inside a real black hole is to have a large level of dissipation inside the black hole, sufficient to reduce the charge (or spin) to zero near the singularity. In that case the central singularity reverts to being spacelike, like the Schwarzschild singularity. While the electrical conductivity of a realistic plasma is more than adequate to neutralize a charged black hole, angular momentum transport is intrinsically a much weaker process, and it is not clear whether the dissipation of angular momentum might be large enough to eliminate the spin near the singularity of a rotating black hole. There has been no research on the latter subject.

24.7 The black hole collider

A good way to think conceptually about mass inflation is that it acts like a particle accelerator. The counter-streaming pressure accelerates outgoing and ingoing streams through each other at an exponential rate, so that a Lagrangian gas element spends equal amounts of proper time accelerating through equal decades of counter-streaming velocity. The centre of mass energy easily exceeds the Planck energy.

Mass inflation is expected to occur just above the inner horizon of a black hole. In a realistic rotating astronomical black hole, the inner horizon is likely to be at a considerable fraction of the radius of the outer horizon. Thus the black hole accelerator operates not near a central singularity, but rather at a macroscopically huge scale. This machine is truly monstrous.

Undoubtedly much fascinating physics occurs in the black hole collider. The situation is far more extreme than anywhere else in our Universe today. Who knows what Nature does there? To my knowledge, there has been no research on the subject.

Concept question 24.1 Which Einstein equations are redundant? RE-ASK THIS IN CONTEXT OF SPHERICAL MODEL. If 4 of the 10 Einstein equations are redundant (after consistent initial conditions are imposed) because of energy-momentum conservation, can any 4 be dropped, or just the 4 with one component the time component?

Exercise 24.2 Can accretion fuel outgoing and ingoing streams at the inner horizon? The inflationary instability is driven by outgoing and ingoing streams at the inner horizon.

1. What are the conditions for collisionless particles accreting from outside the outer horizon to be outgoing or ingoing at the inner horizon of a Kerr black hole?
2. Of particular relevance in astrophysics are collisionless particles that start at effectively infinite radius r , whether massless (Cosmic Microwave Background photons) or massive (nonbaryonic cold dark matter particles). Calculate the maximum latitude to which particles falling from infinite radius can reach and be either outgoing or ingoing at the inner horizon.

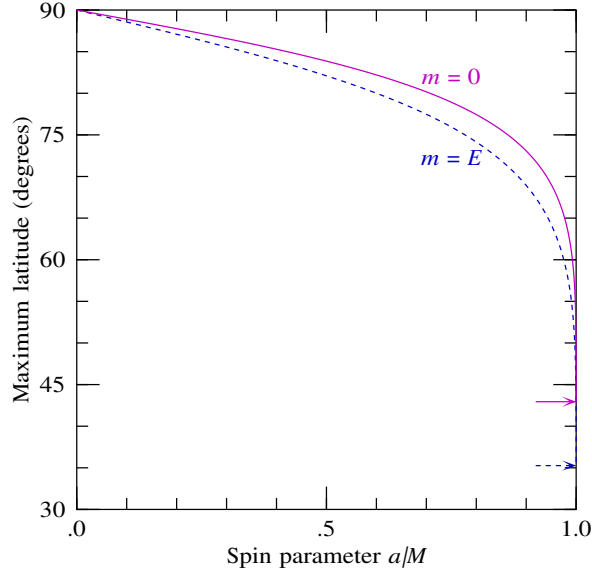


Figure 24.1 Massless (solid) or massive (dashed) collisionless particles that fall from infinite radius can reach the inner horizon and be either outgoing or ingoing only up to a certain maximum latitude on the inner horizon, shown here. At higher latitudes on the inner horizon, particles that free fall from infinity are necessarily ingoing at the inner horizon. The maximum accessible latitude depends on the spin of the black hole. The maximum latitude varies from 90° (all latitudes are accessible) for a Schwarzschild (non-spinning) black hole, to $\arcsin \sqrt{-3 + 2\sqrt{3}} = 42^\circ 9'$ ($m = 0$) or $\arcsin \sqrt{1/3} = 35^\circ 3'$ ($m = E$), arrowed, for an extremal Kerr black hole.

3. What happens at the poles on the inner horizon?

Solution.

1. Particles between the outer and inner horizons are outgoing or ingoing as the Hamilton-Jacobi time parameter P_t , equation (23.5a), is positive or negative, §23.5. Particles that fall through the outer horizon are necessarily ingoing at the outer horizon, requiring $P_t < 0$ at the outer horizon. However, particles with sufficiently positive angular momentum L can turn around and become outgoing at the inner horizon. The division between outgoing and ingoing at the inner horizon $r = r_-$ occurs when P_t is zero at the inner horizon. For a Kerr black hole, where $P_t = -E + L\omega_x$, particles accreted from outside the outer horizon are outgoing or ingoing at the inner horizon as (note that $\omega_x^- > \omega_x^+ > 0$)

$$\begin{cases} L\omega_x^- > E > L\omega_x^+ & \text{outgoing ,} \\ E > \max(L\omega_x^+, L\omega_x^-) & \text{ingoing .} \end{cases} \quad (24.10)$$

2. To reach a given latitude θ at the inner horizon, P_y must be positive, which imposes a lower limit on

the Carter constant $\mathcal{K}$,

$$\mathcal{K} \geq \frac{P_\phi^2}{\Delta_y} + m^2 \rho_y^2 . \quad (24.11)$$

Particles cannot turn around in radius between the outer and inner horizons. Outside the outer horizon, a particle can reach a radius r as long as P_x is positive, which imposes an upper limit on the Carter constant $\mathcal{K}$,

$$\mathcal{K} \leq \frac{P_t^2}{\Delta_x} - m^2 \rho_x^2 . \quad (24.12)$$

A necessary condition for a trajectory to extend to infinite radius is that the particle energy exceed its rest mass, $E \geq m$. Given that condition, the right hand side of equation (24.12) tends to ∞ at $r \rightarrow r_+$ and at $r \rightarrow \infty$, and is a minimum at some radius in between. The condition that a trajectory can start at infinite radius and reach a given latitude inside the outer horizon is

$$\min \left(\frac{P_t^2}{\Delta_x} - m^2 \rho_x^2 \right) \geq \frac{P_\phi^2}{\Delta_y} + m^2 \rho_y^2 , \quad (24.13)$$

where the minimum on the left hand side is over radius r from r_+ to ∞ . The condition (24.13) along with the condition that $P_t = 0$ at the inner horizon translates into a condition on the maximum latitude at any given spin parameter a , illustrated in Figure 24.1.

3. Poles occur where $\Delta_y = 0$. A particle can reach a pole only if $P_y = P_\phi = 0$ there, equation (23.20). This requires that $L = 0$ and

$$\mathcal{K} \geq m^2 \rho_y^2 . \quad (24.14)$$

Since $L = 0$, the time Hamilton-Jacobi parameter P_t defined by equation (23.5a) is a constant, $P_t = \pi_t = -E$. Since the sign of P_t between the horizons determines whether the particle is outgoing ($P_t > 0$) or ingoing ($P_t < 0$), and since a particle falling through the outer horizon is necessarily ingoing, it follows that a particle that falls from outside the outer horizon to a pole on the inner horizon must remain ingoing. The limiting case is for a massless particle, $m = 0$, falling along the principal ingoing null direction along the polar axis. This polar null geodesic has $\mathcal{K} = P_t = E = L = 0$. However, $L/E = \omega_y \rightarrow 0$ on the polar ingoing null geodesic, equation (23.26), which is on the ingoing side of the outgoing/ingoing divide (24.10). Thus there are no geodesics that fall from outside the outer horizon and are outgoing when they reach a pole on the inner horizon. On the other hand it is possible for ingoing photons to scatter off gas or dust inside the outer horizon and thereby become outgoing when they reach the inner horizon, at any latitude.

Exercise 24.3 Inflationary Kasner solution. The inflationary and collapse stages of inflation can be approximated by a Kasner line-element (17.161) with two scale factors equal, $a_2 = a_3$,

$$ds^2 = -dt^2 + a_1^2 dx_1^2 + a_2^2 (dx_2^2 + dx_3^2) . \quad (24.15)$$

All scale factors are functions $a_\alpha(t)$ only of time t . The tetrad-frame inflationary energy-momentum is diagonal with $T_{00} = T_{11}$ and $T_{22} = T_{33} = 0$. The goal is to find scale factors a_1 and a_2 that yield such.

1. Show that the tetrad-frame Einstein tensor that follows from the Kasner line-element (24.15) is diagonal with $G_{22} = G_{33}$.
2. Define the time $T(t)$ (not to be confused with energy-momenta T_{mn}) by

$$dt = -a^3 d \ln |T| . \quad (24.16)$$

In the inflationary context, T is negative, varying from $-\infty$ in the distant past to -0 at the singularity. The minus sign in equation (24.16) ensures that t increases as T increases. Show that the condition $G_{00} = G_{11}$ requires that a_2 be proportional to some power of $|T|$,

$$a_2 \propto |T|^b , \quad (24.17)$$

with b some arbitrary constant.

3. Show that the condition $G_{22} = 0$ implies that

$$a_1 \propto \frac{e^{ca_2^2}}{\sqrt{a_2}} , \quad (24.18)$$

with c some arbitrary constant.

4. Without loss of generality scale the time T so that $b = \frac{1}{2}$ and $c = 1$. With a convenient scaling of the coordinates x_α , the scale factors a_α are

$$a_1 = \frac{e^{|T|}}{|T|^{1/4}} , \quad a_2 = |T|^{1/2} , \quad a^3 \equiv a_1 a_2^2 = |T|^{3/4} e^{|T|} . \quad (24.19)$$

There is a BKL bounce where a_1 goes through its minimum value at $T = -\frac{1}{4}$. Show that

$$G_{00} = G_{11} = \frac{1}{a_1^2 a_2^2} = \frac{e^{-2|T|}}{|T|^{1/2}} . \quad (24.20)$$

Show that the only nonvanishing component of the tetrad-frame Weyl tensor is its spin 0 part,

$$C = -\frac{1}{8a^6} = -\frac{e^{-2|T|}}{8|T|^{3/2}} . \quad (24.21)$$

5. Define the Kasner coefficients q_α by

$$q_\alpha \equiv \frac{d \ln a_\alpha}{d \ln a} , \quad (24.22)$$

which is defined so that $\sum_\alpha q_\alpha = 1$. Show that

$$q_1 = \frac{|T| - \frac{1}{4}}{|T| + \frac{3}{4}} , \quad q_2 = q_3 = \frac{\frac{1}{2}}{|T| + \frac{3}{4}} , \quad (24.23)$$

with asymptotic behaviour

$$\{q_1, q_2\} \rightarrow \begin{cases} \{1, 0\} & T \rightarrow -\infty , \\ \{-\frac{1}{3}, \frac{2}{3}\} & T \rightarrow -0 . \end{cases} \quad (24.24)$$

Conclude that

$$\sum_{\alpha} q_{\alpha}^2 = 1 - \frac{2|T|}{(|T| + \frac{3}{4})^2} . \quad (24.25)$$

6. Geodesics follow from the 3 integrals of motion p_{α} associated with spatial homogeneity, and 1 integral of motion $p^{\mu}p_{\mu} = -m^2$ associated with conservation of rest mass m . Show that the tetrad-frame Einstein tensor can be realised by the sum of energy-momenta of two collisionless streams of massless particles, one outgoing (+) and the other ingoing (−)

$$G_{mn} = 8\pi(T_{mn}^{+} + T_{mn}^{-}) , \quad T_{mn}^{\pm} = N p_m^{\pm} p_n^{\pm} , \quad (24.26)$$

with tetrad-frame momenta

$$p_{\pm}^m = \frac{1}{a_1} \{1, \pm 1, 0, 0\} , \quad (24.27)$$

and tetrad-frame number densities $n_{\pm}^m = N p_{\pm}^m$ where N is the scalar density

$$N = \frac{1}{16\pi a_2^2} . \quad (24.28)$$

The momenta satisfy the geodesic equation $p_{\pm}^m D_m p_{\pm}^n = 0$, and the number densities satisfy number conservation $D_m n_{\pm}^m = 0$.

7. The tetrad-frame 4-momentum along a geodesic of a particle of mass m is

$$p^m = \{p^0, p^a\} = \left\{ \sqrt{\sum_{\alpha} \frac{p_{\alpha}^2}{a_{\alpha}^2} + m^2} , \frac{p_{\alpha}}{a_{\alpha}} \right\} . \quad (24.29)$$

With respect to coordinates $x^{\mu} \equiv \{T, x_{\alpha}\}$, the coordinate 4-momentum along a geodesic is

$$\frac{dx^{\mu}}{d\lambda} \equiv \left\{ \frac{dT}{d\lambda}, \frac{dx_{\alpha}}{d\lambda} \right\} \equiv p^{\mu} = e_m^{\mu} p^m = \left\{ \frac{|T|}{a^3} \sqrt{\sum_{\alpha} \frac{p_{\alpha}^2}{a_{\alpha}^2} + m^2} , \frac{p_{\alpha}}{a_{\alpha}^2} \right\} . \quad (24.30)$$

Draw null geodesics to see what the scene looks like to an observer at rest in the tetrad frame.

8. Show that the ratio of emitted to observed tetrad-frame frequencies $\omega \equiv p^0$ for an observer at rest at time T watching a distant emitter at rest at time $T = T_{\infty} \rightarrow -\infty$ in a direction angled θ away from the 1-axis (x -axis) is

$$\frac{\omega_{\text{em}}}{\omega_{\text{obs}}} = \frac{\omega(T_{\infty})}{\omega(T)} \rightarrow \sqrt{T/T_{\infty}} \sin \theta . \quad (24.31)$$

The proper time experienced by the rest observer is $\tau = t$. Conclude that the acceleration of the distant emitter perceived by the rest observer is

$$\kappa \equiv \frac{d \ln(\omega_{\text{em}}/\omega_{\text{obs}})}{d\tau} = -\frac{1}{2a^3} = -\frac{1}{2}|T|^{-3/4} e^{-|T|} . \quad (24.32)$$

Conclude that the acceleration diverges at the singularity as

$$\kappa \propto -|T|^{-3/4} \propto -|\tau|^{-1} \quad \text{as } \tau \rightarrow 0 . \quad (24.33)$$

Black hole thermodynamics

For an ideal Λ -Kerr-Newman black hole, variations of the black hole's mass M , electric charge Q , angular momentum $J \equiv aM$, and of the cosmological constant $\Lambda \equiv 8\pi G\rho_\Lambda$ are related to variations of the area $A = 4\pi R^2 = 4\pi(r^2 + a^2)$ of the horizon by

$$dM = \frac{\kappa}{8\pi} dA + \Phi dQ + \omega dJ - V d\rho_\Lambda , \quad (25.1)$$

where κ is the acceleration, Φ the electric potential, ω the angular velocity, and V the enclosed volume at the horizon,

$$\kappa = \frac{r - M}{R^2} - \frac{2\Lambda r}{3} , \quad (25.2a)$$

$$\Phi = \frac{Qr}{R^2} , \quad (25.2b)$$

$$\omega = \frac{a}{R^2} , \quad (25.2c)$$

$$V = \frac{4}{3}\pi r R^2 . \quad (25.2d)$$

Equations (25.1) and (25.2) hold at any horizon, wherever the horizon function Δ_x vanishes, including at a cosmological horizon, which exists if the vacuum energy ρ_Λ is positive. The acceleration κ satisfies

$$\kappa = -\frac{1}{2} \frac{d\Delta_x}{dx} . \quad (25.3)$$

The acceleration κ vanishes when the horizon is extremal, that is, where two horizons merge into one, which happens when the horizon function Δ_x is not only zero but also an extremum.

Equation (25.1) can be recast as

$$d(M + \rho_\Lambda V) = \frac{1}{8\pi\kappa} dA + \Phi dQ + \omega dJ - p_\Lambda dV , \quad (25.4)$$

in which the energy within the horizon is taken to be the energy $M + \rho_\Lambda V$ including the contribution from vacuum energy.

Exercise 25.1 Entropy in Hawking radiation. Compare the entropy emitted in Hawking radiation to the Bekenstein-Hawking entropy lost by a black hole when it emits a certain energy dE in Hawking radiation. Assume for simplicity that the emitted radiation carries no charge or angular momentum.

Solution. The entropy S of an ideal gas in thermodynamic equilibrium with zero chemical potential is related to its energy E by, equation (30.18),

$$\frac{T dS}{dE} = (1 + p/\rho) . \quad (25.5)$$

For relativistic radiation, $p/\rho = \frac{1}{3}$, equation (25.6) becomes

$$\frac{T dS}{dE} = \frac{4}{3} . \quad (25.6)$$

If it is assumed that the Hawking radiation carries no charge or angular momentum, then the Bekenstein-Hawking entropy lost by the black hole is

$$\frac{T dS}{dE} = 1 . \quad (25.7)$$

Thus the entropy emitted in Hawking radiation exceeds the Bekenstein-Hawking entropy lost by the black hole by factor $\frac{4}{3}$. A more careful treatment gives a slightly different result (Zurek, 1982; Page, 1983).

Exercise 25.2 Area of the horizon. What is the area of the horizon of a stationary black hole?

Solution. The 2-dimensional angular line-element of the separable line-element (22.1) is

$$dl^2 = \rho^2 \left[\frac{dy^2}{\Delta_y} + \frac{(\Delta_y - \omega_y^2 \Delta_x) d\phi^2}{(1 - \omega_x \omega_y)^2} \right] . \quad (25.8)$$

The angular line-element is diagonal, with proper distances in the two orthogonal y and ϕ directions

$$\frac{\rho dy}{\sqrt{\Delta_y}} , \quad \frac{\rho \sqrt{\Delta_y - \omega_y^2 \Delta_x} d\phi}{1 - \omega_x \omega_y} . \quad (25.9)$$

The area of the angular y - ϕ surface at fixed radius x and time t is obtained by integrating the product of the proper distances over the surface,

$$A = \int \frac{\rho^2 \sqrt{(1 - \omega_y^2 \Delta_x / \Delta_y)}}{1 - \omega_x \omega_y} dy d\phi . \quad (25.10)$$

Horizons occur where the horizon function vanishes, $\Delta_x = 0$, in which case the area simplifies to

$$\begin{aligned}
 A &= \int \frac{\rho^2}{1 - \omega_x \omega_y} dy d\phi \\
 &= 2\pi \int \frac{1}{(f_0 + f_1 \omega_x)(f_1 + f_0 \omega_y)} dy \\
 &= \frac{2\pi}{f_0 + f_1 \omega_x} \int \frac{d\omega_y}{2\sqrt{(f_1 + f_0 \omega_y)^3 (g_1 - g_0 \omega_y)}} \\
 &= \frac{2\pi}{(f_0 g_1 + f_1 g_0)(f_0 + f_1 \omega_x)} \left[\sqrt{\frac{g_1 - g_0 \omega_y}{f_1 + f_0 \omega_y}} \right]. \tag{25.11}
 \end{aligned}$$

The second line of equations invokes equation (22.39a), while the third line uses equation (22.44b) to transform the integral over y to an integral over ω_y . The constants are given by equation (22.72) for Λ -Kerr-Newman, or equation (22.80) for Taub-NUT. The integration over y is from -1 to 1 , north to south pole. For Λ -Kerr-Newman, $\omega_y = 0$ at both poles, but for Taub-NUT, $\omega_y = 2N_\bullet(c_\bullet \pm 1)$ at the poles, equation (22.83). In either case, for both Λ -Kerr-Newman and Taub-NUT, the area of the horizon is

$$A = 4\pi R^2, \tag{25.12}$$

where R is given by equation (22.7) for Λ -Kerr-Newman, and equations (22.83) for Taub-NUT.

Concept Questions

1. Why do general relativistic perturbation theory use the tetrad formalism as opposed to the coordinate approach?
2. Why is the tetrad metric γ_{mn} assumed fixed in the presence of perturbations?
3. Are the tetrad axes γ_m fixed under a perturbation?
4. Is it true that the tetrad components φ_{mn} of a perturbation are (anti-)symmetric in $m \leftrightarrow n$ if and only if its coordinate components $\varphi_{\mu\nu}$ are (anti-)symmetric in $\mu \leftrightarrow \nu$?
5. Does an unperturbed quantity, such as the unperturbed metric $g_{\mu\nu}^0$, change under an infinitesimal coordinate gauge transformation?
6. How can the vierbein perturbation φ_{mn} be considered a tetrad tensor field if it changes under an infinitesimal coordinate gauge transformation?
7. What properties of the unperturbed spacetime allow decomposition of perturbations into independently evolving Fourier modes?
8. What properties of the unperturbed spacetime allow decomposition of perturbations into independently evolving scalar, vector, and tensor modes?
9. In what sense do scalar, vector, and tensor modes have spin 0, 1, and 2 respectively?
10. Tensor modes represent gravitational waves that, in vacuo, propagate at the speed of light. Do scalar and vector modes also propagate at the speed of light in vacuo? If so, do scalar and vector modes also constitute gravitational waves?
11. If scalar, vector, and tensor modes evolve independently, does that mean that scalar modes can exist and evolve in the complete absence of tensor modes? If so, does it mean that scalar modes can propagate causally, in vacuo at the speed of light, without any tensor modes being present?
12. Equation (27.77) defines the mass M of a body as what a distant observer would measure from its gravitational potential. Similarly equation (27.85) defines the angular momentum $\mathbf{L}$ of a body as what a distant observer would measure from the dragging of inertial frames. In what sense are these definitions legitimate?
13. Can an observer far from a body detect the difference between the scalar potentials Ψ and Φ produced by the body?

14. If a gravitational wave is a wave of spacetime itself, distorting the very rulers and clocks that measure spacetime, how is it possible to measure gravitational waves at all?
15. If gravitational waves carry energy-momentum, then can gravitational waves be present in a region of spacetime with vanishing energy-momentum tensor, $T_{mn} = 0$?
16. Have gravitational waves been detected?

What's important?

1. Getting your brain around coordinate and tetrad gauge transformations.
2. A central aim of general relativistic perturbation theory is to identify the coordinate and tetrad gauge-invariant perturbations, since only these have physical meaning.
3. A second central aim is to classify perturbations into independently evolving modes, to the extent that this is possible.
4. In background spacetimes with spatial translation and rotation symmetry, which includes Minkowski space and the Friedmann-Lemaître-Robertson-Walker metric of cosmology, modes decompose into independently evolving scalar (spin 0), vector (spin 1), and tensor (spin 2) modes. In background spacetimes without spatial translation and rotation symmetry, such as black holes, scalar, vector, and tensor modes scatter off the curvature of space, and therefore mix with each other.
5. In background spacetimes with spatial translation and rotation symmetry, there are 6 algebraic combinations of metric coefficients that are coordinate and tetrad gauge-invariant, and therefore represent physical perturbations. There are 2 scalar modes, 2 vector modes, and 2 tensor modes. A spin m mode varies as $e^{-im\chi}$ where χ is the rotational angle about the spatial wavevector $\mathbf{k}$ of the mode.
6. In background spacetimes without spatial translation and rotation symmetry, the coordinate and tetrad gauge-invariant perturbations are not algebraic combinations of the metric coefficients, but rather combinations that involve first and second derivatives of the metric coefficients. Gravitational waves are described by the Weyl tensor, which can be decomposed into 5 complex components, with spin 0, ± 1 , and ± 2 . The spin ± 2 components describe propagating gravitational waves, while the spin 0 and spin ± 1 components describe the non-propagating gravitational field near a source.
7. The preeminent application of general relativistic perturbation theory is to cosmology. Coupled with physics that is either well understood (such as photon-electron scattering) or straightforward to model even without a deep understanding (such as the dynamical behaviour of non-baryonic dark matter and dark energy), the theory has yielded predictions that are in spectacular agreement with observations of fluctuations in the CMB and in the large scale distribution of galaxies and other tracers of the distribution of matter in the Universe.

Perturbations and gauge transformations

This Chapter sets up the basic equations that define perturbations to an arbitrary spacetime in the tetrad formalism of general relativity, and it examines the effect of tetrad and coordinate gauge transformations on those perturbations. The perturbations are supposed to be small, in the sense that quantities quadratic in the perturbations can be neglected. The formalism set up in this Chapter provides a foundation used in subsequent Chapters.

26.1 Notation for perturbations

A 0 (zero) overscript signifies an unperturbed quantity, while a 1 (one) overscript signifies a perturbation. No overscript means the full quantity, including both unperturbed and perturbed parts. An overscript is attached only where necessary. Thus if the unperturbed part of a quantity is zero, then no overscript is needed, and none is attached.

The vierbein of the unperturbed background is e_{μ}^n . In this and the next several sections up to and including §26.7, the unperturbed vierbein e_{μ}^n is an arbitrary differentiable function of arbitrary coordinates x^{μ} .

26.2 Vierbein perturbation

Let the vierbein perturbation φ_{mn} be defined so that the perturbed inverse vierbein is

$$e_m^{\mu} = (\delta_m^n + \varphi_m^n) e_n^{\mu}, \quad (26.1)$$

with corresponding perturbed vierbein

$$e_{\mu}^m = (\delta_n^m - \varphi_n^m) e_{\mu}^n. \quad (26.2)$$

Since the perturbation φ_m^n is already of linear order, to linear order its indices can be raised and lowered with the unperturbed metric, and transformed between tetrad and coordinate frames with the unperturbed

vierbein. In practice it proves convenient to work with the covariant tetrad-frame components φ_{mn} of the vierbein perturbation

$$\varphi_{mn} = \gamma_{nl} \varphi_m^l. \quad (26.3)$$

In terms of the covariant perturbation φ_{mn} , the perturbed inverse vierbein (26.1) is

$$e_m^\mu = (\gamma_{mn} + \varphi_{mn}) e^{n\mu}, \quad (26.4)$$

The perturbation φ_{mn} can be regarded as a tetrad tensor field defined on the unperturbed background.

Concept question 26.1 What are the units of the vierbein perturbation φ_{mn} ? The vierbein perturbation φ_{mn} is dimensionless, as are the coordinate metric $g_{\mu\nu}$, the tetrad metric γ_{mn} , and the vierbein e^m_μ and its inverse e_m^μ .

26.3 Gauge transformations

The vierbein perturbation φ_{mn} has 16 degrees of freedom, but only 6 of these degrees of freedom correspond to real physical perturbations, since 6 degrees of freedom are associated with arbitrary infinitesimal changes in the choice of tetrad, which is to say arbitrary infinitesimal Lorentz transformations, and a further 4 degrees of freedom are associated with arbitrary infinitesimal changes in the coordinates.

In the context of perturbation theory, these infinitesimal tetrad and coordinate transformations are called **gauge transformations**. Real physical perturbations are perturbations that are **gauge-invariant** under both tetrad and coordinate gauge transformations.

26.4 Tetrad metric assumed constant

In the tetrad formalism, tetrad axes γ_m are introduced as locally inertial (or other physically motivated) axes attached to an observer. The axes enable quantities to be projected into the frame of the observer. In a spacetime buffeted by perturbations, it is natural for an observer to cling to the rock provided by the locally inertial (or other) axes, as opposed to allowing the axes to bend with the wind. For example, when a gravitational wave goes by, the tidal compression and rarefaction causes the proper distance between two freely falling test masses to oscillate, Fig. 27.1. It is natural to choose the tetrad so that it continues to measure proper times and distances in the perturbed spacetime.

In the treatment of general relativistic perturbation theory in this book, the tetrad metric is taken to be constant everywhere, and unchanged by a perturbation

$$\boxed{\gamma_{mn} = \overset{0}{\gamma}_{mn} = \text{constant}}. \quad (26.5)$$

For example, if the tetrad is orthonormal, then the tetrad metric is constant, the Minkowski metric η_{mn} .

However, the tetrad could also be some other tetrad for which the tetrad metric is constant, such as a spin tetrad (§40.1), or a Newman-Penrose tetrad (§41.1.1).

26.5 Perturbed coordinate metric

The perturbed coordinate metric is

$$\begin{aligned} g_{\mu\nu} &= \gamma_{mn} e^m_{\mu} e^n_{\nu} \\ &= \gamma_{kl} (\delta_m^k - \varphi_m^k) e^m_{\mu} (\delta_n^l - \varphi_n^l) e^n_{\nu} \\ &= \overset{0}{g}_{\mu\nu} - (\varphi_{\mu\nu} + \varphi_{\nu\mu}) . \end{aligned} \quad (26.6)$$

Thus the perturbation of the coordinate metric depends only on the symmetric part of the vierbein perturbation φ_{mn} , not the antisymmetric part

$$\overset{1}{g}_{\mu\nu} = -(\varphi_{\mu\nu} + \varphi_{\nu\mu}) . \quad (26.7)$$

26.6 Tetrad gauge transformations

Under an infinitesimal tetrad (Lorentz) transformation, the covariant vierbein perturbations φ_{mn} transform as

$$\varphi_{mn} \rightarrow \varphi_{mn} + \epsilon_{mn} , \quad (26.8)$$

where ϵ_{mn} is the generator of a Lorentz transformation, which is to say an arbitrary antisymmetric tensor (Exercise 11.2). Thus the antisymmetric part $\varphi_{mn} - \varphi_{nm}$ of the covariant perturbation φ_{mn} is arbitrarily adjustable through an infinitesimal tetrad transformation, while the symmetric part $\varphi_{mn} + \varphi_{nm}$ is tetrad gauge-invariant. This is consistent with the fact that the perturbation (26.7) of the coordinate metric depends only on the symmetric part of the perturbation, and the coordinate metric is tetrad gauge-invariant.

It is easy to see when a quantity is tetrad gauge-invariant: it is tetrad gauge-invariant if and only if it depends only on the symmetric part of the vierbein perturbation, not on the antisymmetric part. Evidently the perturbation (26.7) to the coordinate metric $g_{\mu\nu}$ is tetrad gauge-invariant. This is as it should be, since the coordinate metric $g_{\mu\nu}$ is a coordinate-frame quantity, independent of the choice of tetrad frame.

If only tetrad gauge-invariant perturbations are physical, why not just discard tetrad perturbations (the antisymmetric part of φ_{mn}) altogether, and work only with the tetrad gauge-invariant part (the symmetric part of φ_{mn})? The answer is that tetrad-frame quantities such as the tetrad-frame Einstein tensor do change under tetrad gauge transformations (infinitesimal Lorentz transformations of the tetrad). It is true that the only physical perturbations of the Einstein tensor are those combinations of it that are tetrad gauge-invariant. But in order to identify these tetrad gauge-invariant combinations, it is necessary to carry through the dependence on the non-tetrad-gauge-invariant part, the antisymmetric part of φ_{mn} .

Much of the professional literature on general relativistic perturbation theory works with the traditional

coordinate formalism, as opposed to the tetrad formalism. The term “gauge-invariant” then means coordinate gauge-invariant, as opposed to both coordinate and tetrad gauge-invariant. This is fine as far as it goes: the coordinate approach is perfectly able to identify physical perturbations versus gauge perturbations. However, there still remains the problem of projecting the perturbations into the frame of an observer, so ultimately the issue of perturbations of the observer’s frame, tetrad perturbations, must be faced.

Concept question 26.2 **Non-infinitesimal tetrad transformations in perturbation theory?** In perturbation theory, can tetrad gauge transformations be non-infinitesimal?

26.7 Coordinate gauge transformations

A coordinate gauge transformation is a transformation of the coordinates x^μ by an infinitesimal shift ϵ^μ

$$x^\mu \rightarrow x'^\mu = x^\mu + \epsilon^\mu . \quad (26.9)$$

You should not think of this as shifting the underlying spacetime around; rather, it is just a change of the coordinate system, which leaves the underlying spacetime unchanged. Because the shift ϵ^μ is, like the vierbein perturbations φ_{mn} , already of linear order, its indices can be raised and lowered with the unperturbed metric, and transformed between coordinate and tetrad frames with the unperturbed vierbein. Thus the shift ϵ^μ can be regarded as a vector field defined on the unperturbed background. The tetrad components ϵ^m of the shift ϵ^μ are

$$\epsilon^m = e^m{}_\mu \epsilon^\mu . \quad (26.10)$$

Physically, the tetrad-frame shift ϵ^m is the shift measured in locally inertial coordinates ξ^m ,

$$\xi^m \rightarrow \xi'^m = \xi^m + \epsilon^m . \quad (26.11)$$

26.7.1 The change in any tensor under a coordinate transformation is minus its Lie derivative

As discussed in §7.34, the change in any coordinate tensor $A^{\kappa\lambda\dots}_{\mu\nu\dots}(x)$ under a coordinate gauge transformation (26.9) is minus its Lie derivative $\mathcal{L}_\epsilon$ with respect to the infinitesimal shift ϵ ,

$$A^{\kappa\lambda\dots}_{\mu\nu\dots}(x) \rightarrow A'^{\kappa\lambda\dots}_{\mu\nu\dots}(x) = A^{\kappa\lambda\dots}_{\mu\nu\dots}(x) - \mathcal{L}_\epsilon A^{\kappa\lambda\dots}_{\mu\nu\dots} . \quad (26.12)$$

The Lie derivative $\mathcal{L}_\epsilon A^{\kappa\lambda\dots}_{\mu\nu\dots}$ is given by formula (7.157). Under a coordinate gauge transformation (26.9), the coordinate of a fixed physical position transforms from x to x' . But in perturbation theory, quantities are considered to be functions of coordinate position x , which does not remain at a fixed physical position under a coordinate transformation. As discussed in §7.34, the Lie derivative is defined such that the transformed tensor $A'^{\kappa\lambda\dots}_{\mu\nu\dots}(x)$ is evaluated at fixed coordinate position x , not at fixed physical position.

26.7.2 Coordinate gauge transformation of a tetrad tensor

A tetrad-frame 4-vector A^m is a coordinate-invariant quantity, and therefore acts like a coordinate scalar under a coordinate gauge transformation (26.9). Thus a tetrad frame 4-vector A^m must be treated as a coordinate scalar when its Lie derivative is taken. Under a coordinate gauge transformation (26.9), a tetrad-frame 4-vector A^m transforms as

$$A^m(x) \rightarrow A'^m(x) = A^m(x) - \mathcal{L}_\epsilon A^m, \quad (26.13)$$

where the Lie derivative is, equation (7.139),

$$\mathcal{L}_\epsilon A^m = \epsilon^\kappa \frac{\partial A^m}{\partial x^\kappa} \quad \text{not a tetrad tensor.} \quad (26.14)$$

The change $\epsilon^\kappa \partial_\kappa A^m$ is a coordinate tensor (specifically, a coordinate scalar), but not a tetrad tensor.

More generally, a tetrad-frame tensor $A_{mn\dots}^{kl\dots}$ transforms under a coordinate gauge transformation (26.9) as

$$A_{mn\dots}^{kl\dots}(x) \rightarrow A'_{mn\dots}{}^{kl\dots}(x) = A_{mn\dots}^{kl\dots}(x) - \mathcal{L}_\epsilon A_{mn\dots}^{kl\dots}, \quad (26.15)$$

where the Lie derivative is

$$\mathcal{L}_\epsilon A_{mn\dots}^{kl\dots} = \epsilon^\pi \partial_\pi A_{mn\dots}^{kl\dots} \quad \text{not a tensor.} \quad (26.16)$$

Again, the change $-\epsilon^\pi \partial_\pi A_{mn\dots}^{kl\dots}$ is a coordinate tensor (a coordinate scalar), but not a tetrad tensor.

Concept question 26.3 Should not the Lie derivative of a tetrad tensor be a tetrad tensor?

The Lie derivative of a tetrad tensor, as defined in this book, is a coordinate tensor but not a tetrad tensor. Would it not be better to define the Lie derivative so it is a tetrad tensor as well as a coordinate tensor?

Answer. In this book, the Lie derivative of any quantity is *defined* to be minus the variation of the quantity under a coordinate transformation. This definition is unambiguous; and it implies that the Lie derivative of a tetrad tensor is not a tetrad tensor.

26.7.3 Coordinate gauge transformation of the vierbein

The inverse vierbein $e_m{}^\mu$ is a coordinate vector and a tetrad vector. It transforms under a coordinate gauge transformation (26.9) as

$$e_m{}^\mu(x) \rightarrow e'_m{}^\mu(x) - \mathcal{L}_\epsilon e_m{}^\mu, \quad (26.17)$$

where the Lie derivative of the inverse vierbein is, equation (7.143),

$$\begin{aligned}
 \mathcal{L}_\epsilon e_m{}^\mu &= -e_m{}^\kappa \frac{\partial \epsilon^\mu}{\partial x^\kappa} + \epsilon^\kappa \frac{\partial e_m{}^\mu}{\partial x^\kappa} \\
 &= -\partial_m (e^{n\mu} \epsilon_n) + \epsilon^k \partial_k e_m{}^\mu \\
 &= -e^{n\mu} (\partial_m \epsilon_n - \epsilon^k d_{nkm} + \epsilon^k d_{nmk}) \\
 &= -e^{n\mu} \left[\partial_m \epsilon_n + \epsilon^k (\overset{\circ}{\Gamma}_{nkm} - \overset{\circ}{\Gamma}_{nmk}) \right].
 \end{aligned} \tag{26.18}$$

On the third line the vierbein derivatives have been replaced by d_{nkm} defined by equation (11.33), while on the fourth line $\overset{\circ}{\Gamma}_{nkm}$ is the torsion-free tetrad-frame connection, defined in terms of the vierbein derivatives by equation (11.54).

26.7.4 Coordinate transformation of the vierbein perturbation

According to equation (26.1), the perturbation $\overset{1}{e}_m{}^\mu$ of the inverse vierbein may be expressed in terms of a covariant vierbein perturbation field φ_{mn} ,

$$\overset{1}{e}_m{}^\mu = \varphi_{mn} \overset{0}{e}^{n\mu}. \tag{26.19}$$

The perturbation induced by a coordinate gauge transformation (26.9) equals the Lie derivative given by equation (26.18), $\overset{1}{e}_m{}^\mu = \mathcal{L}_\epsilon e_m{}^\mu$. Consequently the vierbein perturbation φ_{mn} transforms under a coordinate gauge transformation (26.9) as

$$\boxed{\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \partial_m \epsilon_n + (\overset{\circ}{\Gamma}_{nkm} - \overset{\circ}{\Gamma}_{nmk}) \epsilon^k}, \tag{26.20}$$

with $\overset{\circ}{\Gamma}_{nkm}$ the torsion-free tetrad-frame connection. This is the fundamental formula that gives the effect of coordinate transformations on the vierbein perturbations in any background spacetime.

Concept question 26.4 Variation of unperturbed quantities under coordinate gauge transformations? How does an unperturbed quantity, such as the unperturbed coordinate metric $\overset{0}{g}_{\mu\nu}$, vary under an infinitesimal coordinate gauge transformation? **Answer.** It doesn't. Any variation is considered to be part of the perturbation.

26.8 Scalar, vector, tensor decomposition of perturbations

In the particular case that the unperturbed spacetime is spatially homogeneous and isotropic, which includes not only Minkowski space but also the important case of the cosmological Friedmann-Lemaître-Robertson-Walker metric, perturbations decompose into independently evolving scalar (spin 0), vector (spin 1), and tensor (spin 2) modes.

Similarly to Fourier decomposition, decomposition into scalar, vector, and tensor modes is non-local, in

principle requiring knowledge of perturbation amplitudes simultaneously throughout all of space. In practical problems however, an adequate decomposition is possible as long as the scales probed are sufficiently larger than the wavelengths of the modes probed. Ultimately, the fact that an adequate decomposition is possible is a consequence of the fact that gravitational fluctuations in the real Universe appear to converge at the cosmological horizon, so that what happens locally is largely independent of what is happening far away.

26.8.1 Decomposition of a vector in flat 3D space

Theorem: In flat 3-dimensional space, a 3-vector field $\mathbf{w}(\mathbf{x})$ can be decomposed uniquely (subject to the boundary condition that $\mathbf{w}$ vanishes sufficiently rapidly at infinity) into a sum of scalar and vector parts

$$\boxed{\mathbf{w} = \underset{\text{scalar}}{\nabla w_{\parallel}} + \underset{\text{vector}}{\mathbf{w}_{\perp}}} . \quad (26.21)$$

In this context, the term **vector** signifies a 3-vector $\mathbf{w}_{\perp}$ that is transverse, that is to say, it has vanishing divergence,

$$\boxed{\nabla \cdot \mathbf{w}_{\perp} = 0} . \quad (26.22)$$

Here $\nabla \equiv \partial/\partial\mathbf{x} \equiv \nabla_a \equiv \partial/\partial x^a$ is the gradient in flat 3D space. The scalar and vector parts are also known as spin 0 and spin 1, or gradient and curl, or longitudinal and transverse. The scalar part $\nabla w_{\parallel}$ contains 1 degree of freedom, while the vector part $\mathbf{w}_{\perp}$ contains 2 degrees of freedom. Together they account for the 3 degrees of freedom of the vector $\mathbf{w}$.

Proof: Take the divergence of equation (26.21)

$$\nabla \cdot \mathbf{w} = \nabla^2 w_{\parallel} . \quad (26.23)$$

The operator ∇^2 on the right hand side of equation (26.23) is the 3D Laplacian. The solution of equation (26.23) is

$$w_{\parallel}(\mathbf{x}) = - \int \frac{\nabla' \cdot \mathbf{w}(\mathbf{x}')}{|\mathbf{x}' - \mathbf{x}|} \frac{d^3 x'}{4\pi} . \quad (26.24)$$

The solution (26.24) is valid subject to boundary conditions that the vector $\mathbf{w}$ vanish sufficiently rapidly at infinity. In cosmology, the required boundary conditions, which are set at the Big Bang, are apparently satisfied because fluctuations at the Big Bang were small. Equation (26.21) then immediately implies that the vector part is $\mathbf{w}_{\perp} = \mathbf{w} - \nabla w_{\parallel}$.

It is sometimes convenient to abbreviate $\nabla w_{\parallel} = \mathbf{w}_{\parallel}$ (distinguished by bold face $\mathbf{w}_{\parallel}$ instead of normal face $w_{\parallel}$), so that the decomposition (26.21) is

$$\mathbf{w} = \underset{\text{scalar}}{\mathbf{w}_{\parallel}} + \underset{\text{vector}}{\mathbf{w}_{\perp}} . \quad (26.25)$$

26.8.2 Fourier version of the decomposition of a vector in flat 3D space

When the background has some symmetry, it is natural to expand perturbations in eigenmodes of the symmetry. If the background space is flat, then it is translation symmetric. Eigenmodes of the translation operator ∇ are Fourier modes.

A function $a(\mathbf{x})$ in flat 3D space and its Fourier transform $a(\mathbf{k})$ are related by (the signs and disposition of factors of 2π in the following definition follows the convention most commonly adopted by cosmologists; beware that, with the $-+++$ signature adopted in this book, the convention is opposite to the quantum mechanics convention $\mathbf{p} = \hbar\mathbf{k} = -i\hbar\nabla$ for spatial momentum)

$$a(\mathbf{k}) = \int a(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x, \quad a(\mathbf{x}) = \int a(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}} \frac{d^3k}{(2\pi)^3}. \quad (26.26)$$

You may not be familiar with the practice of using the same symbol a in both real and Fourier space; but a is the same vector in Hilbert space, with components $a_{\mathbf{x}} = a(\mathbf{x})$ in real space, and $a_{\mathbf{k}} = a(\mathbf{k})$ in Fourier space.

Taking the gradient ∇ in real space is equivalent to multiplying by $-i\mathbf{k}$ in Fourier space

$$\boxed{\nabla \rightarrow -i\mathbf{k}}. \quad (26.27)$$

Thus the decomposition (26.21) of the 3D vector $\mathbf{w}$ translates into Fourier space as

$$\mathbf{w} = \underbrace{-i\mathbf{k} w_{\parallel}}_{\text{scalar}} + \underbrace{\mathbf{w}_{\perp}}_{\text{vector}}, \quad (26.28)$$

where the vector part $\mathbf{w}_{\perp}$ satisfies

$$\mathbf{k} \cdot \mathbf{w}_{\perp} = 0. \quad (26.29)$$

In other words, in Fourier space the scalar part $\nabla w_{\parallel}$ of the vector $\mathbf{w}$ is the part parallel (longitudinal) to the wavevector $\mathbf{k}$, while the vector part $\mathbf{w}_{\perp}$ is the part perpendicular (transverse) to the wavevector $\mathbf{k}$.

26.8.3 Decomposition of a tensor in flat 3D space

Similarly, the 9 components of a 3×3 spatial matrix h_{ab} can be decomposed into 3 scalars, 2 vectors, and 1 tensor:

$$h_{ab} = \underbrace{\delta_{ab} \phi}_{\text{scalar}} + \underbrace{\nabla_a \nabla_b h}_{\text{scalar}} + \underbrace{\varepsilon_{abc} \nabla_c \tilde{h}}_{\text{scalar}} + \underbrace{\nabla_a h_b}_{\text{vector}} + \underbrace{\nabla_b \tilde{h}_a}_{\text{vector}} + \underbrace{h_{ab}^T}_{\text{tensor}}. \quad (26.30)$$

In this context, the term **tensor** signifies a 3×3 matrix h_{ab}^T that is traceless, symmetric, and transverse:

$$h^T_a{}^a = 0, \quad h_{ab}^T = h_{ba}^T, \quad \nabla_a h_{ab}^T = 0. \quad (26.31)$$

The transverse-traceless-symmetric matrix h_{ab}^T has two degrees of freedom. The vector components h_a and $\tilde{h}_a$ are by definition transverse,

$$\nabla_a h_a = \nabla_a \tilde{h}_a = 0. \quad (26.32)$$

The tildes on $\tilde{h}$ and $\tilde{h}_a$ simply distinguish those symbols (from h and h_a); the tildes have no other significance. The trace of the 3×3 matrix h_{ab} is

$$h_a^a = 3\phi + \nabla^2 h . \quad (26.33)$$

Perturbations in a flat space background

General relativistic perturbation theory is simplest in the case that the unperturbed background space is Minkowski space. In Cartesian coordinates $x^\mu \equiv \{x^0, x^1, x^2, x^3\} \equiv \{t, x, y, z\}$, the unperturbed coordinate metric is the Minkowski metric

$${}^0g_{\mu\nu} = \eta_{\mu\nu} . \quad (27.1)$$

In this Chapter the tetrad γ_m is taken to be orthonormal, and aligned with the unperturbed coordinate axes ${}^0e_\mu$, so that the unperturbed inverse vierbein is the unit matrix,

$${}^0e_m{}^\mu = \delta_m^\mu , \quad (27.2)$$

and the tetrad metric is Minkowski, $\gamma_{mn} = \eta_{mn}$. Let overdot denote partial differentiation with respect to time t ,

$$\text{overdot} \equiv \frac{\partial}{\partial t} , \quad (27.3)$$

and let ∇ denote the spatial gradient

$$\nabla \equiv \frac{\partial}{\partial \mathbf{x}} \equiv \nabla_a \equiv \frac{\partial}{\partial x^a} . \quad (27.4)$$

Sometimes it will also be convenient to use ∇_m to denote the 4-dimensional spacetime derivative

$$\nabla_m \equiv \left\{ \frac{\partial}{\partial t}, \nabla \right\} . \quad (27.5)$$

27.1 Classification of vierbein perturbations

The aims of this section are two-fold. First, decompose perturbations into scalar, vector, and tensor parts. Second, identify the coordinate and tetrad gauge-invariant perturbations. It will be found, equations (27.13), that there are 6 coordinate and tetrad gauge-invariant perturbations, comprising 2 scalars Ψ and Φ , 1 vector W_a containing 2 degrees of freedom, and 1 tensor h_{ab} containing 2 degrees of freedom.

The vierbein perturbations φ_{mn} defined by equation (26.1) decompose, §26.8, into 6 scalars, 4 vectors, and 1 tensor, a total of $6 + 4 \times 2 + 1 \times 2 = 16$ degrees of freedom,

$$\varphi_{00} = \underset{\text{scalar}}{\psi} , \quad (27.6a)$$

$$\varphi_{0a} = \underset{\text{scalar}}{\nabla_a w} + \underset{\text{vector}}{w_a} , \quad (27.6b)$$

$$\varphi_{a0} = \underset{\text{scalar}}{\nabla_a \tilde{w}} + \underset{\text{vector}}{\tilde{w}_a} , \quad (27.6c)$$

$$\varphi_{ab} = \underset{\text{scalar}}{\delta_{ab} \Phi} + \underset{\text{scalar}}{\nabla_a \nabla_b h} + \underset{\text{scalar}}{\varepsilon_{abc} \nabla_c \tilde{h}} + \underset{\text{vector}}{\nabla_a h_b} + \underset{\text{vector}}{\nabla_b \tilde{h}_a} + \underset{\text{tensor}}{h_{ab}} . \quad (27.6d)$$

The tildes on $\tilde{w}$ and $\tilde{h}$ simply distinguish those symbols (from w and h); the tildes have no other significance. The vector components are by definition transverse (have vanishing divergence), while the tensor component h_{ab} is by definition traceless, symmetric, and transverse. For a single Fourier mode whose wavevector $\mathbf{k}$ is taken without loss of generality to lie in the z -direction, so that $\nabla_x = \nabla_y = 0$, equations (27.6) are

$$\varphi_{mn} = \begin{pmatrix} \psi & w_x & w_y & \nabla_z w \\ \tilde{w}_x & \Phi + h_{xx} & h_{xy} + \nabla_z \tilde{h} & \nabla_z \tilde{h}_x \\ \tilde{w}_y & h_{xy} - \nabla_z \tilde{h} & \Phi - h_{xx} & \nabla_z \tilde{h}_y \\ \nabla_z \tilde{w} & \nabla_z h_x & \nabla_z h_y & \Phi + \nabla_z^2 h \end{pmatrix} . \quad (27.7)$$

To identify coordinate gauge-invariant quantities, it is necessary to consider infinitesimal coordinate gauge transformations (26.9). The 4 tetrad-frame components ϵ_m of the coordinate shift of the coordinate gauge transformation decompose into 2 scalars and 1 vector

$$\epsilon_m = \left\{ \underset{\text{scalar}}{\epsilon_0} , \quad \underset{\text{scalar}}{\nabla_a \epsilon} + \underset{\text{vector}}{\epsilon_a} \right\} . \quad (27.8)$$

In the flat space background space being considered, the coordinate gauge transformation (26.20) of the vierbein perturbation simplifies to

$$\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \nabla_m \epsilon_n . \quad (27.9)$$

In terms of the scalar, vector, and tensor potentials introduced in equations (27.6), the gauge transformations (27.9) are

$$\varphi_{00} \rightarrow \underset{\text{scalar}}{\psi} + \underset{\text{scalar}}{\dot{\epsilon}_0} , \quad (27.10a)$$

$$\varphi_{0a} \rightarrow \underset{\text{scalar}}{\nabla_a (w + \dot{\epsilon})} + \underset{\text{vector}}{(w_a + \dot{\epsilon}_a)} , \quad (27.10b)$$

$$\varphi_{a0} \rightarrow \underset{\text{scalar}}{\nabla_a (\tilde{w} + \epsilon_0)} + \underset{\text{vector}}{\tilde{w}_a} , \quad (27.10c)$$

$$\varphi_{ab} \rightarrow \underset{\text{scalar}}{\delta_{ab} \Phi} + \underset{\text{scalar}}{\nabla_a \nabla_b (h + \epsilon)} + \underset{\text{scalar}}{\varepsilon_{abc} \nabla_c \tilde{h}} + \underset{\text{vector}}{\nabla_a (h_b + \epsilon_b)} + \underset{\text{vector}}{\nabla_b \tilde{h}_a} + \underset{\text{tensor}}{h_{ab}} . \quad (27.10d)$$

Equations (27.10a) imply that under an infinitesimal coordinate gauge transformation the potentials trans-

form as

$$\psi \rightarrow \psi + \dot{\epsilon}_0 , \quad (27.11a)$$

$$w \rightarrow w + \dot{\epsilon} , \quad w_a \rightarrow w_a + \dot{\epsilon}_a , \quad (27.11b)$$

$$\tilde{w} \rightarrow \tilde{w} + \epsilon_0 , \quad \tilde{w}_a \rightarrow \tilde{w}_a , \quad (27.11c)$$

$$\Phi \rightarrow \Phi , \quad h \rightarrow h + \epsilon , \quad \tilde{h} \rightarrow \tilde{h} , \quad h_a \rightarrow h_a + \epsilon_a , \quad \tilde{h}_a \rightarrow \tilde{h}_a , \quad h_{ab} \rightarrow h_{ab} . \quad (27.11d)$$

Eliminating the coordinate shift ϵ_m from the transformations (27.11) yields 12 coordinate gauge-invariant combinations of the potentials

$$\begin{array}{cccccccc} \psi - \dot{\tilde{w}} , & w - \dot{h} , & w_a - \dot{h}_a , & \tilde{w}_a , & \Phi , & \tilde{h} , & \tilde{h}_a , & h_{ab} . \\ \text{scalar} & \text{scalar} & \text{vector} & \text{vector} & \text{scalar} & \text{scalar} & \text{vector} & \text{tensor} \end{array} \quad (27.12)$$

Physical perturbations are not only coordinate but also tetrad gauge-invariant. A quantity is tetrad gauge-invariant if and only if it depends only on the symmetric part of the vierbein perturbations, not on the antisymmetric part, §26.6. There are 6 combinations of the coordinate gauge-invariant perturbations (27.12) that are symmetric, and therefore not only coordinate but also tetrad gauge-invariant. These 6 coordinate and tetrad gauge-invariant perturbations comprise 2 scalars, 1 vector, and 1 tensor

$$\boxed{\begin{array}{c} \Psi \\ \text{scalar} \end{array}} \equiv \psi - \dot{w} - \dot{\tilde{w}} + \ddot{h} , \quad (27.13a)$$

$$\boxed{\begin{array}{c} \Phi \\ \text{scalar} \end{array}} , \quad (27.13b)$$

$$\boxed{\begin{array}{c} W_a \\ \text{vector} \end{array}} \equiv w_a + \tilde{w}_a - \dot{h}_a - \dot{\tilde{h}}_a , \quad (27.13c)$$

$$\boxed{\begin{array}{c} h_{ab} \\ \text{tensor} \end{array}} . \quad (27.13d)$$

Since only the 6 tetrad and coordinate gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} have physical significance, it is legitimate to choose a particular **gauge**, a set of conditions on the non-gauge-invariant potentials, arranged to simplify the equations, or to bring out some physical aspect. Three gauges considered later are harmonic gauge (§27.7), Newtonian gauge (§27.8), and synchronous gauge (§27.9). However, for the next several sections, no gauge will be chosen: the exposition will continue to be completely general.

Exercise 27.1 Classification of perturbations in arbitrary dimensions. Classify and enumerate general relativistic perturbations in N spacetime dimensions.

Solution. In N spacetime dimensions, there are $N-2$ transverse directions. In N spacetime dimensions, the vierbein perturbations φ_{mn} , equations (27.6), decompose into: 5 scalars ψ , w , $\tilde{w}$, Φ , h ; 4 vectors w_a , $\tilde{w}_a$, h_a , $\tilde{h}_a$; 1 transverse antisymmetric tensor $\tilde{h}_{ab}$ (which for $N = 4$ reduces to a scalar $\varepsilon_{abc}\nabla_c\tilde{h}$); and 1 transverse traceless symmetric tensor h_{ab} ; for a total of $5 + 4(N-2) + \frac{1}{2}(N-2)(N-3) + \frac{1}{2}N(N-3) = N^2$ degrees of freedom. Coordinate transformations, equation (27.8), decompose into 2 scalars ϵ_0 , ϵ , and 1 vector ϵ_a , a total of $2 + (N-2) = N$ degrees of freedom, leaving 3 scalars, 3 vectors, 1 transverse antisymmetric tensor,

and 1 transverse traceless symmetric tensor. Tetrad (Lorentz) transformations remove a further 1 scalar, 2 vectors, and 1 transverse antisymmetric tensor, a total of $1 + 2(N-2) + \frac{1}{2}(N-2)(N-3) = \frac{1}{2}(N-1)(N-2)$ degrees of freedom, leaving as physical degrees of freedom 2 scalars Ψ and Φ , 1 vector W_a , and 1 transverse traceless symmetric tensor h_{ab} , a total of $2 + (N-2) + \frac{1}{2}N(N-3) = \frac{1}{2}(N-1)(N-2)$ degrees of freedom. The transverse traceless symmetric tensor h_{ab} carries propagating gravitational waves, §27.13. Gravitational waves have $\frac{1}{2}N(N-3)$ degrees of freedom, and exist only in spacetime dimensions $N \geq 4$.

27.2 Metric, tetrad connections, and Einstein and Weyl tensors

This section gives expressions in a completely general gauge for perturbed quantities in flat background Minkowski space.

27.2.1 Metric

The unperturbed metric $\overset{0}{g}_{\mu\nu}$ is the Minkowski metric, equation (27.1). The perturbation $\overset{1}{g}_{\mu\nu}$ of the coordinate metric is, from equation (26.6),

$$\overset{1}{g}_{tt} = -2 \underset{\text{scalar}}{\psi} , \quad (27.14a)$$

$$\overset{1}{g}_{ta} = -\underset{\text{scalar}}{\nabla_a}(w + \tilde{w}) - (\underset{\text{vector}}{w_a} + \tilde{w}_a) , \quad (27.14b)$$

$$\overset{1}{g}_{ab} = -\underset{\text{scalar}}{\delta_{ab}} 2\Phi - 2\underset{\text{scalar}}{\nabla_a}\underset{\text{scalar}}{\nabla_b}h - \underset{\text{vector}}{\nabla_a}(\underset{\text{vector}}{h_b} + \tilde{h}_b) - \underset{\text{vector}}{\nabla_b}(\underset{\text{vector}}{h_a} + \tilde{h}_a) - \underset{\text{tensor}}{2h_{ab}} . \quad (27.14c)$$

The coordinate metric is tetrad gauge-invariant, but not coordinate gauge-invariant.

27.2.2 Tetrad-frame connections

The tetrad-frame connections Γ_{kmn} can be calculated from the usual formula (11.54). The unperturbed tetrad connections $\overset{0}{\Gamma}_{kmn}$ all vanish in the flat background. The perturbations $\overset{1}{\Gamma}_{kmn}$ of the tetrad connections are

$$\overset{1}{\Gamma}_{0a0} = -\underset{\text{scalar}}{\nabla_a}(\underset{\text{scalar}}{\psi} - \underset{\text{vector}}{\dot{w}}) + \underset{\text{vector}}{\dot{w}_a} , \quad (27.15a)$$

$$\overset{1}{\Gamma}_{0ab} = \underset{\text{scalar}}{\delta_{ab}}\underset{\text{scalar}}{\dot{\Phi}} - \underset{\text{scalar}}{\nabla_a}\underset{\text{scalar}}{\nabla_b}(w - \dot{h}) - \frac{1}{2}(\underset{\text{vector}}{\nabla_a}W_b + \underset{\text{vector}}{\nabla_b}W_a) + \underset{\text{vector}}{\nabla_b}\tilde{w}_a + \underset{\text{tensor}}{\dot{h}_{ab}} , \quad (27.15b)$$

$$\overset{1}{\Gamma}_{ab0} = \frac{1}{2}(\underset{\text{vector}}{\nabla_a}W_b - \underset{\text{vector}}{\nabla_b}W_a) - \frac{\partial}{\partial t}(\underset{\text{scalar}}{\varepsilon_{abd}}\underset{\text{scalar}}{\nabla_d}\tilde{h} - \underset{\text{vector}}{\nabla_a}\tilde{h}_b + \underset{\text{vector}}{\nabla_b}\tilde{h}_a) , \quad (27.15c)$$

$$\overset{1}{\Gamma}_{abc} = (\underset{\text{scalar}}{\delta_{bc}}\underset{\text{scalar}}{\nabla_a} - \underset{\text{scalar}}{\delta_{ac}}\underset{\text{scalar}}{\nabla_b})\Phi - \underset{\text{scalar}}{\nabla_c}(\underset{\text{scalar}}{\varepsilon_{abd}}\underset{\text{scalar}}{\nabla_d}\tilde{h} - \underset{\text{vector}}{\nabla_a}\tilde{h}_b + \underset{\text{vector}}{\nabla_b}\tilde{h}_a) + \underset{\text{tensor}}{\nabla_a}h_{bc} - \underset{\text{tensor}}{\nabla_b}h_{ac} . \quad (27.15d)$$

The perturbations of the tetrad connections are all coordinate gauge-invariant, as is evident from the fact that they depend only on, and on all 12 of, the coordinate gauge-invariant combinations (27.12). The coordinate

gauge-invariance of the tetrad connections follows more fundamentally from the fact that any quantity that vanishes in the unperturbed background is coordinate gauge-invariant. According to the rule established in §26.7, the change in a quantity under an infinitesimal coordinate gauge transformation equals its Lie derivative $\mathcal{L}_\epsilon$ with respect to the infinitesimal coordinate shift ϵ . Any quantity that vanishes in the unperturbed background has, to linear order, vanishing Lie derivative, therefore is coordinate gauge-invariant.

However, the perturbations $\overset{1}{\Gamma}_{kmn}$ of the tetrad connections are not tetrad gauge-invariant, as is evident from the fact that they (all) depend on antisymmetric parts of the vierbein perturbations φ_{mn} .

27.2.3 Tetrad-frame Einstein tensor

The tetrad-frame Einstein tensor G_{mn} in perturbed Minkowski space follows from the usual formulae (11.62), (11.79), and (11.81). The unperturbed Einstein tensor $\overset{0}{G}_{mn}$ vanishes identically. The perturbations $\overset{1}{G}_{mn}$ of the tetrad-frame Einstein tensor are

$$\boxed{\overset{1}{G}_{00} = \underset{\text{scalar}}{2 \nabla^2 \Phi}}, \quad (27.16a)$$

$$\boxed{\overset{1}{G}_{0a} = \underset{\text{scalar}}{2 \nabla_a \dot{\Phi}} + \underset{\text{vector}}{\frac{1}{2} \nabla^2 W_a}}, \quad (27.16b)$$

$$\boxed{\overset{1}{G}_{ab} = \underset{\text{scalar}}{2 \delta_{ab} \ddot{\Phi}} - \underset{\text{scalar}}{(\nabla_a \nabla_b - \delta_{ab} \nabla^2)(\Psi - \Phi)} + \underset{\text{vector}}{\frac{1}{2} (\nabla_a \dot{W}_b + \nabla_b \dot{W}_a)} + \underset{\text{tensor}}{\square h_{ab}}}, \quad (27.16c)$$

where $\square$ is the d'Alembertian, the 4-dimensional wave operator

$$\square \equiv \nabla_m \nabla^m = -\frac{\partial^2}{\partial t^2} + \nabla^2. \quad (27.17)$$

All the perturbations $\overset{1}{G}_{mn}$ of the Einstein tensor are both coordinate and tetrad gauge-invariant, as follows from the fact that the expressions (27.16) depend only on the coordinate and tetrad gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} . The property that the perturbations of the Einstein tensor are coordinate and tetrad gauge-invariant is a feature of flat (Minkowski) background spacetime, and does not persist to more general spacetimes, such as the Friedmann-Lemaître-Robertson-Walker spacetime.

In a frame with the wavevector $\mathbf{k}$ taken along the z -axis, so that $\nabla_x = \nabla_y = 0$, the perturbations of the Einstein tensor are

$$\overset{1}{G}_{mn} = \begin{pmatrix} 2 \nabla_z^2 \Phi & \frac{1}{2} \nabla_z^2 W_x & \frac{1}{2} \nabla_z^2 W_y & 2 \nabla_z \dot{\Phi} \\ \frac{1}{2} \nabla_z^2 W_x & 2 \ddot{\Phi} + \nabla_z^2 (\Psi - \Phi) + \square h_+ & \square h_\times & \frac{1}{2} \nabla_z \dot{W}_x \\ \frac{1}{2} \nabla_z^2 W_y & \square h_\times & 2 \ddot{\Phi} + \nabla_z^2 (\Psi - \Phi) - \square h_+ & \frac{1}{2} \nabla_z \dot{W}_y \\ 2 \nabla_z \dot{\Phi} & \frac{1}{2} \nabla_z \dot{W}_x & \frac{1}{2} \nabla_z \dot{W}_y & 2 \ddot{\Phi} \end{pmatrix}, \quad (27.18)$$

where h_+ and $h_\times$ are the two linear polarizations of gravitational waves, discussed further in §27.13,

$$h_+ \equiv h_{xx} = -h_{yy}, \quad h_\times \equiv h_{xy} = h_{yx}. \quad (27.19)$$

The tetrad-frame complexified Weyl tensor is

$$\begin{aligned}\tilde{C}_{0a0b} = & \frac{1}{4}(\nabla_a \nabla_b - \frac{1}{3} \underset{\text{scalar}}{\delta_{ab}} \nabla^2)(\Psi + \Phi) \\ & + \frac{1}{8}[-(\nabla_a \dot{W}_b + \nabla_b \dot{W}_a) + i(\varepsilon_{acd} \nabla_b + \varepsilon_{bcd} \nabla_a) \nabla_c W_d] \\ & + \frac{1}{4}[\ddot{h}_{ab} - \varepsilon_{acd} \varepsilon_{bef} \nabla_c \nabla_e h_{df} - i(\varepsilon_{acd} \nabla_c \dot{h}_{bd} + \varepsilon_{bcd} \nabla_c \dot{h}_{ad})] .\end{aligned}\quad (27.20)$$

vector

Like the tetrad-frame Einstein tensor, the tetrad-frame Weyl tensor is both coordinate and tetrad gauge-invariant, depending only on the coordinate and tetrad gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} .

27.3 Spin components of the Einstein tensor

Scalar, vector, and tensor perturbations correspond respectively to perturbations of spin 0, 1, and 2. An object has spin s if it is unchanged by a rotation of $2\pi/s$ about a prescribed direction. In perturbed Minkowski space, the prescribed direction is the direction of the wavevector $\mathbf{k}$ in the Fourier decomposition of the modes. The spin components may be projected out by working in a spin tetrad, §40.1.

In a frame where the wavevector $\mathbf{k}$ is taken along the z -axis, the spin components of the perturbations $\overset{1}{G}_{mn}$ of the Einstein tensor (27.16) are

$$\overset{1}{G}_{00} = 2 \underset{\text{spin-0}}{\nabla_z^2 \Phi} , \quad \overset{1}{G}_{0z} = 2 \underset{\text{spin-0}}{\nabla_z \dot{\Phi}} , \quad \overset{1}{G}_{zz} = 2 \underset{\text{spin-0}}{\ddot{\Phi}} , \quad (27.21a)$$

$$\overset{1}{G}_{+-} - \overset{1}{G}_{zz} = \underset{\text{spin-0}}{\nabla_z^2 (\Psi - \Phi)} , \quad (27.21b)$$

$$\overset{1}{G}_{0\pm} = \frac{1}{2} \underset{\text{spin-}\pm 1}{\nabla_z^2 W_{\pm}} , \quad \overset{1}{G}_{z\pm} = \frac{1}{2} \underset{\text{spin-}\pm 1}{\nabla_z \dot{W}_{\pm}} , \quad (27.21c)$$

$$\overset{1}{G}_{\pm\pm} = \underset{\text{spin-}\pm 2}{\square} h_{\pm\pm} , \quad (27.21d)$$

where $W_{\pm}$ are the spin ± 1 components of the vector perturbation W_a ,

$$W_{\pm} = \frac{1}{\sqrt{2}}(W_x \pm i W_y) , \quad (27.22)$$

and $h_{\pm\pm}$ are the spin ± 2 components of the tensor perturbation h_{ab} ,

$$h_{\pm\pm} = h_{xx} \pm i h_{xy} = h_+ \pm i h_{\times} . \quad (27.23)$$

The spin +2 and -2 components h_{++} and h_{--} of the tensor perturbation are called the right- and left-handed circular polarizations. The spin +2 and -2 circular polarizations h_{++} and h_{--} transform as $e^{-i2\chi}$ and $e^{i2\chi}$ under a right-handed rotation by angle χ about the z -axis, while the linear polarizations h_+ and $h_{\times}$ transform as $\cos 2\chi$ and $-\sin 2\chi$.

27.4 Too many Einstein equations?

The Einstein equations are as usual (units $c = G = 1$; in the remainder of this Chapter, perturbation overscripts ¹ on the Einstein and energy-momentum tensors are dropped for brevity, which is fine because the unperturbed tensors vanish identically in the Minkowski background)

$$G_{mn} = 8\pi T_{mn} . \quad (27.24)$$

There are 10 Einstein equations, but the Einstein tensor (27.16) depends on only 6 independent potentials: the two scalars Ψ and Φ , the vector W_a , and the tensor h_{ab} . The system of Einstein equations is thus overcomplete. Why? The answer is that 4 of the Einstein equations enforce conservation of energy-momentum, and can therefore be considered as governing the evolution of the energy-momentum as opposed to being equations for the gravitational potentials. For example, the form of equations (27.16a) and (27.16b) for G_{00} and G_{0a} enforces conservation of energy

$$D^m G_{m0} = 0 , \quad (27.25)$$

while the form of equations (27.16b) and (27.16c) for G_{0a} and G_{ab} enforces conservation of momentum

$$D^m G_{ma} = 0 . \quad (27.26)$$

Normally, the equations governing the evolution of the energy-momentum T_X^{mn} of each species X of mass-energy would be set up so as to ensure overall conservation of energy-momentum. If this is done, then the conservation equations (27.25) and (27.26) can be regarded as redundant. Since equations (27.25) and (27.26) are equations for the time evolution of G_{00} and G_{0a} , one might think that the Einstein equations for G_{00} and G_{0a} would become redundant, but this is not quite true. In fact the Einstein equations for G_{00} and G_{0a} impose constraints that must be satisfied on the initial spatial hypersurface. Conservation of energy-momentum guarantees that those constraints will continue to be satisfied on subsequent spatial hypersurfaces, but still the initial conditions must be arranged to satisfy the constraints. Because the Einstein equations for G_{00} and G_{0a} must be satisfied as constraints on the initial conditions, but thereafter can be ignored, the equations are called constraint equations. The Einstein equation for G_{00} is called the energy constraint, or Hamiltonian constraint. The Einstein equations for G_{0a} are called the momentum constraints.

27.5 Action at a distance?

The tensor component of the Einstein equations shows that, in a vacuum $T_{mn} = 0$, the tensor perturbations h_{ab} propagate at the speed of light, satisfying the wave equation

$$\square h_{ab} = 0 . \quad (27.27)$$

The tensor perturbations represent propagating gravitational waves.

It is to be expected that scalar and vector perturbations would also propagate at the speed of light, yet this is not obvious from the form of the Einstein tensor (27.16). Specifically, there are 4 components of the

Einstein tensor (27.16) that apparently depend only on spatial derivatives, not on time derivatives. The 4 corresponding Einstein equations are

$$\nabla^2 \Phi = 4\pi T_{00}^{\text{scalar}} , \quad (27.28a)$$

$$\nabla^2 W_a = 16\pi T_{0a}^{\text{vector}} , \quad (27.28b)$$

$$\nabla^2 (\Psi - \Phi) = -8\pi Q_{ab} T_{ab}^{\text{scalar}} , \quad (27.28c)$$

where Q_{ab} in equation (27.28c) is the quadrupole operator defined below, equation (27.102). These conditions must be satisfied everywhere at every instant of time, giving the impression that signals are travelling instantaneously from place to place.

27.6 Comparison to electromagnetism

The previous two sections §27.4 and §27.5 brought up two issues:

1. There are 10 Einstein equations, but only 6 independent gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} . The additional 4 Einstein equations serve to enforce conservation of energy-momentum.
2. Only 2 of the gauge-invariant potentials, the tensor potentials h_{ab} , satisfy causal wave equations. The remaining 4 gauge-invariant potentials Ψ , Φ , and W_a satisfy equations (27.28) that depend on the instantaneous distribution of energy-momentum throughout space, on the face of it violating causality.

These facts may seem surprising, but in fact the equations of electromagnetism have a similar structure, as will now be shown. In this section, the spacetime is assumed for simplicity to be flat Minkowski space. The discussion in this section is based in part on the exposition by Bertschinger (1993).

In accordance with the usual procedure, the electromagnetic field may be defined in terms of an electromagnetic 4-potential A^m , whose time and spatial parts constitute the scalar potential ϕ and the vector potential $\mathbf{A}$:

$$A^m \equiv \{\phi, \mathbf{A}\} . \quad (27.29)$$

In flat (Minkowski) space, the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ are defined in terms of the potentials ϕ and $\mathbf{A}$ by

$$\mathbf{E} \equiv -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t} , \quad (27.30a)$$

$$\mathbf{B} \equiv \nabla \times \mathbf{A} . \quad (27.30b)$$

Given their definition (27.30), the electric and magnetic fields automatically satisfy the two source-free Maxwell's equations

$$\nabla \cdot \mathbf{B} = 0 , \quad (27.31a)$$

$$\nabla \times \mathbf{E} + \frac{\partial\mathbf{B}}{\partial t} = 0 . \quad (27.31b)$$

The remaining two Maxwell's equations, the sourced ones, are

$$\nabla \cdot \mathbf{E} = 4\pi q , \quad (27.32a)$$

$$\nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} = 4\pi \mathbf{j} , \quad (27.32b)$$

where q and $\mathbf{j}$ are the electric charge and current density, the time and space components of the electric 4-current density j^m

$$j^m \equiv \{q, \mathbf{j}\} . \quad (27.33)$$

The electromagnetic potentials ϕ and $\mathbf{A}$ are not unique, but rather are defined only up to a gauge transformation by some arbitrary gauge field θ

$$\phi \rightarrow \phi - \frac{\partial \theta}{\partial t} , \quad \mathbf{A} \rightarrow \mathbf{A} + \nabla \theta . \quad (27.34)$$

The gauge transformation (27.34) evidently leaves the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$, equations (27.30), invariant.

Following the path of previous sections, §27.1 and thereafter, decompose the vector potential $\mathbf{A}$ into its scalar and vector parts

$$\mathbf{A} = \underbrace{\nabla A_{\parallel}}_{\text{scalar}} + \underbrace{\mathbf{A}_{\perp}}_{\text{vector}} , \quad (27.35)$$

in which the vector part by definition satisfies the transversality condition $\nabla \cdot \mathbf{A}_{\perp} = 0$. Under a gauge transformation (27.34), the potentials transform as

$$\phi \rightarrow \phi - \frac{\partial \theta}{\partial t} , \quad (27.36a)$$

$$A_{\parallel} \rightarrow A_{\parallel} + \theta , \quad (27.36b)$$

$$\mathbf{A}_{\perp} \rightarrow \mathbf{A}_{\perp} . \quad (27.36c)$$

Eliminating the gauge field θ yields 3 gauge-invariant potentials, comprising 1 scalar Φ , and 1 vector $\mathbf{A}_{\perp}$ containing 2 degrees of freedom:

$$\boxed{\begin{array}{c} \Phi \\ \text{scalar} \end{array}} \equiv \phi + \frac{\partial A_{\parallel}}{\partial t} , \quad (27.37a)$$

$$\boxed{\begin{array}{c} \mathbf{A}_{\perp} \\ \text{vector} \end{array}} . \quad (27.37b)$$

This shows that the electromagnetic field contains 3 independent degrees of freedom, consisting of 1 scalar and 1 vector.

Concept question 27.2 **Are gauge-invariant potentials Lorentz-invariant?** The potentials Φ and $\mathbf{A}_{\perp}$, equations (27.37), are by construction gauge-invariant, but is this construction Lorentz-invariant? Do Φ and $\mathbf{A}_{\perp}$ constitute the components of a 4-vector? **Answer.** No.

In terms of the gauge-invariant potentials Φ and $\mathbf{A}_\perp$, equations (27.37), the electric and magnetic fields are

$$\mathbf{E} = -\nabla\Phi - \frac{\partial\mathbf{A}_\perp}{\partial t} , \quad (27.38a)$$

$$\mathbf{B} = \nabla \times \mathbf{A}_\perp . \quad (27.38b)$$

The sourced Maxwell's equations (27.32) thus become, in terms of Φ and $\mathbf{A}_\perp$,

$$\underset{\text{scalar}}{-\nabla^2\Phi} = \underset{\text{scalar}}{4\pi q} , \quad (27.39a)$$

$$\underset{\text{scalar}}{\nabla\dot{\Phi}} - \underset{\text{vector}}{\square\mathbf{A}_\perp} = \underset{\text{scalar}}{4\pi\nabla j_\parallel} + \underset{\text{vector}}{4\pi\mathbf{j}_\perp} , \quad (27.39b)$$

where $\nabla j_\parallel$ and $\mathbf{j}_\perp$ are the scalar and vector parts of the current density $\mathbf{j}$. Equations (27.39) bear a striking similarity to the Einstein equations (27.16). Only the vector part $\mathbf{A}_\perp$ satisfies a wave equation,

$$\square\mathbf{A}_\perp = -4\pi\mathbf{j}_\perp , \quad (27.40)$$

while the scalar part Φ satisfies an instantaneous equation (27.39a), $\nabla\dot{\Phi} = 4\pi\nabla j_\parallel$, that seemingly violates causality. And just as Einstein's equations (27.16) enforce conservation of energy-momentum, so also Maxwell's equations (27.39) enforce conservation of electric charge,

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{j} = 0 , \quad (27.41)$$

or in 4-dimensional form

$$\nabla_m j^m = 0 . \quad (27.42)$$

The fact that only the vector part $\mathbf{A}_\perp$ satisfies a wave equation (27.40) reflects physically the fact that electromagnetic waves are transverse, and they contain only two propagating degrees of freedom, the vector, or spin ± 1 , components.

Why do Maxwell's equations (27.39) have this structure? Although equation (27.40) appears to be a local wave equation for the vector part $\mathbf{A}_\perp$ of the potential sourced by the vector part $\mathbf{j}_\perp$ of the current, in fact the wave equation is non-local because the decomposition of the potential and current into scalar and vector parts is non-local (it involves the solution of a Laplacian equation, eq. (26.23)). It is only the sum $\mathbf{j} = \nabla j_\parallel + \mathbf{j}_\perp$ of the scalar and vector parts of the current density that is local. Therefore, the Maxwell's equation (27.39b) must have a scalar part to go along with the vector part, such that the source on the right hand side, the current density $\mathbf{j}$, is local. Given this Maxwell equation (27.39b), the Maxwell equation (27.39a) then serves precisely to enforce conservation of electric charge, equation (27.41).

Just as it is possible to regard the Einstein equations (27.16a) and (27.16b) as constraint equations whose continued satisfaction is guaranteed by conservation of energy-momentum, so also the Maxwell equation (27.39a) for Φ can be regarded as a constraint equation whose continued satisfaction is guaranteed

by conservation of electric charge. For charge conservation (27.41) coupled with the spatial Maxwell equation (27.39b) ensures that

$$\frac{\partial}{\partial t} (4\pi q + \nabla^2 \Phi) = 0 , \quad (27.43)$$

the solution of which, subject to the condition that $4\pi q + \nabla^2 \Phi = 0$ initially, is $4\pi q + \nabla^2 \Phi = 0$ at all times, which is precisely the Maxwell equation (27.39a).

In a system of charges and electromagnetic fields, equations of motion for the charges in the electromagnetic field must be adjoined to the (Maxwell) equations of motion for the electromagnetic field. If the equations of motion for the charges are arranged to conserve charge, as they should, then the scalar Maxwell equation (27.39a) determines the scalar potential Φ on the initial hypersurface of constant time, but can be discarded thereafter as redundant.

Concept question 27.3 **What parts of Maxwell's equations can be discarded?** Is it possible to discard the scalar part of the spatial Maxwell equation (27.39b), rather than the scalar equation (27.39a) for Φ ? Project out the scalar part of equation (27.39b) by taking its divergence,

$$\nabla^2 (4\pi j_{\parallel} - \dot{\Phi}) = 0 . \quad (27.44)$$

Argue that the Maxwell equation (27.39a), coupled with charge conservation (27.41), ensures that equation (27.44) is true, subject to boundary condition that the current $\mathbf{j}$ vanish sufficiently rapidly at spatial infinity, in accordance with the decomposition theorem of §26.8.1.

Since only gauge-invariant quantities have physical significance, it is legitimate to impose any condition on the gauge field θ . A gauge in which the potentials ϕ and $\mathbf{A}$ individually satisfy wave equations is **Lorenz** (not Lorentz!) **gauge**, which consists of the Lorentz-invariant condition

$$\nabla_m A^m = 0 . \quad (27.45)$$

Under a gauge transformation (27.34), the left hand side of equation (27.45) transforms as

$$\nabla_m A^m \rightarrow \nabla_m A^m + \square \theta , \quad (27.46)$$

and the Lorenz gauge condition (27.45) can be accomplished as a particular solution of the wave equation for the gauge field θ . In terms of the potentials ϕ and $A_{\parallel}$, the Lorenz gauge condition (27.45) is

$$\frac{\partial \phi}{\partial t} + \nabla^2 A_{\parallel} = 0 . \quad (27.47)$$

In Lorenz gauge, Maxwell's equations (27.39) become

$$\square \phi = -4\pi q , \quad (27.48a)$$

$$\square \mathbf{A} = -4\pi \mathbf{j} , \quad (27.48b)$$

which are manifestly wave equations for the potentials ϕ and $\mathbf{A}$.

Does the fact that the potentials ϕ and $\mathbf{A}$ in one particular gauge, Lorenz gauge, satisfy wave equations necessarily guarantee that the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ satisfy wave equations? Yes, because it follows from the definitions (27.30) of $\mathbf{E}$ and $\mathbf{B}$ that if the potentials ϕ and $\mathbf{A}$ satisfy wave equations, then so also must the fields $\mathbf{E}$ and $\mathbf{B}$ themselves; but the fields $\mathbf{E}$ and $\mathbf{B}$ are gauge-invariant, so if they satisfy wave equations in one gauge, then they must satisfy the same wave equations in any gauge.

In electromagnetism, the most physical choice of gauge is one in which the potentials ϕ and $\mathbf{A}$ coincide with the gauge-invariant potentials Φ and $\mathbf{A}_\perp$, equations (27.37). This gauge, known as **Coulomb gauge**, is accomplished by setting

$$A_\parallel = 0 , \quad (27.49)$$

or equivalently

$$\nabla \cdot \mathbf{A} = 0 . \quad (27.50)$$

The gravitational analogue of this gauge is the Newtonian gauge discussed in the next section but one, §27.8.

Does the fact that in Lorenz gauge the potentials ϕ and $\mathbf{A}$ propagate at the speed of light (in the absence of sources, $j^m = 0$) imply that the gauge-invariant potentials Φ and $\mathbf{A}_\perp$ propagate at the speed of light? No. The gauge-invariant potentials Φ and $\mathbf{A}_\perp$, equations (27.37), are related to the Lorenz gauge potentials ϕ and $\mathbf{A}$ by a non-local decomposition.

27.7 Harmonic gauge

The fact that all locally measurable gravitational perturbations do propagate causally, at the speed of light in the absence of sources, can be demonstrated by choosing a particular gauge, **harmonic gauge**, equation (27.51), which can be considered an analogue of the Lorenz gauge of electromagnetism, equation (27.45). In harmonic gauge, all 10 of the tetrad gauge-variant (i.e. symmetric) combinations $\varphi_{mn} + \varphi_{nm}$ of the vierbein perturbations satisfy wave equations (27.56), and therefore propagate causally. This does not imply that the scalar, vector, and tensor components of the vierbein perturbations individually propagate causally, because the decomposition into scalar, vector, and tensor modes is non-local. In particular, of the coordinate and tetrad-gauge invariant potentials Ψ , Φ , W_a , and h_{ab} defined by equations (27.13), only the tensor potential h_{ab} propagates causally. The situation is entirely analogous to that of electromagnetism, §27.6, where in Lorenz gauge the potentials ϕ and $\mathbf{A}$ propagate causally, equations (27.48), yet of the gauge-invariant potentials Φ and $\mathbf{A}_\perp$ defined by equations (27.37), only the vector potential $\mathbf{A}_\perp$ propagates causally.

Harmonic gauge is the set of 4 coordinate conditions

$$\nabla^m (\varphi_{mn} + \varphi_{nm} - \eta_{mn} \varphi_k^k) = 0 , \quad (27.51)$$

equivalent to the vanishing of Fock's (1957) harmonic function (17.213). The conditions (27.51) are arranged in a form that is tetrad gauge-invariant (the conditions depend only on the symmetric part of φ_{mn}). The quantities on the left hand side of equations (27.51) transform under a coordinate gauge transformation, in

accordance with (27.9), as

$$\nabla^m(\varphi_{mn} + \varphi_{nm} - \eta_{mn}\varphi_k^k) \rightarrow \nabla^m(\varphi_{mn} + \varphi_{nm} - \eta_{mn}\varphi_k^k) - \square\epsilon_n . \quad (27.52)$$

The change $\square\epsilon_n$ resulting from the coordinate gauge transformation is the 4-dimensional wave operator $\square$ acting on the coordinate shift ϵ_n . Indeed, the harmonic gauge conditions (27.51) follow uniquely from the requirements (a) that the change produced by a coordinate gauge transformation be $\square\epsilon_n$, as suggested by the analogous electromagnetic transformation (27.46), and (b) that the conditions be tetrad gauge-invariant. The harmonic gauge conditions (27.51) can be accomplished as a particular solution of the wave equation for the coordinate shift ϵ_n . In terms of the potentials defined by equations (27.6) and (27.13), the 4 harmonic gauge conditions (27.51) are

$$\dot{\Psi} + 3\dot{\Phi} - \square(w + \tilde{w} - \dot{h}) = 0 , \quad (27.53a)$$

$$\dot{W}_a - \square(h_a + \tilde{h}_a) = 0 , \quad (27.53b)$$

$$-\Psi + \Phi - \square h = 0 , \quad (27.53c)$$

or equivalently

$$\square(w + \tilde{w}) = 4\dot{\Phi} , \quad (27.54a)$$

$$\square(h_a + \tilde{h}_a) = \dot{W}_a , \quad (27.54b)$$

$$\square h = -\Psi + \Phi . \quad (27.54c)$$

Substituting equations (27.54) into the Einstein tensor G_{mn} , equation (27.16), leads, after some calculation, to the result that in harmonic gauge,

$$\frac{1}{2} \square (\varphi_{mn} + \varphi_{nm} - \eta_{mn}\varphi_k^k) = G_{mn} , \quad (27.55)$$

or equivalently

$$\boxed{\frac{1}{2} \square (\varphi_{mn} + \varphi_{nm}) = R_{mn}} , \quad (27.56)$$

where R_{mn} is the Ricci tensor. Equation (27.56) shows that in harmonic gauge, all tetrad gauge-invariant (i.e. symmetric) combinations $\varphi_{mn} + \varphi_{nm}$ of the vierbein potentials propagate causally, at the speed of light in vacuo, $R_{mn} = 0$. Although the result (27.56) is true only in a particular gauge, harmonic gauge, it follows that all quantities that are (coordinate and tetrad) gauge-invariant, and that can be constructed from the vierbein potentials φ_{mn} and their derivatives (and are therefore local), must also propagate at the speed of light.

The 4 coordinate gauge conditions (27.51) still leave 6 tetrad gauge conditions to be chosen at will. A natural choice, in the sense that it leads to the greatest simplification of the tetrad connections Γ_{kmn} , equations (27.15), is the 6 tetrad gauge conditions

$$\tilde{w} = \tilde{w}_a = \tilde{h} = \tilde{h}_a = 0 . \quad (27.57)$$

Exercise 27.4 Einstein tensor in harmonic gauge. Confirm equation (27.56).

27.8 Newtonian (Copernican) gauge

If the unperturbed background is Minkowski space, then the most physical gauge is one in which the 6 perturbations retained coincide with the 6 coordinate and tetrad gauge-invariant perturbations (27.13). This gauge is called **Newtonian gauge**. Because in Newtonian gauge the perturbations are precisely the physical perturbations, if the perturbations are physically weak (small), then the perturbations in Newtonian gauge will necessarily be small.

I think Newtonian gauge should be called **Copernican gauge**. Even though the solar system is a highly non-linear system, from the perspective of general relativity it is a weakly perturbed gravitating system. Applied to the solar system, Newtonian gauge effectively keeps the coordinates aligned with the classical Sun-centred Copernican coordinate frame. By contrast, the coordinates of synchronous gauge (§27.9), which are chosen to follow freely-falling bodies, would quickly collapse or get wound up by orbital motions if applied to the solar system, and would cease to provide a useful description.

Newtonian (Copernican) gauge sets

$$w = \tilde{w} = \tilde{w}_a = h = \tilde{h} = h_a = \tilde{h}_a = 0 , \quad (27.58)$$

so that the retained perturbations are the 6 coordinate and tetrad gauge-invariant perturbations (27.13)

$$\begin{array}{l} \Psi \\ \text{scalar} \end{array} = \psi , \quad (27.59a)$$

$$\begin{array}{l} \Phi \\ \text{scalar} \end{array} , \quad (27.59b)$$

$$\begin{array}{l} W_a \\ \text{vector} \end{array} = w_a , \quad (27.59c)$$

$$\begin{array}{l} h_{ab} \\ \text{tensor} \end{array} . \quad (27.59d)$$

In matrix form, the vierbein perturbation in Newtonian gauge, in a frame where the wavevector $\mathbf{k}$ is along the z -direction, are, from equation (27.7),

$$\varphi_{mn} = \begin{pmatrix} \Psi & W_x & W_y & 0 \\ 0 & \Phi + h_{xx} & h_{xy} & 0 \\ 0 & h_{xy} & \Phi - h_{xx} & 0 \\ 0 & 0 & 0 & \Phi \end{pmatrix} . \quad (27.60)$$

The Newtonian line-element is, in a form that keeps the Newtonian tetrad manifest,

$$ds^2 = -[(1 + \Psi) dt]^2 + \delta_{ab}[(1 - \Phi)dx^a - h_c^a dx^c - W^a dt][(1 - \Phi)dx^b - h_d^b dx^d - W^b dt] , \quad (27.61)$$

which reduces to the Newtonian metric

$$ds^2 = -(1 + 2\Psi) dt^2 - 2W_a dt dx^a + [\delta_{ab}(1 - 2\Phi) - 2h_{ab}] dx^a dx^b . \quad (27.62)$$

Since scalar, vector, and tensor perturbations evolve independently, it is legitimate to consider each in isolation. For example, if one is interested only in scalar perturbations, then it is fine to keep only the scalar potentials Ψ and Φ non-zero. Furthermore, as discussed in §27.12, since the difference $\Psi - \Phi$ in scalar potentials is sourced by anisotropic relativistic pressure, which is typically small, it is often a good approximation to set $\Psi = \Phi$.

The tetrad-frame 4-velocity of a person at rest in the tetrad frame is by definition $u^m = \{1, 0, 0, 0\}$, and the corresponding coordinate 4-velocity u^μ is, in Newtonian gauge,

$$u^\mu = e_0^\mu = \{1 - \Psi, W_a\} . \quad (27.63)$$

This shows that W_a can be interpreted as a 3-velocity at which the tetrad frame is moving through the coordinates. This is the “dragging of inertial frames” discussed in §27.11. The proper acceleration experienced by a person at rest in the tetrad frame, with tetrad 4-velocity $u^m = \{1, 0, 0, 0\}$, is

$$\frac{Du^a}{D\tau} = u^0 D_0 u^a = u^0 (\partial_0 u^a + \Gamma_{00}^a u^0) = \Gamma_{00}^a = \nabla_a \Psi . \quad (27.64)$$

This shows that the “gravity,” or minus the proper acceleration, experienced by a person at rest in the tetrad frame is minus the gradient of the potential Ψ .

Concept question 27.5 Independent evolution of scalar, vector, and tensor modes. If the decomposition into scalar, vector, and tensor modes is non-local, how can it be legitimate to consider the evolution of the modes in isolation from each other?

27.9 Synchronous gauge

One of the earliest gauges used in general relativistic perturbation theory, and still (in its conformal version) widely used in cosmology, is **synchronous gauge**. As will be seen below, equations (27.71) and (27.72), synchronous gauge effectively chooses a coordinate system and tetrad that is attached to the locally inertial frames of freely falling observers. This is fine as long as the observers move only slightly from their initial positions, but the coordinate system will fail when the system evolves too far, even if, as in the solar system, the gravitational perturbations remain weak and therefore treatable in principle with perturbation theory.

Synchronous gauge sets the time components φ_{mn} with $m = 0$ or $n = 0$ of the vierbein perturbations to zero

$$\psi = w = \tilde{w} = w_a = \tilde{w}_a = 0 , \quad (27.65)$$

and makes the additional tetrad gauge choices

$$\tilde{h} = \tilde{h}_a = 0 , \quad (27.66)$$

with the result that the retained perturbations are the spatial perturbations

$$\begin{array}{c} \Phi \\ \text{scalar} \end{array}, \quad \begin{array}{c} h \\ \text{scalar} \end{array}, \quad \begin{array}{c} h_a \\ \text{vector} \end{array}, \quad \begin{array}{c} h_{ab} \\ \text{tensor} \end{array}. \quad (27.67)$$

In terms of these spatial perturbations, the gauge-invariant perturbations (27.13) are

$$\begin{array}{c} \Psi \\ \text{scalar} \end{array} = \ddot{h}, \quad (27.68a)$$

$$\begin{array}{c} \Phi \\ \text{scalar} \end{array}, \quad (27.68b)$$

$$\begin{array}{c} W_a \\ \text{vector} \end{array} = -\dot{h}_a, \quad (27.68c)$$

$$\begin{array}{c} h_{ab} \\ \text{tensor} \end{array}. \quad (27.68d)$$

The synchronous line-element is, in a form that keeps the synchronous tetrad manifest,

$$ds^2 = -dt^2 + \delta_{ab} [(1 - \Phi)dx^a - (\nabla_c \nabla^a h + \nabla_c h^a + h_c^a)dx^c] [(1 - \Phi)dx^b - (\nabla_d \nabla^b h + \nabla_d h^b + h_d^b)dx^d], \quad (27.69)$$

which reduces to the synchronous metric

$$ds^2 = -dt^2 + [(1 - 2\Phi)\delta_{ab} - 2\nabla_a \nabla_b h - \nabla_a h_b - \nabla_b h_a - 2h_{ab}] dx^a dx^b. \quad (27.70)$$

In synchronous gauge, a person at rest in the tetrad frame has coordinate 4-velocity

$$u^\mu = e_0^\mu = \{1, 0, 0, 0\}, \quad (27.71)$$

so that the tetrad rest frame coincides with the coordinate rest frame, and proper time in the rest frame coincides with coordinate time, $\tau = t$. Moreover a person at rest in the tetrad frame is freely falling, which follows from the fact that the acceleration experienced by a person at rest in the tetrad frame is zero,

$$\frac{Du^a}{D\tau} = u^0 (\partial_0 u^a + \Gamma_{00}^a u^0) = \Gamma_{00}^a = 0, \quad (27.72)$$

in which $\partial_0 u^a = 0$ because the 4-velocity at rest in the tetrad frame is constant, $u^a = \{1, 0, 0, 0\}$, and $\Gamma_{00}^a = 0$ from equations (27.15a) with the synchronous gauge choices (27.65) and (27.66). However, the freely falling person's locally inertial frame is rotated relative to the tetrad frame. The cumulative rotation is described by a rotor $R = e^{-\theta/2}$ generated by a bivector $\theta \equiv \frac{1}{2} \theta_{ab} \gamma^a \wedge \gamma^b$ (the factor of $\frac{1}{2}$ would disappear if the sum were over distinct pairs ab of antisymmetric indices) that is the integral of the tetrad connection $\Gamma_0 \equiv \frac{1}{2} \Gamma_{ab0} \gamma^a \wedge \gamma^b$ over time, as follows from $\partial_0 \mathbf{a} = -\frac{1}{2} [\Gamma_0, \mathbf{a}]$ for any multivector $\mathbf{a}$, equation (15.15). From equations (27.15c) and (27.68c) for Γ_{ab0} , the bivector θ_{ab} is

$$\theta_{ab} = \int \Gamma_{ab0} d\tau = \frac{1}{2} (\nabla_a h_b - \nabla_b h_a), \quad (27.73)$$

which is the curl of the vector potential h_a .

27.10 Newtonian potential

The next few sections examine the physical meaning of each of the gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} by looking at the potentials at large distances produced by a finite body containing energy-momentum, such as the Sun.

Einstein's equations $G_{mn} = 8\pi T_{mn}$ applied to the time-time component G_{00} of the Einstein tensor, equation (27.16a), imply Poisson's equation

$$\nabla^2 \Phi = 4\pi \rho , \quad (27.74)$$

where ρ is the mass-energy density

$$\rho \equiv T_{00} . \quad (27.75)$$

The solution of Poisson's equation (27.74) is

$$\Phi(\mathbf{x}) = - \int \frac{\rho(\mathbf{x}') d^3 x'}{|\mathbf{x}' - \mathbf{x}|} . \quad (27.76)$$

Consider a finite body, for example the Sun, whose energy-momentum is confined within a certain region. Define the mass M of the body to be the integral of the mass-energy density ρ ,

$$M \equiv \int \rho(\mathbf{x}') d^3 x' . \quad (27.77)$$

Equation (27.77) agrees with what the definition of the mass M would be in the nonrelativistic limit, and as seen below, equation (27.80), it is what a distant observer would infer the mass of the body to be based on its gravitational potential Φ far away. Thus equation (27.77) can be taken as the definition of the mass of the body even when the energy-momentum is relativistic. Choose the origin of the coordinates to be at the centre of mass, meaning that

$$\int \mathbf{x}' \rho(\mathbf{x}') d^3 x' = 0 . \quad (27.78)$$

Consider the potential Φ at a point $\mathbf{x}$ far outside the body. Expand the denominator of the integral on the right hand side of equation (27.76) as a Taylor series in $1/x$ where $x \equiv |\mathbf{x}|$

$$\frac{1}{|\mathbf{x}' - \mathbf{x}|} = \frac{1}{x} \sum_{\ell=0}^{\infty} \left(\frac{x'}{x} \right)^{\ell} P_{\ell}(\hat{\mathbf{x}} \cdot \hat{\mathbf{x}}') = \frac{1}{x} + \frac{\hat{\mathbf{x}} \cdot \mathbf{x}'}{x^2} + \dots \quad (27.79)$$

where $P_{\ell}(\mu)$ are Legendre polynomials. Then

$$\begin{aligned} \Phi(\mathbf{x}) &= - \frac{1}{x} \int \rho(\mathbf{x}') d^3 x' - \frac{1}{x^2} \hat{\mathbf{x}} \cdot \int \mathbf{x}' \rho(\mathbf{x}') d^3 x' - O(x^{-3}) \\ &= - \frac{M}{x} - O(x^{-3}) . \end{aligned} \quad (27.80)$$

Equation (27.80) shows that the potential far from a body goes as $\Phi = -M/x$, reproducing the usual Newtonian formula.

27.11 Dragging of inertial frames

In Newtonian gauge, the vector potential $\mathbf{W} \equiv W_a$ is the velocity at which the locally inertial tetrad frame moves through the coordinates, equation (29.50). This is called the dragging of inertial frames. As shown below, a body of angular momentum $\mathbf{L}$ drags frames around it with an angular velocity that goes to $2\mathbf{L}/x^3$ at large distances x .

Einstein's equations applied to the vector part of the time-space component G_{0a} of the Einstein tensor, equation (27.16b), imply

$$\nabla^2 \mathbf{W} = -16\pi \mathbf{f} , \quad (27.81)$$

where $\mathbf{W} \equiv W_a$ is the gauge-invariant vector potential, and $\mathbf{f}$ is the vector part of the energy flux T^{0a}

$$\mathbf{f} \equiv f_a = f^a \equiv \underset{\text{vector}}{T^{0a}} = -\underset{\text{vector}}{T_{0a}} . \quad (27.82)$$

The solution of equation (27.81) is

$$\mathbf{W}(\mathbf{x}) = 4 \int \frac{\mathbf{f}(\mathbf{x}') d^3 x'}{|\mathbf{x}' - \mathbf{x}|} . \quad (27.83)$$

As in the previous section, §27.10, consider a finite body, such as the Sun, whose energy-momentum is confined within a certain region. Work in the rest frame of the body, defined to be the frame where the energy flux $\mathbf{f}$ integrated over the body is zero,

$$\int \mathbf{f}(\mathbf{x}') d^3 x' = 0 . \quad (27.84)$$

Define the angular momentum $\mathbf{L}$ of the body to be

$$\mathbf{L} \equiv \int \mathbf{x}' \times \mathbf{f}(\mathbf{x}') d^3 x' . \quad (27.85)$$

Equation (27.85) agrees with what the definition of angular momentum would be in the nonrelativistic limit, where the mass-energy flux of a mass density ρ moving at velocity $\mathbf{v}$ is $\mathbf{f} = \rho \mathbf{v}$. As will be seen below, the angular momentum (27.85) is what a distant observer would infer the angular momentum of the body to be based on the potential $\mathbf{W}$ far away, and equation (27.85) can be taken to be the definition of the angular momentum of the body even when the energy-momentum is relativistic. As will be proven momentarily, equation (27.86), the integral $\int x'_a f_b(\mathbf{x}') d^3 x'$ is antisymmetric in ab . To show this, write $f_b = \varepsilon_{bcd} \nabla'_c \phi_d$ for some potential ϕ_d , which is valid because f_b is the vector (curl) part of the energy flux. Then

$$\int x'_a f_b(\mathbf{x}') d^3 x' = \int x'_a \varepsilon_{bcd} \nabla'_c \phi_d(\mathbf{x}') d^3 x' = - \int \varepsilon_{bcd} \phi_d(\mathbf{x}') \nabla'_c x'_a d^3 x' = \int \varepsilon_{abd} \phi_d(\mathbf{x}') d^3 x' , \quad (27.86)$$

where the third expression follows from the second by integration by parts, the surface term vanishing because of the assumption that the energy-momentum of the body is confined within a certain region.

Taylor expanding equation (27.83) using equation (27.79) gives

$$\begin{aligned}
 \mathbf{W}(\mathbf{x}) &= \frac{4}{x} \int \mathbf{f}(\mathbf{x}') d^3x' + \frac{4}{x^2} \int (\hat{\mathbf{x}} \cdot \mathbf{x}') \mathbf{f}(\mathbf{x}') d^3x + O(x^{-3}) \\
 &= \frac{2}{x^2} \int [(\hat{\mathbf{x}} \cdot \mathbf{x}') \mathbf{f}(\mathbf{x}') - (\hat{\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}')) \mathbf{x}'] d^3x + O(x^{-3}) \\
 &= \frac{2}{x^2} \mathbf{L} \times \hat{\mathbf{x}} + O(x^{-3}) ,
 \end{aligned} \tag{27.87}$$

where the first integral on the right hand side of the first line of equation (27.87) vanishes because the frame is the rest frame of the body, equation (27.84), and the second integral on the right hand side of the first line equals the first integral on the second line thanks to the antisymmetry of $\int \mathbf{x}' \mathbf{f}(\mathbf{x}') d^3x$, equation (27.86). The vector potential $\mathbf{W} \equiv W_a$ points in the direction of rotation, right-handedly about the axis of angular momentum $\mathbf{L}$. Equation (27.87) says that a body of angular momentum $\mathbf{L}$ drags frames around it at angular velocity $\boldsymbol{\Omega}$ at large distances x

$$\mathbf{W} = \boldsymbol{\Omega} \times \mathbf{x} , \quad \boldsymbol{\Omega} = \frac{2\mathbf{L}}{x^3} . \tag{27.88}$$

Exercise 27.6 Gravity Probe B and the geodetic and frame-dragging precession of gyroscopes.

The purpose of Gravity Probe B was to measure the predicted general relativistic precession of a gyroscope in the gravitational field of the Earth. Consider a gyroscope that is in free fall in a spacecraft in orbit around the Earth. In the gyro rest frame, the spin 4-vector σ^m of the gyro has only spatial components

$$\sigma^m = \{0, \sigma^a\} . \tag{27.89}$$

If the gyroscope is moving at 4-velocity u^m relative to the tetrad (Earth) frame, then the components s^m of the spin vector in the tetrad frame are related to those σ^m in the gyro frame by a Lorentz boost at 4-velocity $-u^m$ (early alphabet indices $a, b, \dots$ signify spatial components):

$$\{s^0, s^a\} = \left\{ \sigma_b u^b, \sigma^a + \frac{\sigma_b u^b u^a}{1 + u^0} \right\} . \tag{27.90}$$

Conversely, the components σ^m of the spin vector in the gyro frame are related to those s^m in the tetrad frame by

$$\sigma^a = s^a - \frac{s^0 u^a}{1 + u^0} . \tag{27.91}$$

The gyro is in free-fall in orbit about the earth, so its 4-velocity u^m and 4-spin s^m satisfy the geodesic equations of motion

$$\frac{du^k}{d\tau} + \Gamma_{mn}^k u^m u^n = 0 , \quad \frac{ds^k}{d\tau} + \Gamma_{mn}^k s^m u^n = 0 . \tag{27.92}$$

1. **Spin equation.** Show that

$$\begin{aligned} \frac{d\sigma^a}{d\tau} &= \sigma^b u^n \left[-\Gamma_{abn} + \frac{1}{1+u^0} (u^b \Gamma_{0an} - u^a \Gamma_{0bn}) \right] \\ &= \sigma^b \left[-u^c \Gamma_{abc} - u^0 \Gamma_{ab0} + \frac{u^c}{1+u^0} (u^b \Gamma_{0ac} - u^a \Gamma_{0bc}) + \frac{u^0}{1+u^0} (u^b \Gamma_{0a0} - u^a \Gamma_{0b0}) \right] . \end{aligned} \quad (27.93)$$

[Hint: The first step is to convert σ^a to s^k and u^k , using equation (27.91). Then apply the geodesic equations (27.92). Then convert s^k back to σ^a using equation (27.90).]

2. **Spin precession.** Gravitational fields in the solar system are weak, so perturbation theory in Minkowski space is valid. The tetrad connections Γ_{kmn} in Newtonian gauge are, from equations (27.15),

$$\Gamma_{0a0} = -\nabla_a \Psi , \quad (27.94a)$$

$$\Gamma_{0ab} = \delta_{ab} \dot{\Phi} - \frac{1}{2} (\nabla_a W_b + \nabla_b W_a) + \dot{h}_{ab} , \quad (27.94b)$$

$$\Gamma_{ab0} = \frac{1}{2} (\nabla_a W_b - \nabla_b W_a) , \quad (27.94c)$$

$$\Gamma_{abc} = (\delta_{bc} \nabla_a - \delta_{ac} \nabla_b) \Phi + \nabla_a h_{bc} - \nabla_b h_{ac} . \quad (27.94d)$$

Show from equation (27.93) that the spin $\sigma \equiv \sigma^a$ of a freely-falling gyroscope moving at 3-velocity $\mathbf{v} \equiv \mathbf{u}/u^0$ in a weak gravitational field evolves as (the proper time derivative $d/d\tau$ in equation (27.93) can be converted to the coordinate time derivative d/dt by dividing by $u^0 = dt/d\tau$)

$$\frac{d\sigma}{dt} = \sigma \times \left\{ \mathbf{v} \times \left[\nabla \Phi + \frac{u^0}{1+u^0} \nabla \Psi + \frac{u^c}{1+u^0} \left(\frac{1}{2} (\nabla W_c + \nabla_c \mathbf{W}) - \nabla \times \dot{\mathbf{h}}_c \right) \right] - \frac{1}{2} \nabla \times \mathbf{W} - v^c \nabla \times \mathbf{h}_c \right\} . \quad (27.95)$$

where the vector of vectors $\mathbf{h}_c$ is shorthand for the tensor potential, $\mathbf{h}_c \equiv h_{ac}$. Conclude that at nonrelativistic velocities, $|\mathbf{u}| \ll u^0 \approx 1$, and for $\Psi = \Phi$, equation (27.95) reduces to

$$\frac{d\sigma}{dt} = \sigma \times \left(\frac{3}{2} \mathbf{v} \times \nabla \Phi - \frac{1}{2} \nabla \times \mathbf{W} \right) . \quad (27.96)$$

By comparing your equation (27.96) to the equation of motion of a 3-vector rotating at angular velocity $\boldsymbol{\omega}$,

$$\frac{d\sigma}{dt} = \boldsymbol{\omega} \times \sigma , \quad (27.97)$$

deduce the angular velocity $\boldsymbol{\omega}$ with which the spin $\mathbf{s}$ precesses. The term depending on Φ is the geodetic, or de Sitter (de Sitter, 1916), precession, while the term depending on $\mathbf{W}$ is the frame-dragging, or Lense-Thirring (Thirring, 1918; Lense and Thirring, 1918), precession. [Hint: Recall the 3-vector formula $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$. If the object is nonrelativistic, then $|\mathbf{u}| \ll u^0 \approx 1$.]

3. **Angular velocities.** A body of mass M and angular momentum $\mathbf{L}$ produces scalar and vector perturbations Φ and $\mathbf{W}$ at spatial position $\mathbf{x}$ of, equations (27.80) and (27.87),

$$\Phi(\mathbf{x}) = -\frac{M}{x} , \quad \mathbf{W}(\mathbf{x}) = \frac{2}{x^2} \mathbf{L} \times \hat{\mathbf{x}} . \quad (27.98)$$

Show that for a circular orbit right-handed about direction $\mathbf{n}$, so that $\mathbf{v} = v(\mathbf{n} \times \hat{\mathbf{x}})$, the geodetic/de Sitter precession is, with units restored,

$$\boldsymbol{\omega}_{\text{dS}} = \frac{3(GM)^{3/2}}{2c^2 x^{5/2}} \mathbf{n} , \quad (27.99)$$

while the frame-dragging/Lense-Thirring precession is

$$\boldsymbol{\omega}_{\text{LT}} = \frac{G}{c^2 x^3} [-\mathbf{L} + 3\hat{\mathbf{x}}(\hat{\mathbf{x}} \cdot \mathbf{L})] . \quad (27.100)$$

[Hint: You will need to use the relation between velocity v and potential Φ in a circular orbit.]

4. **Orbit.** What is the orbit-averaged angular velocity for frame-dragging precession in the cases of (i) an equatorial circular orbit, (ii) a polar circular orbit? Compare the directions of the geodetic and frame-dragging precessions in the two cases. Gravity Probe B occupied a polar orbit. Why was that a good strategy?
5. **Gravity Probe B.** Estimate the angular velocity of the geodetic and frame-dragging precessions for Gravity Probe B. Express your answer in arcseconds per year. [Hint: The GPB fact sheet at https://einstein.stanford.edu/content/fact_sheet/GPB_FactSheet-0405.pdf gives the semi-major axis of GPB's orbit as 7027.4 km. The IAU 2009 system of astronomical constants (Luzum et al., 2009) gives $GM = 3.9860044 \times 10^{14} \text{ m}^3 \text{ s}^{-2}$ for the Earth. The Earth fact sheet at <https://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html> gives needed information about the Earth, including its moment of inertia.]
6. **Quadrupole precession.** There is also a purely Newtonian precession that is produced by plain old Newtonian gravity on an object with a quadrupole moment. If you wanted to test the geodetic and frame-dragging effects with a gyroscope in orbit around the Earth, what would you do to avoid contamination by Newtonian quadrupole precession?

27.12 Quadrupole pressure

Einstein's equations applied to the part of the Einstein tensor (27.16c) involving $\Psi - \Phi$ imply

$$\nabla^2(\Psi - \Phi) = -8\pi Q_{ab} T_{ab} , \quad (27.101)$$

where Q_{ab} is the quadrupole operator (an integro-differential operator) defined by

$$Q_{ab} \equiv \frac{3}{2} \nabla_a \nabla_b \nabla^{-2} - \frac{1}{2} \delta_{ab} , \quad (27.102)$$

with ∇^{-2} the inverse spatial Laplacian operator. In Fourier space, the quadrupole operator is

$$Q_{ab} = \frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} . \quad (27.103)$$

The quadrupole operator Q_{ab} yields zero when acting on δ_{ab} (that is, Q_{ab} is traceless), and the Laplacian operator ∇^2 when acting on $\nabla_a \nabla_b$

$$Q_{ab} \delta_{ab} = 0 , \quad Q_{ab} \nabla_a \nabla_b = \nabla^2 . \quad (27.104)$$

The solution of equation (27.101) is

$$\Psi - \Phi = - \int \left[\frac{3}{2} \frac{(x_a - x'_a)(x_b - x'_b)}{|\mathbf{x} - \mathbf{x}'|^2} - \frac{1}{2} \delta_{ab} \right] \frac{T_{ab}(x') d^3 x'}{|\mathbf{x} - \mathbf{x}'|} . \quad (27.105)$$

Taylor expanding equation (27.105) using equation (27.79) yields $\Psi - \Phi$ at large distance in the x -direction from a finite body,

$$\Psi - \Phi = - \frac{1}{x} \int [T_{xx} - \frac{1}{2}(T_{yy} + T_{zz})] d^3 x' + O(x^{-2}) . \quad (27.106)$$

Equation (27.101) shows that the source of the difference $\Psi - \Phi$ between the two scalar potentials is the quadrupole pressure. Since the quadrupole pressure is small if either there are no relativistic sources, or any relativistic sources are isotropic, it is often a good approximation to set $\Psi = \Phi$. An exception is where there is a significant anisotropic relativistic component. For example, the energy-momentum tensor of a static electric field is relativistic and anisotropic.

One situation where the difference between Ψ and Φ is appreciable is the case of freely-streaming photons (and neutrinos) at around the time of recombination in cosmology. The 2008 analysis of the CMB by the WMAP team claims to detect a non-zero value of $\Psi - \Phi$ from a slight shift in the third acoustic peak.

Exercise 27.7 Scalar potentials outside a spherical body. Argue that the traceless part of the spatial energy-momentum tensor of a spherically symmetric distribution must take the form

$$T_{ab}(r) = (\hat{r}_a \hat{r}_b - \frac{1}{3} \delta_{ab}) (p(r) - p_{\perp}(r)) , \quad (27.107)$$

where $p(r)$ and $p_{\perp}(r)$ are the radial and transverse pressures at radius r . From equation (27.105), show that $\Psi - \Phi$ at radial distance x from the centre of a spherically symmetric distribution is

$$\Psi(x) - \Phi(x) = - \int_x^{\infty} (r^2 - x^2) (p(r) - p_{\perp}(r)) \frac{4\pi dr}{r} . \quad (27.108)$$

Notice that the integral is over $r > x$, that is, only energy-momentum outside radius x produces non-vanishing $\Psi - \Phi$. Show that if the only source of energy-momentum outside the body is an electric charge Q , for which $-p = p_{\perp} = Q^2/r^4$, then

$$\Psi(x) - \Phi(x) = \frac{2\pi Q^2}{x^2} . \quad (27.109)$$

27.13 Gravitational waves

The tensor perturbations h_{ab} describe propagating gravitational waves. The two independent components of the tensor perturbations describe two polarizations. The two components are commonly designated h_+ and $h_{\times}$, equations (27.19). Gravitational waves induce a quadrupole tidal oscillation transverse to the direction of propagation, and the subscripts $+$ and $\times$ represent the shape of the quadrupole oscillation, as illustrated

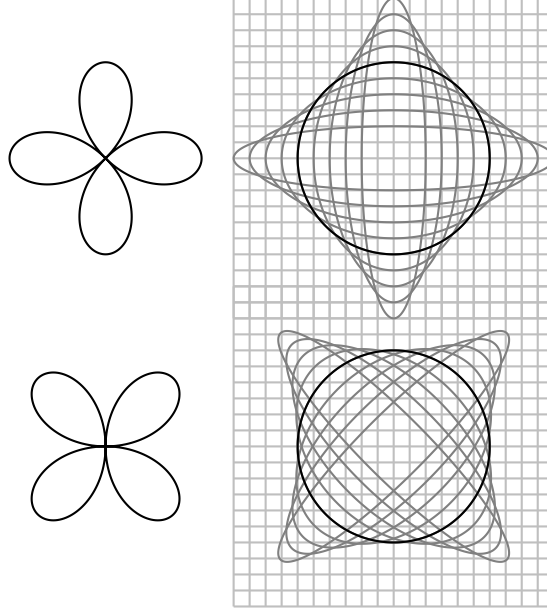


Figure 27.1 The two polarizations of gravitational waves. The (top) polarization h_+ varies as $\cos 2\chi$ under a right-handed rotation by angle χ about the direction of propagation (into the paper), while the (bottom) polarization $h_\times$ varies as $-\sin 2\chi$. A gravitational wave causes a system of freely falling test masses to oscillate relative to a grid of points a fixed proper distance apart.

by Figure 27.1. The h_+ polarization varies as $\cos 2\chi$ under a right-handed rotation by angle χ about the direction of propagation (the z -direction), while the $h_\times$ polarization varies as $-\sin 2\chi$.

Einstein's equations applied to the tensor component of the spatial Einstein tensor (27.16c) imply that gravitational waves are sourced by the tensor component of the energy-momentum

$$\square h_{ab} = 8\pi \underset{\text{tensor}}{T_{ab}} . \quad (27.110)$$

The solution of the wave equation (27.110) can be obtained from the Green's function of the d'Alembertian wave operator $\square$ defined by equation (27.17). The Green's function is by definition the solution of the wave equation with a delta-function source. There are retarded solutions, which propagate into the future along the future light cone, and advanced solutions, which propagate into the past along the past light cone. In the present case, the solutions of interest are the retarded solutions, since these represent gravitational waves emitted by a source. Because of the time and space translation symmetry of the d'Alembertian in flat (Minkowski) space, the delta-function source of the Green's function can without loss of generality be taken at the origin $t = \mathbf{x} = 0$. Thus the Green's function F is the solution of

$$\square F = \delta^4(x) , \quad (27.111)$$

where $\delta^4(x) \equiv \delta(t)\delta^3(\mathbf{x})$ is the 4-dimensional Dirac delta-function. The solution of equation (27.111) subject to retarded boundary conditions is (a standard exercise in mathematics) the retarded Green's function

$$F = \frac{\delta(t - |\mathbf{x}|)\Theta(t)}{4\pi|\mathbf{x}|}, \quad (27.112)$$

where $\Theta(t)$ is the Heaviside function, $\Theta(t) = 0$ for $t < 0$ and $\Theta(t) = 1$ for $t \geq 0$. The solution of the sourced gravitational wave equation (27.110) is thus

$$h_{ab}(t, \mathbf{x}) = -2 \int \frac{T_{ab}(t', \mathbf{x}') d^3x'}{|\mathbf{x}' - \mathbf{x}|}, \quad (27.113)$$

where t' is the retarded time

$$t' \equiv t - |\mathbf{x}' - \mathbf{x}|, \quad (27.114)$$

which lies on the past light cone of the observer, and is the time at which the source emitted the signal. The solution (27.113) resembles the solution of Poisson's equation, except that the source is evaluated along the past light cone of the observer.

As in §§27.10 and 27.11, consider a finite body, whose energy-momentum is confined within a certain region, and which is a source of gravitational waves. The Hulse-Taylor binary pulsar, Exercise 27.9, is a fine example. Far from the body, the leading order contribution to the tensor potential h_{ab} is, from the multipole expansion (27.79),

$$h_{ab}(t, \mathbf{x}) = -\frac{2}{x} \int T_{ab}(t', \mathbf{x}') d^3x'. \quad (27.115)$$

The integral (27.115) is hard to solve in general, but there is a simple solution for gravitational waves whose wavelengths are large compared to the size of the body. To obtain this solution, first consider that conservation of energy-momentum implies that

$$\frac{\partial^2 T^{00}}{\partial t^2} - \nabla_a \nabla_b T^{ba} = \frac{\partial}{\partial t} \left(\frac{\partial T^{00}}{\partial t} + \nabla_a T^{0a} \right) - \nabla_a \left(\frac{\partial T^{0a}}{\partial t} + \nabla_b T^{ba} \right) = 0. \quad (27.116)$$

Multiply by $x^a x^b$ and integrate

$$\int x^a x^b \frac{\partial^2 T^{00}}{\partial t^2} d^3x = \int x^a x^b \nabla_c \nabla_d T^{cd} d^3x = \int T^{cd} \nabla_c \nabla_d (x^a x^b) d^3x = 2 \int T^{ab} d^3x, \quad (27.117)$$

where the third expression follows from the second by a double integration by parts. For wavelengths that are long compared to the size of the body, the first expression of equations (27.117) is

$$\int x_a x_b \frac{\partial^2 T^{00}}{\partial t^2} d^3x \approx \frac{\partial^2}{\partial t^2} \int x_a x_b T^{00} d^3x = \frac{\partial^2 I_{ab}}{\partial t^2}, \quad (27.118)$$

where I_{ab} is the second moment of the mass

$$I_{ab} \equiv \int x_a x_b T^{00} d^3x. \quad (27.119)$$

The tensor (spin 2) part of the energy-momentum is trace-free. The trace-free part $\mathcal{I}_{ab}$ of the second moment I_{ab} is the quadrupole moment of the mass distribution (this definition is conventional, but differs by a factor of 2/3 from what is called the quadrupole moment in spherical harmonics)

$$\mathcal{I}_{ab} \equiv I_{ab} - \frac{1}{3} \delta_{ab} I_c^c = \int (x_a x_b - \frac{1}{3} \delta_{ab} x^2) T^{00} d^3x . \quad (27.120)$$

Substituting the last expression of equations (27.117) into equation (27.115) gives the quadrupole formula for gravitational radiation at wavelengths long compared to the size of the emitting body

$$\boxed{h_{ab}(t, \mathbf{x}) = -\frac{1}{x} \underset{\text{tensor}}{\ddot{\mathcal{I}}_{ab}}(t-x)} . \quad (27.121)$$

Equation (27.121) is valid for long wavelength modes observed at distances x far from the source of gravitational radiation. The right hand side is evaluated at retarded time $t-x$: the observer is looking at the source as it used to be at time $t-x$.

If the gravitational wave is moving in the z -direction, then the tensor components of the quadrupole moment $\mathcal{I}_{ab}$ are

$$\mathcal{I}_+ = \frac{1}{2}(I_{xx} - I_{yy}) , \quad \mathcal{I}_\times = \frac{1}{2}(I_{xy} + I_{yx}) . \quad (27.122)$$

Concept question 27.8 Units of the gravitational quadrupole radiation formula. Restore units to the quadrupole formula (27.121) for gravitational radiation. **Answer:**

$$h_{ab}(t, \mathbf{x}) = -\frac{G}{c^4 x} \underset{\text{tensor}}{\ddot{\mathcal{I}}_{ab}}(t-x) . \quad (27.123)$$

27.14 Energy-momentum carried by gravitational waves

The gravitational wave equation (27.27) in empty space appears to describe gravitational waves propagating in a region where the energy-momentum tensor T_{mn} is zero. However, gravitational waves do carry energy-momentum, just as do other kinds of waves, such as electromagnetic waves. The energy-momentum is quadratic in the tensor perturbation h_{ab} , and so vanishes to linear order.

To determine the energy-momentum in gravitational waves, calculate the Einstein tensor G_{mn} to second order, imposing the vacuum conditions that the unperturbed and linear parts of the Einstein tensor vanish

$$\overset{0}{G}_{mn} = \overset{1}{G}_{mn} = 0 . \quad (27.124)$$

The parts of the second-order perturbation that depend on the tensor perturbation h_{ab} are

$$\overset{2}{G}_{mn} = -(\nabla_m h_{ab})(\nabla_n h^{ab}) + \frac{1}{4} \eta_{mn} \square h^2 . \quad (27.125)$$

where

$$h^2 \equiv h_{ab}h^{ab} = 2(h_+^2 + h_\times^2) = 2h_{++}h_{--} . \quad (27.126)$$

Since the Einstein tensor vanishes to linear order, equations (27.124), the Lie derivative of the linear order Einstein tensor is zero, and consequently the quadratic order expression (27.125) is coordinate gauge-invariant. It is also tetrad gauge-invariant since it depends only on the (coordinate and) tetrad gauge-invariant perturbation h_{ab} . In a frame where the wavevector $\mathbf{k}$ is along the z -axis, the rightmost term on the right hand side of equation (27.125) is a sum of total derivatives with respect to time t or space z ,

$$\square h^2 = (-\partial_t^2 + \nabla_z^2) h^2 . \quad (27.127)$$

These terms yield surface terms when integrated over a region, and tend to average to zero when integrated over a region much larger than a wavelength. On the other hand, the leftmost set of terms on the right hand side of each of equations (27.125) do not average to zero for m and/or n equal to 0 or z . Notably, the terms for G_{00} and G_{zz} are $-(\partial_t h_{ab})(\partial_t h^{ab})$ and $-(\nabla_z h_{ab})(\nabla_z h^{ab})$, which being minus sums of squares are negative everywhere. A negative energy density? The interpretation is that these terms are to be taken over to the right hand side of the Einstein equations, and re-interpreted as the energy-momentum T_{mn}^{gw} in gravitational waves

$$T_{mn}^{\text{gw}} \equiv \frac{1}{8\pi} [(\nabla_m h_{ab})(\nabla_n h^{ab}) - \frac{1}{4}\eta_{mn}\square h^2] \quad (27.128)$$

The terms involving total derivatives, although they vanish when averaged over a region larger than many wavelengths, ensure that the energy-momentum T_{mn}^{gw} in gravitational waves satisfies conservation of energy-momentum in the flat background space,

$$\nabla^m T_{mn}^{\text{gw}} = 0 . \quad (27.129)$$

Averaged over a region larger than many wavelengths, the energy-momentum in gravitational waves is

$$\langle T_{mn}^{\text{gw}} \rangle = \frac{1}{8\pi} (\nabla_m h_{ab})(\nabla_n h^{ab}) . \quad (27.130)$$

Equation (27.130) may also be written explicitly as a sum over the two linear or circular polarizations

$$\begin{aligned} \langle T_{mn}^{\text{gw}} \rangle &= \frac{1}{4\pi} [(\nabla_m h_+)(\nabla_n h_+) + (\nabla_m h_\times)(\nabla_n h_\times)] \\ &= \frac{1}{8\pi} [(\nabla_m h_{++})(\nabla_n h_{--}) + (\nabla_m h_{+-})(\nabla_n h_{-+})] . \end{aligned} \quad (27.131)$$

Exercise 27.9 Hulse-Taylor binary.

1. **Quadrupole moment.** Consider a pair of masses M_1 and M_2 in circular orbit, with position vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ relative to their center of mass. Argue that the quadrupole moment $\mathcal{I}_{ab}$ of the mass distribution defined by

$$\mathcal{I}_{ab} \equiv \sum_{\text{masses } X} M_X (r_{X,a} r_{X,b} - \frac{1}{3} \delta_{ab} r_X^2) \quad (27.132)$$

is

$$\mathcal{I}_{ab} = mr^2(\hat{r}_a\hat{r}_b - \frac{1}{3}\delta_{ab}) , \quad (27.133)$$

where $\mathbf{r} \equiv r\hat{\mathbf{x}} \equiv \mathbf{r}_2 - \mathbf{r}_1$ is the orbital separation, and m is the reduced mass

$$m \equiv \frac{M_1 M_2}{M} , \quad M \equiv M_1 + M_2 . \quad (27.134)$$

[Hint: Assume for simplicity that the orbit is described by classical Newtonian mechanics.]

2. **Tensor components.** Suppose that the orbital plane is inclined at inclination angle ι to the line-of-sight. Choose the observer's locally inertial frame so that the z -axis $\hat{\mathbf{z}}$ is the line-of-sight direction from the center of mass of the binary to the observer, and the x -axis $\hat{\mathbf{x}}$ points in the plane of the orbit. Argue that the orbital separation $\mathbf{r}$ is

$$\mathbf{r} = r[(\hat{\mathbf{z}} \cos \iota - \hat{\mathbf{y}} \sin \iota) \cos \omega t + \hat{\mathbf{x}} \sin \omega t] \quad (27.135)$$

where ω is the orbital frequency. Deduce that the tensor components of the quadrupole moment are

$$\mathcal{I}_+ \equiv \frac{1}{2}(\mathcal{I}_{xx} - \mathcal{I}_{yy}) = \frac{1}{4}mr^2 [\cos^2 \iota - (1 + \sin^2 \iota) \cos 2\omega t] , \quad (27.136a)$$

$$\mathcal{I}_\times \equiv \mathcal{I}_{xy} = -\frac{1}{2}mr^2 \sin \iota \sin 2\omega t . \quad (27.136b)$$

[Hint: Recall the trigonometric formulae $\cos^2 \phi = \frac{1}{2}(1 + \cos 2\phi)$ and $\sin^2 \phi = \frac{1}{2}(1 - \cos 2\phi)$.]

3. **Tensor perturbation.** Deduce the tensor perturbations h_+ and $h_\times$ at large distance z from the orbiting masses from the quadrupole formula

$$h_{ab} = -\frac{1}{z} \ddot{\mathcal{I}}_{ab}(t - z) . \quad (27.137)$$

Notice that $t - z$ is the retarded time: an observer at distance z is looking at the orbiting masses as they used to be at time $t - z$.

4. **Energy momentum in gravitational waves.** The energy-momentum T_{mn}^{gw} in gravitational waves is given by the quadrupole formula

$$4\pi T_{mn}^{\text{gw}} = (\nabla_m h_+)(\nabla_n h_+) + (\nabla_m h_\times)(\nabla_n h_\times) , \quad (27.138)$$

where $\nabla_m = \{\partial/\partial t, \partial/\partial x^a\}$. Show that the non-vanishing components of the gravitational wave energy-momentum tensor are

$$T_{00}^{\text{gw}} = -T_{0z}^{\text{gw}} = T_{zz}^{\text{gw}} = \frac{m^2 r^4 \omega^6}{2\pi z^2} (1 + 6\sin^2 \iota + \sin^4 \iota - \cos^4 \iota \cos 4\omega(t - z)) . \quad (27.139)$$

[Hint: The quadrupole formula is valid for large z , so you need keep only the leading term in powers of z .]

5. **Energy flux in gravitational waves** The energy loss $\dot{E}$ by gravitational waves is given by the integral of the energy flux over all directions (note that energy flux is T^{0z} with raised indices, and there is a minus sign from $T^{0z} = -T_{0z}$),

$$\dot{E} = - \int_{-\pi/2}^{\pi/2} T_{0z}^{\text{gw}} 2\pi z^2 \cos \iota d\iota . \quad (27.140)$$

Show that (with units of c and G restored)

$$\dot{E} = \frac{32Gm^2r^4\omega^6}{5c^5} . \quad (27.141)$$

6. **Rate of change of orbital frequency.** If the orbit of the binary is described adequately by a Keplerian orbit, then the orbital energy E is

$$E = -\frac{GmM}{2r} , \quad (27.142)$$

and the radius r and angular frequency ω are related by Kepler's third law

$$r^3 = \frac{GM}{\omega^2} . \quad (27.143)$$

The orbital period P is related to the angular frequency ω by

$$P \equiv \frac{2\pi}{\omega} . \quad (27.144)$$

Conclude that

$$\frac{\dot{P}}{P} = -\frac{\dot{\omega}}{\omega} = \frac{3}{2} \frac{\dot{E}}{E} = -\frac{96(Gm)(GM)^{2/3}\omega^{8/3}}{5c^5} , \quad (27.145)$$

the minus sign in the third expression coming from the fact that the orbit is losing energy.

7. **Hulse-Taylor binary.** The so-called binary pulsar PSR B1913+16 discovered by Hulse and Taylor (1975) consists of two neutron stars, one a pulsar, in orbit. The masses of the pulsar and its companion are measured from the orbital motion to be (Weisberg and Taylor, 2005)

$$M_1 = 1.4414 M_\odot , \quad M_2 = 1.3867 M_\odot . \quad (27.146)$$

The orbital period is

$$P = 0.322997448930 \text{ day} . \quad (27.147)$$

What is the predicted general relativistic rate of change $\dot{P}$ of the period, in dimensionless units (or s/s, if you prefer)? [Hint: The heliocentric gravitational constant is $GM_\odot = 1.3271244 \times 10^{20} \text{ m}^3 \text{ s}^{-2}$ according to the IAU 2009 system of astronomical constants at <https://link.springer.com/article/10.1007%2F978-94-007-9352-4>.]

8. **Eccentricity correction.** Actually PSR B1913+16 has a substantial eccentricity,

$$e = 0.6171338 . \quad (27.148)$$

The correct general relativistic formula including the effects of eccentricity is equation (27.145) multiplied by a function $f(e)$ of the eccentricity

$$\frac{\dot{P}}{P} = -\frac{96(Gm)(GM)^{2/3}\omega^{8/3}}{5c^5} f(e) , \quad (27.149)$$

with

$$f(e) = \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right) (1 - e^2)^{-7/2} . \quad (27.150)$$

Compare the eccentricity-corrected predicted numerical result for $\dot{P}$ with the measured value

$$\dot{P} = -2.4184 \times 10^{-12} . \quad (27.151)$$

Exercise 27.10 Will you be torn apart when two black holes merge? The book “Death from the Skies!” by Phil Plait (the Bad Astronomer) contains a Chapter “Seven ways a black hole can kill you.” One of the ways, says Phil, is to stand near a pair of merging black holes, and be torn apart by the tidal forces from the gravitational waves. Is it true?

1. **Tidal forces.** For a gravitational wave propagating in the z -direction in empty space, the non-zero components of the Riemann tensor of the perturbed Minkowski space are

$$R_{0x0x} = -R_{0y0y} = -R_{0xzx} = R_{0yzy} = R_{zxzx} = -R_{zyzy} = \ddot{h}_+ , \quad (27.152a)$$

$$R_{0x0y} = -R_{0xzy} = -R_{0yzx} = R_{zxzy} = \ddot{h}_\times . \quad (27.152b)$$

From the expression (27.137) for h_{ab} that you derived in Exercise 27.9, and from the equation of geodesic deviation

$$\frac{D^2 \delta \xi_m}{D\tau^2} + R_{klmn} \delta \xi^k u^l u^n = 0 \quad (27.153)$$

deduce the tidal forces on a person moving nonrelativistically. [Hint: If a person is moving nonrelativistically, it is legitimate to take the person’s 4-velocity to be $u^m = \{1, 0, 0, 0\}$. Why?]

2. **Comment.** What is your advice to Phil Plait? [Hint: What you need here is rough estimates. Consider both supermassive and stellar-sized black holes. To make things sensible, you should require that you, the observer, be (a) outside the horizon, and (b) outside the point at which the static tidal force of the black hole would tear you apart even without gravitational waves. You may find it convenient to define the mass M_g of a black hole whose tidal force at the horizon is 1 gee per metre

$$g = \frac{1}{M_g^2} \quad (27.154)$$

which you figured out in Exercise 11.11.]

Concept Questions

1. Why do the wavelengths of perturbations in cosmology expand with the Universe, whereas perturbations in Minkowski space do not expand?
2. What does power spectrum mean?
3. Why is the power spectrum a good way to characterize the amplitude of fluctuations?
4. Why is the power spectrum of fluctuations of the Cosmic Microwave Background (CMB) plotted as a function of harmonic number?
5. What causes the acoustic peaks in the power spectrum of fluctuations of the CMB?
6. Are there acoustic peaks in the power spectrum of matter (galaxies) today?
7. What sets the scale of the first peak in the power spectrum of the CMB? [What sets the physical scale? Then what sets the angular scale?]
8. The odd peaks (including the first peak) in the CMB power spectrum are compression peaks, while the even peaks are rarefaction peaks. Why does a rarefaction produce a peak, not a trough?
9. Why is the first peak the most prominent? Why do higher peaks generally get progressively weaker?
10. The third peak is about as strong as the second peak? Why?
11. The matter power spectrum reaches a maximum at a scale that is slightly larger than the scale of the first baryonic acoustic peak. Why?
12. The physical density of species x at the time of recombination is proportional to $\Omega_x h^2$ where Ω_x is the ratio of the actual to critical density of species x at the present time, and $h \equiv H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$ is the present-day Hubble constant. Explain.
13. How does changing the baryon density $\Omega_b h^2$ affect the CMB power spectrum?
14. How does changing the non-baryonic cold dark matter density $\Omega_c h^2$, without changing the baryon density $\Omega_b h^2$, affect the CMB power spectrum?
15. What effects do neutrinos have on perturbations?
16. How does changing the curvature Ω_k affect the CMB power spectrum?
17. How does changing the dark energy Ω_Λ affect the CMB power spectrum?

An overview of cosmological perturbations

Undoubtedly the preeminent application of general relativistic perturbation theory is to cosmology. Fluctuations in the temperature and polarization of the Cosmic Microwave Background (CMB) provide an observational window on the Universe at 400,000 years old that, coupled with other astronomical observations, has yielded impressively precise measurements of cosmological parameters.

The theory of cosmological perturbations is based principally on general relativistic perturbation theory coupled to the physics of 5 species of energy-momentum: photons, baryons, non-baryonic cold dark matter, neutrinos, and dark energy.

Dark energy was not important at the time of recombination, where the CMB that we see comes from, but it is important today. If dark energy has a vacuum equation of state, $p = -\rho$, then dark energy does not cluster (vacuum energy density is a constant), but it affects the evolution of the cosmic scale factor, and thereby does affect the clustering of baryons and dark matter today. Moreover the evolution of the gravitational potential along the line of sight to the CMB does affect the observed power spectrum of the CMB, the so-called integrated Sachs-Wolfe effect.

1. **Inflationary initial conditions.** The theory of inflation has been remarkably successful in accounting for many aspects of observational cosmology, even though a fundamental understanding of the inflaton scalar field that supposedly drove inflation is missing. The current paradigm holds that primordial fluctuations were generated by vacuum quantum fluctuations in the inflaton field at the time of inflation. The theory makes the generic predictions that the gravitational potentials generated by vacuum fluctuations were (a) **Gaussian**, (b) **adiabatic** (meaning that all species of mass-energy fluctuated together, as opposed to in opposition to each other), and (c) **scale-free**, or rather almost scale-free (the fact that inflation came to an end modifies slightly the scale-free character). The three predictions fit the observed power spectrum of the CMB astonishingly well.
2. **Comoving Fourier modes.** The spatial homogeneity of the Friedmann-Lemaître-Robertson-Walker background spacetime means that its linear perturbations are characterized by Fourier modes of constant comoving wavevector. Each Fourier mode generated by inflation evolved independently, and its wavelength expanded with the Universe.
3. **Scalar, vector, tensor modes.** Spatial isotropy on top of spatial homogeneity means that the pertur-

bations comprised independently evolving scalar, vector, and tensor modes. Scalar modes dominate the fluctuations of the CMB, and caused the clustering of matter today. Vector modes are usually assumed to vanish, because there is no mechanism to generate the rotation that sources vector modes, and the expansion of the Universe tends to redshift away any vector modes that might have been present. Inflation generates gravitational waves, which then propagate essentially freely to the present time. Gravitational waves leave an observational imprint in the “*B*” (magnetic $(-)^{\ell+1}$ parity) mode of polarization of the CMB, whereas scalar modes produce only an “*E*” (electric $(-)^{\ell}$ parity) mode of polarization.

4. **Power spectrum.** The primary quantity measurable from observations is the **power spectrum**, which is the variance of fluctuations of the CMB or of matter (as traced by galaxies, galaxy clusters, the Lyman alpha forest, peculiar velocities, weak lensing, or 21 centimetre observations at high redshift). The statistics of a Gaussian field are completely characterized by its mean and variance. The mean characterizes the unperturbed background, while the variance characterizes the fluctuations. For a 3-dimensional statistically homogeneous and isotropic field, the variance of Fourier modes $\delta_{\mathbf{k}}$ defines the power spectrum $P(k)$,

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = \mathbf{1}_{\mathbf{k}\mathbf{k}'} P(k) , \quad (28.1)$$

where $\mathbf{1}_{\mathbf{k}\mathbf{k}'}$ is the unit matrix in the Hilbert space of Fourier modes,

$$\mathbf{1}_{\mathbf{k}\mathbf{k}'} \equiv (2\pi)^3 \delta_D^3(\mathbf{k} + \mathbf{k}') . \quad (28.2)$$

The “momentum-conserving” Dirac delta function in equation (28.2) is a consequence of statistical spatial translation symmetry. Isotropy implies that the power spectrum $P(k)$ is a function only of the magnitude $k \equiv |\mathbf{k}|$ of the wavevector. For a statistically rotation-invariant field projected on the sky, such as the CMB, the variance of spherical harmonic modes $\Theta_{\ell m} \equiv \delta T_{\ell m}/T$ defines the power spectrum C_{ℓ} ,

$$\langle \Theta_{\ell m} \Theta_{\ell' m'} \rangle = \mathbf{1}_{\ell m, \ell' m'} C_{\ell} \quad (28.3)$$

where $\mathbf{1}_{\ell m, \ell' m'}$ is the unit matrix in the Hilbert space of spherical harmonics (distinguish the three usages of δ in this paragraph: δ meaning fluctuation, δ_D meaning Dirac delta function, and δ meaning Kronecker delta, as in the following equation),

$$\mathbf{1}_{\ell m, \ell' m'} \equiv \delta_{\ell \ell'} \delta_{m, -m'} . \quad (28.4)$$

Again, the “angular momentum-preserving” condition (28.4) that $\ell = \ell'$ and $m + m' = 0$ is a consequence of rotational symmetry. The same rotational symmetry implies that the power spectrum C_{ℓ} is a function only of the harmonic number ℓ , not of the directional harmonic number m .

5. **Reheating.** Early Universe inflation evidently came to an end. It is presumed that the vacuum energy released by the decay of the inflaton field, an event called **reheating**, somehow efficiently produced the matter and radiation fields that we see today. After reheating, the Universe was dominated by relativistic fields, collectively called “radiation.” Reheating changed the evolution of the cosmic scale factor from acceleration to deceleration, but is presumed not to have generated additional fluctuations.

6. **Photon-baryon fluid and the sound horizon.** Photon-electron (Thomson) scattering kept photons and baryons tightly coupled to each other, so that they behaved like a relativistic fluid. As long as the radiation density exceeded the baryon density, which remained true up to near the time of recombination, the speed of sound in the photon-baryon fluid was $\sqrt{p/\rho} \approx \sqrt{\frac{1}{3}}$ of the speed of light, §30.6. Fluctuations with wavelengths outside the sound horizon grew by gravity. As time went by, the sound horizon expanded in comoving radius, and fluctuations thereby came inside the sound horizon. Once inside the sound horizon, sound waves could propagate, which tended to decrease the gravitational potential. However, each individual sound wave itself continued to oscillate, its oscillation amplitude $\delta T/T$ relative to the background temperature T remaining approximately constant, Fig. 30.7, at least well before recombination, when damping is unimportant (point 11 below). The suppression of the potential at small scales is responsible for the turnover in the observed power spectrum of matter fluctuations today from large to small scales, Fig. 30.15.
7. **Acoustic peaks in the power spectrum.** The oscillations of the photon-baryon fluid produced the characteristic pattern of peaks and troughs in the CMB power spectrum observed today. The same peaks and troughs occur in the matter power spectrum, but are much less prominent, at a level of about 10% as opposed to the order unity oscillations observed in the CMB power spectrum. For adiabatic fluctuations, the amplitude of the temperature fluctuations follows a pattern $\sim -\cos(k\eta_s)$ where η_s is the comoving sound horizon, Fig. 30.7. The n th peak occurs at a wavenumber k where $k\eta_s \approx n\pi$. In the observed CMB power spectrum, the relevant value of the sound horizon η_s is its value $\eta_{s,\text{rec}}$ at recombination. Thus the wavenumber k of the first peak of the observed CMB power spectrum occurs where $k\eta_{s,\text{rec}} \approx \pi$. Two competing forces cause a mode to evolve: a gravitational force that amplifies compression, and a restoring pressure force that counteracts compression, §32.10. When a mode enters the sound horizon for the first time, the compressing gravitational force beats the restoring pressure force, so the first thing that happens is that the mode compresses further. Consequently the first peak is a compression peak. This sets the subsequent pattern: odd peaks are compression peaks, while even peaks are rarefaction peaks. The observed temperature fluctuations of the CMB are produced by a combination of intrinsic temperature fluctuations, Doppler shifts, and gravitational redshifting out of potential wells. The Doppler shift produced by the velocity of a perturbation is 90° out of phase with the temperature fluctuation, and so tends to fill in the troughs in the power spectrum of the temperature fluctuation. This is the main reason that the observed CMB power spectrum remains above zero at all scales.
8. **Logarithmic growth of matter fluctuations.** Non-baryonic cold dark matter interacts weakly except by gravity, and is needed to explain the observed clustering of matter in the Universe today in spite of the small amplitude of temperature fluctuations in the CMB. The adjective “cold” refers to the requirement that the dark matter became nonrelativistic ($p = 0$) at some early time. If the dark matter is both non-baryonic and cold, then it did not participate in the oscillations of the photon-baryon fluid. During the radiation-dominated phase prior to matter-radiation equality, dark matter matter fluctuations inside the sound horizon grow logarithmically, Fig 30.10. The logarithmic growth translates into a logarithmic increase in the amplitude of matter fluctuations at small scales, and is a characteristic signature of non-

baryonic cold dark matter. Unfortunately this signature is not readily discernible in the power spectrum of matter today, because of nonlinear clustering.

9. **Epoch of matter-radiation equality.** The density of nonrelativistic matter decreases more slowly than the density of relativistic radiation. There came a point where the matter density equaled the radiation density, an epoch called matter-radiation equality, after which the matter density exceeded the radiation density. The observed ratio of the density of matter and radiation (CMB) today require that matter-radiation equality occurred at a redshift of $z_{\text{eq}} \approx 3400$, equation (30.42), a factor of 3 higher in redshift than recombination at $z_{\text{rec}} \approx 1100$. After matter-radiation equality, dark matter perturbations grew more rapidly, linearly instead of just logarithmically with cosmic scale factor. A larger dark matter density causes matter-radiation equality to occur earlier. The sound horizon at matter-radiation equality corresponds to a scale roughly around the 2.5th peak in the CMB power spectrum. For adiabatic fluctuations, the way that the temperature and gravitational perturbations interact when a mode first enters the sound horizon means that the temperature oscillation is 5 times larger for modes that enter the horizon well into the radiation-dominated epoch versus well into the matter-dominated epoch, Fig. 32.3. The effect enhances the amplitude of observed CMB peaks higher than 2.5 relative to those lower than 2.5. The observed relative strengths of the 3rd versus the 2nd peak of the CMB power spectrum provides a measurement of the redshift of matter-radiation equality, and direct evidence for the presence of non-baryonic cold dark matter.
10. **Sound speed.** The density of baryons decreased more slowly than the density of radiation, so that at around recombination the baryon density was becoming comparable to the radiation density. The sound speed $\sqrt{p/\rho}$ depends on the ratio of pressure p , which was essentially entirely that of the photons, to the density ρ , which was produced by both photons and baryons. The sound speed consequently decreased below $\sqrt{\frac{1}{3}}$, §32.4. Increasing the baryon-to-photon ratio at recombination has several observational effects on the acoustic peaks of the CMB power spectrum, making it a prime measurable parameter from the CMB. First, an increased baryon fraction increases the gravitational forcing (baryon loading), which enhances the compression (odd) peaks while reducing the rarefaction (even) peaks. Second, increasing the baryon fraction reduces the sound speed, which: (a) decreases the amplitude of the radiation velocity relative to the radiation density, so increasing the prominence of the peaks; and (b) reduces the oscillation frequency of the photon-baryon fluid, which shifts the peaks to larger scales. The reduced sound speed also causes an adiabatic reduction of the amplitudes of all modes by the square root of the sound speed, but this effect is degenerate with an overall reduction in the initial amplitudes of modes produced by inflation.
11. **Electron-photon scattering.**

Prior to recombination, photons are coupled to the baryonic plasma mainly by nonrelativistic electron-photon (Thomson) scattering. The finite mean free path to scattering damps oscillations of the photon-baryon fluid. As recombination approaches, the mean free path grows longer, and the damping becomes greater, Fig. 32.3. Damping by Thomson scattering is responsible for the decline in the CMB power spectrum at smaller scales.

12. **Recombination.** As the temperature cooled below about 3,000 K, electrons combined with hydrogen and helium nuclei into neutral atoms, Fig. 31.4. This drastically reduced the amount of photon-electron scattering, releasing the CMB to propagate almost freely. At the same time, the baryons were released from the photons. Without radiation pressure to support them, fluctuations in the baryons began to grow like the dark matter fluctuations.
13. **Neutrinos.** Probably all three species of neutrino have mass less than 0.2 eV and were therefore relativistic up to and at the time of recombination, equation (10.112). Each of the 3 species of neutrino had an abundance comparable to that of photons, and therefore made an important contribution to the relativistic background and its fluctuations. Unlike photons, neutrinos streamed freely, without scattering, Fig. 33.2. The relativistic free-streaming of neutrinos provided the main source of the quadrupole pressure that produces a non-vanishing difference $\Psi - \Phi$ between the scalar potentials, Fig. 33.4. However, the neutrino quadrupole pressure was still only $\sim 10\%$ of the neutrino monopole pressure. To the extent that the neutrino quadrupole pressure can be approximated as negligible, the neutrinos and their fluctuations can be treated the same as photons.
14. **CMB fluctuations.** The CMB fluctuations seen on the sky today represent a projection of fluctuations on a thin but finite shell at a redshift of about 1100, Fig. 34.1, corresponding to an age of the Universe of about 400,000 yr. The temperature, and the degrees of polarization in two different directions, provide 3 independent observables at each point on the sky. The isotropy of the unperturbed radiation means that it is most natural to measure the fluctuations in spherical harmonics, which are the eigenmodes of the rotation operator. Similarly, it is natural to measure the CMB polarization in spin harmonics.
15. **Matter fluctuations.** After recombination, perturbations in the non-baryonic and baryonic matter grew by gravity, essentially unaffected any longer by photon pressure, Fig. 32.3. If one or more of the neutrino types had a mass small enough to be relativistic but large enough to contribute appreciable density, then its relativistic streaming could have suppressed power in matter fluctuations at small scales, but observations show no evidence of such suppression, which places an upper limit of about an eV on the mass of the most massive neutrino. The matter power spectrum measured from the clustering of galaxies contains acoustic oscillations like the CMB power spectrum, but because the non-baryonic dark matter dominates the baryons, the oscillations are much smaller.
16. **Integrated Sachs-Wolfe effect.** Variations in the gravitational potential along the line of sight to the CMB affect the CMB power spectrum at large scales. This is called the **integrated Sachs-Wolfe** (ISW) effect, §34.2.2. If matter dominates the background, then the gravitational potential Φ has the property that it remains constant in time for linear fluctuations, and there is no ISW effect. In practice, ISW effects are produced by at least three distinct causes. First, an early-time ISW effect is produced by the fact that the Universe at recombination still has an appreciable component of radiation, and is not yet wholly matter-dominated. Second, a late-time ISW effect is produced either by curvature or by a cosmological constant. Third, a non-linear ISW effect is produced by non-linear evolution of the potential.

Cosmological perturbations in a flat FLRW background

For simplicity, this book considers only a flat (not closed or open) Friedmann-Lemaître-Robertson-Walker (FLRW) background. The comoving Hubble distance at recombination was much smaller than today, and consequently the cosmological density Ω was much closer to 1 at recombination than it is today. Since observations indicate that the Universe today is within 0.1% of being spatially flat, Table 10.1, it is an excellent approximation to treat the Universe at the time of recombination as being spatially flat. The generalization to perturbations in open or closed universes is covered by Hu et al. (1998).

With some modifications arising from cosmological expansion, perturbation theory on a flat FLRW background is quite similar to perturbation theory in flat (Minkowski) space, Chapter 27.

The strategy is to start in a completely general gauge, and to discover how the conformal Newtonian (Copernican) gauge, §29.8, which is used in subsequent Chapters, emerges naturally as that gauge in which the perturbations are precisely the physical perturbations.

29.1 Unperturbed line-element

It is convenient to choose the coordinate system $x^\mu \equiv \{x^0, x^1, x^2, x^3\} \equiv \{\eta, x, y, z\}$ to consist of conformal time η together with comoving Cartesian coordinates $\mathbf{x} \equiv x^\alpha \equiv \{x, y, z\}$. The coordinate metric of the unperturbed background flat FLRW geometry is then

$$ds^2 = a(\eta)^2 (-d\eta^2 + dx^2 + dy^2 + dz^2) \ , \quad (29.1)$$

where $a(\eta)$ is the cosmic scale factor. The unperturbed coordinate metric is thus the conformal Minkowski metric

$${}^0g_{\mu\nu} = a(\eta)^2 \eta_{\mu\nu} \ . \quad (29.2)$$

The tetrad is taken to be orthonormal, with the unperturbed tetrad axes $\boldsymbol{\gamma}_m \equiv \{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3\}$ being aligned with the unperturbed coordinate axes ${}^0\mathbf{e}_\mu \equiv \{{}^0\mathbf{e}_0, {}^0\mathbf{e}_1, {}^0\mathbf{e}_2, {}^0\mathbf{e}_3\}$ so that the unperturbed vierbein and inverse

vierbein are respectively a and $1/a$ times the unit matrix,

$$e^m_{\mu} = a \delta^m_{\mu} , \quad e_m^{\mu} = \frac{1}{a} \delta_m^{\mu} , \quad (29.3)$$

and the tetrad metric is Minkowski, $\gamma_{mn} = \eta_{mn}$.

Let ∇_a denote spatial derivatives with respect to comoving spatial coordinates,

$$\nabla_a \equiv \delta_a^{\alpha} \frac{\partial}{\partial x^{\alpha}} = \frac{\partial}{\partial x^a} , \quad (29.4)$$

which should be distinguished from the directed derivatives $\partial_a \equiv e_a^{\mu} \partial / \partial x^{\mu} \approx (1/a) \partial / \partial x^a$. Because the background FLRW geometry is spatially homogeneous, comoving spatial gradients ∇_a are of first order, and can be treated as spatial vectors whose tetrad-frame components can be raised and lowered with the Euclidean metric. Further, let overdot denote partial differentiation with respect to conformal time η ,

$$\text{overdot} \equiv \frac{\partial}{\partial \eta} , \quad (29.5)$$

so that for example $\dot{a} \equiv da/d\eta$. The Hubble parameter H in the unperturbed background is

$$H \equiv \frac{\dot{a}}{a^2} . \quad (29.6)$$

29.2 Comoving Fourier modes

Since the unperturbed Friedmann-Lemaître-Robertson-Walker spacetime is spatially homogeneous and isotropic, it is natural to work in comoving Fourier modes. Comoving Fourier modes have the key property that they evolve independently of each other, as long as perturbations remain linear. Equations in Fourier space are obtained by replacing the comoving spatial gradient ∇_a by $-i$ times the comoving wavevector k_a (the choice of sign is the standard convention in cosmology)

$$\nabla_a \rightarrow -ik_a . \quad (29.7)$$

By this means, the spatial derivatives become algebraic, so that the partial differential equations governing the evolution of perturbations become ordinary differential equations.

In what follows, the comoving spatial gradient ∇_a will be used interchangeably with $-ik_a$, whichever is most convenient.

29.3 Classification of vierbein perturbations

The definition (26.1) of the vierbein perturbations φ_{mn} implies that the perturbed inverse vierbein in the perturbed FLRW spacetime is

$$e_m^{\mu} = \frac{1}{a} (\delta_m^n + \varphi_m^n) \delta_n^{\mu} = (\eta_{mn} + \varphi_{mn}) e^{n\mu} , \quad (29.8)$$

while the perturbed vierbein is

$$e^m{}_\mu = a(\delta_n^m - \varphi_n{}^m)\delta_\mu^n = (\eta^{mn} - \varphi^{nm})e_{n\mu}^0. \quad (29.9)$$

The covariant tetrad-frame components φ_{mn} of the vierbein perturbation of the FLRW geometry decompose in much the same way as in flat Minkowski case into 6 scalars, 4 vectors, and 1 tensor, a total of $6 + 4 \times 2 + 1 \times 2 = 16$ degrees of freedom (the following equations are essentially the same as those (27.6) for the flat Minkowski background),

$$\varphi_{00} = \underset{\text{scalar}}{\psi}, \quad (29.10a)$$

$$\varphi_{0a} = \underset{\text{scalar}}{\nabla_a w} + \underset{\text{vector}}{w_a}, \quad (29.10b)$$

$$\varphi_{a0} = \underset{\text{scalar}}{\nabla_a \tilde{w}} + \underset{\text{vector}}{\tilde{w}_a}, \quad (29.10c)$$

$$\varphi_{ab} = \underset{\text{scalar}}{\delta_{ab} \phi} + \underset{\text{scalar}}{\nabla_a \nabla_b h} + \underset{\text{scalar}}{\varepsilon_{abc} \nabla_c \tilde{h}} + \underset{\text{vector}}{\nabla_a h_b} + \underset{\text{vector}}{\nabla_b \tilde{h}_a} + \underset{\text{tensor}}{h_{ab}}. \quad (29.10d)$$

For a single Fourier mode whose wavevector $\mathbf{k}$ is taken without loss of generality to lie in the z -direction, so that $\nabla_x = \nabla_y = 0$, equations (29.10) are essentially the same as those (27.7) in a Minkowski background. The 4 covariant tetrad-frame components ϵ_m of the coordinate shift of the coordinate gauge transformation (26.9) similarly decompose into 2 scalars and 1 vector (2 degrees of freedom) (the following equation is essentially the same as that (27.8) for the flat Minkowski background),

$$\epsilon_m = \left\{ \underset{\text{scalar}}{\epsilon_0}, \underset{\text{scalar}}{\nabla_a \epsilon} + \underset{\text{vector}}{\epsilon_a} \right\}. \quad (29.11)$$

The vierbein perturbations φ_{mn} transform under a coordinate gauge transformation (26.9) as, equation (29.12),

$$\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \partial_m \epsilon_n = \varphi_{mn} + \frac{1}{a} \nabla_m \epsilon_n, \quad (29.12)$$

with vanishing contribution from the unperturbed tetrad-frame connection, equation (29.23), since the latter is symmetric whereas equation (26.20) depends on an antisymmetric combination of connections. The individual components of the vierbein perturbations transform under a coordinate gauge transformation as

$$\varphi_{00} \rightarrow \underset{\text{scalar}}{\psi} + \frac{1}{a} \frac{\partial \epsilon_0}{\partial \eta}, \quad (29.13a)$$

$$\varphi_{0a} \rightarrow \underset{\text{scalar}}{\nabla_a \left(w + \frac{1}{a} \left(\frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon \right)} + \underset{\text{vector}}{\left(w_a + \frac{1}{a} \left(\frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon_a \right)}, \quad (29.13b)$$

$$\varphi_{a0} \rightarrow \underset{\text{scalar}}{\nabla_a \left(\tilde{w} + \frac{1}{a} \epsilon_0 \right)} + \underset{\text{vector}}{\tilde{w}_a}, \quad (29.13c)$$

$$\varphi_{ab} \rightarrow \underset{\text{scalar}}{\delta_{ab} \left(\phi - \frac{\dot{a}}{a^2} \epsilon_0 \right)} + \underset{\text{scalar}}{\nabla_a \nabla_b \left(h + \frac{1}{a} \epsilon \right)} + \underset{\text{scalar}}{\varepsilon_{abc} \nabla_c \tilde{h}} + \underset{\text{vector}}{\nabla_a \left(h_b + \frac{1}{a} \epsilon_b \right)} + \underset{\text{vector}}{\nabla_b \tilde{h}_a} + \underset{\text{tensor}}{h_{ab}}, \quad (29.13d)$$

or equivalently

$$\psi \rightarrow \psi + \frac{1}{a} \frac{\partial \epsilon_0}{\partial \eta} , \quad (29.14a)$$

$$w \rightarrow w + \frac{1}{a} \left(\frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon , \quad w_a \rightarrow w_a + \frac{1}{a} \left(\frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon_a , \quad (29.14b)$$

$$\tilde{w} \rightarrow \tilde{w} + \frac{1}{a} \epsilon_0 , \quad \tilde{w}_a \rightarrow \tilde{w}_a , \quad (29.14c)$$

$$\phi \rightarrow \phi - \frac{\dot{a}}{a^2} \epsilon_0 , \quad h \rightarrow h + \frac{1}{a} \epsilon , \quad \tilde{h} \rightarrow \tilde{h} , \quad h_a \rightarrow h_a + \frac{1}{a} \epsilon_a , \quad \tilde{h}_a \rightarrow \tilde{h}_a , \quad h_{ab} \rightarrow h_{ab} . \quad (29.14d)$$

Eliminating the coordinate shift ϵ_m from the transformations (29.14) yields 12 coordinate gauge-invariant combinations of the perturbations,

$$\underbrace{\psi - \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) \tilde{w}}_{\text{scalar}} , \quad \underbrace{w - \dot{h}}_{\text{scalar}} , \quad \underbrace{w_a - \dot{h}_a}_{\text{vector}} , \quad \underbrace{\tilde{w}_a}_{\text{vector}} , \quad \underbrace{\phi + \frac{\dot{a}}{a} \tilde{w}}_{\text{scalar}} , \quad \underbrace{\tilde{h}}_{\text{scalar}} , \quad \underbrace{\tilde{h}_a}_{\text{vector}} , \quad \underbrace{h_{ab}}_{\text{tensor}} . \quad (29.15)$$

Six combinations of these coordinate gauge-invariant perturbations depend only on the symmetric part $\varphi_{mn} + \varphi_{nm}$ of the vierbein perturbations, and are therefore tetrad gauge-invariant as well as coordinate gauge-invariant. These 6 coordinate and tetrad gauge-invariant perturbations comprise 2 scalars, 1 vector, and 1 tensor

$$\boxed{\Psi_{\text{scalar}} \equiv \psi - \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) (w + \tilde{w} - \dot{h})} , \quad (29.16a)$$

$$\boxed{\Phi_{\text{scalar}} \equiv \phi + \frac{\dot{a}}{a} (w + \tilde{w} - \dot{h})} , \quad (29.16b)$$

$$\boxed{W_a_{\text{vector}} \equiv w_a + \tilde{w}_a - \dot{h}_a - \dot{\tilde{h}}_a} , \quad (29.16c)$$

$$\boxed{h_{ab}_{\text{tensor}}} . \quad (29.16d)$$

The coordinate and tetrad gauge-invariant perturbations (29.16) reduce to those (27.13) in Minkowski space when the cosmic scale factor does not change, $\dot{a} = 0$.

29.4 Residual global gauge freedoms

There are residual global gauge freedoms associated with (a) uncertainty in the cosmic scale factor $a(\eta)$ in the background FLRW geometry, and (b) addition of spatially uniform but time-dependent contributions to vierbein components that are spatial gradients in equations (29.13), namely w , $\tilde{w}$, h , $\tilde{h}$, h_a , and $\tilde{h}_a$. The freedoms are global in the sense that they are spatially uniform functions of time η . The global gauge freedoms mean that the scalar and vector perturbations Ψ , Φ , and W_a are gauge-invariant only up to the addition of spatially uniform functions of time. The tensor perturbation h_{ab} remains fully gauge-invariant.

The scalar global gauge freedoms translate into freedoms to adjust the scalar perturbations $\Psi(\eta)$ and $\Phi(\eta)$ by arbitrary functions of time η , and are associated with a rescaling of the cosmic scale factor $a(\eta)$, along with a gauge transformation of the conformal time η . The residual vector global gauge freedoms translate into freedom to adjust the vector perturbation $W_a(\eta)$ by time-dependent global spatial transformations that leave the spatial FLRW geometry unchanged, which include a time-dependent spatially uniform motion in any direction, and a time-dependent spatially uniform rotation about any axis. The tensor perturbation h_{ab} remains fully gauge-invariant.

To exhibit the scalar global gauge freedoms, consider the line-element

$$ds^2 = a(\eta)^2 \left\{ -[1 + \Psi(\eta)]^2 d\eta^2 + [1 - \Phi(\eta)]^2 \delta_{ab} dx^a dx^b \right\} , \quad (29.17)$$

in which $\Psi(\eta)$ and $\Phi(\eta)$ are functions only of conformal time η . A rescaling of the cosmic scale factor a , together with a coordinate transformation of conformal time η ,

$$a \rightarrow a' = a(1 - \Phi) , \quad (29.18a)$$

$$d\eta \rightarrow d\eta' = \left(\frac{1 + \Psi}{1 - \Phi} \right) d\eta , \quad (29.18b)$$

brings the line-element (29.17) to FLRW form,

$$ds^2 = a'(\eta')^2 (-d\eta'^2 + \delta_{ab} dx^a dx^b) . \quad (29.19)$$

The rescaling (29.18a) of the cosmic scale factor a is distinct from any coordinate transformation, and constitutes an additional global gauge freedom over and above the coordinate and tetrad gauge freedoms discussed in §29.3. The transformation (29.18b) of the time coordinate is allowed because Ψ and Φ are functions only of time. The argument in §29.3 that Ψ and Φ are gauge-invariant is spoiled because the combination $w + \tilde{w} - \dot{h}$ upon which both Ψ and Φ depend, equations (29.16a) and (29.16b), can be adjusted by an arbitrary function of time, since each of w , $\tilde{w}$, and h appear inside spatial gradients in equations (29.13).

The vector global gauge freedoms translate into freedoms to adjust the vector perturbation $W_a(\eta)$ by a spatially constant vector and a spatially constant rotation. For example, a spatially constant vector in the x -direction is $W_a(\eta) = W(\eta)\{1, 0, 0\}$, while a spatially constant rotation about the x -direction is $W_a(\eta) = \omega(\eta)\{0, z, -y\}$. Both transformations are transverse, satisfying $\nabla_a W_a = 0$. A spatially uniform W_a corresponds to a uniform velocity of the entire Universe, which is observationally undetectable. A spatially uniform rotation W_a is also observationally undetectable, an expression of Mach's principle, Concept question 29.4.

Modes whose wavelengths are larger than the horizon size of an observer look spatially uniform to the observer. The observer cannot distinguish such modes from a change in the parameters of the background FLRW geometry. Thus an observer cannot measure the amplitudes Ψ , Φ , W_a , or h_{ab} of modes outside their horizon.

Of course, an observer can measure modes that were outside the horizon of an earlier observer. For example, astronomers on Earth today can and do measure in both the CMB and in galaxy clustering “superhorizon” modes that were outside the horizon of an observer at the time of recombination.

The residual global gauge freedoms mean that the intrinsic monopole mode of the observed CMB is unmeasurable, being indistinguishable from a rescaling of the temperature of the FLRW background. Moreover the intrinsic CMB dipole is unmeasurable, being indistinguishable from an adjustment of the rest frame of the FLRW background.

In the remainder of this book, the perturbations Ψ , Φ , W_a , and h_{ab} will be referred to as gauge-invariant on the understanding that this refers to modes that are measurable by (within the horizon of) the observer.

Concept question 29.1 Global curvature as a perturbation? The usual FLRW metric contains a curvature constant κ in addition to a cosmic scale factor a . Can curvature κ , if small, be treated as a perturbation to a flat FLRW geometry, and if so, how? Does the curvature perturbation represent a residual global gauge freedom? **Answer.** Yes, κ , if small, can be treated as a perturbation. The isotropic (Poincaré) form of the FLRW line-element, equation (10.26), takes the form

$$ds^2 = a(\eta)^2 \left(-d\eta^2 + \frac{1}{1 + \frac{1}{4}\kappa x^2} \delta_{ab} dx^a dx^b \right), \quad (29.20)$$

where $x^2 \equiv \sum_a x_a^2$ is the square of the comoving radial distance from the origin. If the curvature scale is much smaller than the horizon distance, $\frac{1}{2}\sqrt{|\kappa|}\eta \ll 1$, then the curvature looks like a perturbation proportional to the square of the comoving distance,

$$\Phi(x) = \frac{1}{8}\kappa x^2. \quad (29.21)$$

Is this a residual global gauge freedom? Equation (29.21) states that only the sum $\frac{1}{8}\kappa x^2 - \Phi$ is gauge-invariant, so yes there is a residual global gauge freedom associated with the ambiguity between κ and Φ . In pre-1998 days when astronomers were measuring $\Omega_m \approx 0.3$ and only the reckless contemplated non-zero Ω_Λ , it was necessary to consider that Nature might have chosen a substantial curvature $\Omega_k \approx 0.7$, in which case κ was decidedly non-zero (and negative), certainly not a perturbation. Post dark-energy, observations are stubbornly consistent with zero curvature. Occam's razor would then prefer the simpler of two models that fit the data, a flat background geometry $\kappa = 0$.

29.5 Metric, tetrad connections, Einstein tensor, and energy-momentum tensor

This section gives expressions in a completely general gauge for perturbed quantities in the flat Friedmann-Lemaître-Robertson-Walker background geometry.

29.5.1 Metric

The unperturbed metric is the FLRW metric (29.2). The perturbation $\overset{1}{g}_{\mu\nu}$ to the coordinate metric is, equation (26.6),

$$\overset{1}{g}_{\eta\eta} = -a^2 \underset{\text{scalar}}{2\psi} \ , \quad (29.22a)$$

$$\overset{1}{g}_{\eta a} = -a^2 \left[\underset{\text{scalar}}{\nabla_a(w + \tilde{w})} + \underset{\text{vector}}{(w_a + \tilde{w}_a)} \right] \ , \quad (29.22b)$$

$$\overset{1}{g}_{ab} = -a^2 \left[\underset{\text{scalar}}{2\phi\delta_{ab}} + \underset{\text{scalar}}{2\nabla_a\nabla_b h} + \underset{\text{vector}}{\nabla_a(h_b + \tilde{h}_b)} + \underset{\text{vector}}{\nabla_b(h_a + \tilde{h}_a)} + \underset{\text{tensor}}{2h_{ab}} \right] \ . \quad (29.22c)$$

The coordinate metric is tetrad gauge-invariant, but not coordinate gauge-invariant.

29.5.2 Tetrad-frame connections

The tetrad-frame connections Γ_{kmn} are obtained from the usual formula (11.54). The non-vanishing unperturbed tetrad-frame connections are

$$\overset{0}{\Gamma}_{0ab} = -\frac{\dot{a}}{a^2}\delta_{ab} \ . \quad (29.23)$$

The perturbations $\overset{1}{\Gamma}_{kmn}$ to the tetrad-frame connections are

$$\overset{1}{\Gamma}_{0a0} = \frac{1}{a} \left[-\underset{\text{scalar}}{\nabla_a \left(\psi - \left(\frac{\partial}{\partial\eta} + \frac{\dot{a}}{a} \right) \tilde{w} \right)} + \underset{\text{vector}}{\left(\frac{\partial}{\partial\eta} + \frac{\dot{a}}{a} \right) \tilde{w}_a} \right] \ , \quad (29.24a)$$

$$\overset{1}{\Gamma}_{0ab} = \frac{1}{a} \left[\underset{\text{scalar}}{F\delta_{ab}} - \underset{\text{scalar}}{\nabla_a\nabla_b(w - \dot{h})} - \frac{1}{2}(\underset{\text{vector}}{\nabla_a W_b} + \underset{\text{vector}}{\nabla_b W_a}) + \underset{\text{vector}}{\nabla_b \tilde{w}_a} + \underset{\text{tensor}}{\dot{h}_{ab}} \right] \ , \quad (29.24b)$$

$$\overset{1}{\Gamma}_{ab0} = \frac{1}{a} \left[\frac{1}{2}(\underset{\text{vector}}{\nabla_a W_b} - \underset{\text{vector}}{\nabla_b W_a}) - \frac{\partial}{\partial\eta}(\underset{\text{scalar}}{\varepsilon_{abd}\nabla_d \tilde{h}} - \underset{\text{vector}}{\nabla_a \tilde{h}_b} + \underset{\text{vector}}{\nabla_b \tilde{h}_a}) \right] \ , \quad (29.24c)$$

$$\begin{aligned} \overset{1}{\Gamma}_{abc} = \frac{1}{a} \left[\underset{\text{scalar}}{(\delta_{bc}\nabla_a - \delta_{ac}\nabla_b) \left(\phi + \frac{\dot{a}}{a}\tilde{w} \right)} - \frac{\dot{a}}{a}(\underset{\text{vector}}{\delta_{ac}\delta_{bd} - \delta_{bc}\delta_{ad}})\tilde{w}_d \right. \\ \left. - \underset{\text{scalar}}{\nabla_c(\varepsilon_{abd}\nabla_d \tilde{h}} - \underset{\text{vector}}{\nabla_a \tilde{h}_b} + \underset{\text{vector}}{\nabla_b \tilde{h}_a}) + \underset{\text{tensor}}{\nabla_a h_{bc} - \nabla_b h_{ac}} \right] \ , \end{aligned} \quad (29.24d)$$

where F is defined by

$$F \equiv \frac{\dot{a}}{a}\psi + \dot{\phi} \ . \quad (29.25)$$

Equations (29.24) show that the perturbations $\overset{1}{\Gamma}_{klm}$ of the tetrad-frame connections depend on all 12 of the coordinate gauge-invariant potentials (29.15). The only non-coordinate-gauge-invariant dependence of the tetrad-frame connections is on F defined by equation (29.25). The quantity F transforms under a coordinate

gauge transformation (26.9) as, from equations (29.14),

$$F \rightarrow F - \epsilon_0 \frac{d}{d\eta} \frac{\dot{a}}{a^2} . \quad (29.26)$$

Thus perturbations $\overset{1}{\Gamma}_{0a0}$, $\overset{1}{\Gamma}_{0ab}$ with $a \neq b$, $\overset{1}{\Gamma}_{ab0}$, and $\overset{1}{\Gamma}_{abc}$ are coordinate gauge-invariant, while the transformation (29.26) of F implies that Γ_{0ab} with $a = b$ transforms under an infinitesimal coordinate transformation (26.9) as

$$\overset{1}{\Gamma}_{0ab} \rightarrow \overset{1}{\Gamma}_{0ab} - \frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{\dot{a}}{a^2} \delta_{ab} . \quad (29.27)$$

The transformation of the tetrad-frame connections under coordinate transformations can be checked another way. According to the rule established in §26.7, the change in a quantity under an infinitesimal coordinate gauge transformation equals minus its Lie derivative $\mathcal{L}_\epsilon$ with respect to the infinitesimal coordinate shift ϵ . Any quantity that vanishes in the unperturbed background has, to linear order, vanishing Lie derivative, so is coordinate gauge-invariant. Thus the perturbations $\overset{1}{\Gamma}_{0a0}$, $\overset{1}{\Gamma}_{ab0}$, and $\overset{1}{\Gamma}_{abc}$ are coordinate gauge-invariant, confirming the previous conclusion. The only tetrad-frame connections that are finite in the unperturbed background, and are therefore not coordinate gauge-invariant, are Γ_{0ab} . Although tetrad-frame connections are generically not tetrad-frame tensors, the unperturbed connection $\overset{0}{\Gamma}_{0ab} \equiv -(\dot{a}/a^2)\delta_{ab}$, equation (29.23), is a tetrad-frame tensor, because the spatial unit matrix δ_{ab} can be expressed as the tensor $u_m u_n + \eta_{mn}$, where u_m is the tetrad-frame 4-velocity of the Lorentz-transformed tetrad frame relative to the rest tetrad frame. The tetrad-frame connections Γ_{0ab} transform as

$$\overset{1}{\Gamma}_{0ab} \rightarrow \overset{1}{\Gamma}_{0ab} - \mathcal{L}_\epsilon \Gamma_{0ab} , \quad \mathcal{L}_\epsilon \Gamma_{0ab} = \epsilon^k \partial_k \Gamma_{0ab} = \frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{\dot{a}}{a^2} \delta_{ab} , \quad (29.28)$$

in agreement with the transformation (29.27).

29.5.3 Tetrad-frame Einstein tensor

The tetrad-frame Einstein tensor G_{mn} follows from the usual formulae (11.62), (11.79), and (11.81). The unperturbed tetrad-frame Einstein tensor $\overset{0}{G}_{mn}$ is (equations (29.29) differ from equations (10.29) because the time coordinate here is the conformal time η , not the cosmic time t)

$$\overset{0}{G}_{00} = 3 \frac{\dot{a}^2}{a^4} , \quad (29.29a)$$

$$\overset{0}{G}_{0a} = 0 , \quad (29.29b)$$

$$\overset{0}{G}_{ab} = \left(-2 \frac{\ddot{a}}{a^3} + \frac{\dot{a}^2}{a^4} \right) \delta_{ab} . \quad (29.29c)$$

The perturbation $\overset{1}{G}_{mn}$ of the tetrad-frame Einstein tensor is

$$\overset{1}{G}_{00} = \frac{1}{a^2} \left[-6 \underset{\text{scalar}}{\frac{\dot{a}}{a} F + 2 \nabla^2 \Phi} \right], \quad (29.30a)$$

$$\overset{1}{G}_{0a} = \frac{1}{a^2} \left[2 \underset{\text{scalar}}{\nabla_a \left(F + \left(\frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \tilde{w} \right)} + \underset{\text{vector}}{\frac{1}{2} \nabla^2 W_a + 2 \left(\frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \tilde{w}_a} \right], \quad (29.30b)$$

$$\begin{aligned} \overset{1}{G}_{ab} = \frac{1}{a^2} \left[\left(2 \underset{\text{scalar}}{\left(\frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) F + 2 \left(\frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \psi} \right) \delta_{ab} - (\nabla_a \nabla_b - \underset{\text{scalar}}{\delta_{ab} \nabla^2}) (\Psi - \Phi) \right. \\ \left. + \underset{\text{vector}}{\frac{1}{2} \left(\frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) (\nabla_a W_b + \nabla_b W_a)} - \underset{\text{tensor}}{\left(\frac{\partial^2}{\partial \eta^2} + 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} - \nabla^2 \right) h_{ab}} \right]. \end{aligned} \quad (29.30c)$$

According to the rule established in §26.7, the variation of the Einstein tensor under a coordinate transformation equals minus its Lie derivative,

$$\overset{1}{G}_{mn} \rightarrow \overset{1}{G}_{mn} - \mathcal{L}_\epsilon G_{mn}. \quad (29.31)$$

Consequently, as with the tetrad-frame connections, the tetrad-frame Einstein components that vanish in the background, namely the off-diagonal components G_{mn} with $m \neq n$, are coordinate gauge-invariant, while the components that are finite in the background, namely the diagonal components G_{mn} with $m = n$, are not coordinate gauge-invariant. The variations of the non-coordinate-gauge-invariant Einstein components under an infinitesimal coordinate transformation (26.9) are

$$\mathcal{L}_\epsilon G_{00} = \epsilon^k \partial_k G_{00} = -\frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{3\dot{a}^2}{a^4}, \quad (29.32a)$$

$$\mathcal{L}_\epsilon G_{ab} = \epsilon^k \partial_k G_{ab} = \frac{\epsilon_0}{a} \frac{d}{d\eta} \left(\frac{2\ddot{a}}{a^3} - \frac{\dot{a}^2}{a^4} \right) \delta_{ab}. \quad (29.32b)$$

It can be checked that the same transformations of the tetrad-frame Einstein components under a coordinate transformation follow from the expressions (29.30) for the perturbed Einstein components and the coordinate transformations (29.14) of the potentials.

The time-time and space-space perturbations $\overset{1}{G}_{00}$ and $\overset{1}{G}_{ab}$ are tetrad gauge-invariant, as follows from the fact that these components depend only on symmetric combinations of the vierbein potentials. However, the time-space perturbations $\overset{1}{G}_{0a}$ are not tetrad gauge-invariant, as is evident from the fact that equation (29.30b) involves the non-tetrad-gauge-invariant perturbations $\tilde{w}$ and $\tilde{w}_a$. Physically, under a tetrad boost by a velocity v of linear order, the time-space components G_{0a} change by first order v , but G_{00} and G_{ab} change only to second order v^2 . Thus to linear order, only G_{0a} changes under a tetrad boost. Note that G_{0a} changes under a tetrad boost ($\tilde{w}$ and $\tilde{w}_a$), but not under a tetrad rotation ($\tilde{h}$ and $\tilde{h}_a$).

29.5.4 Tetrad-frame energy-momentum tensor

The unperturbed components $\overset{0}{T}{}^{mn}$ of the tetrad-frame energy-momentum comprise the energy density $\bar{\rho}(\eta)$ and isotropic pressure $\bar{p}(\eta)$ of the FLRW background,

$$\overset{0}{T}{}^{00} \equiv \bar{\rho} , \quad (29.33a)$$

$$\overset{0}{T}{}^{0a} \equiv 0 , \quad (29.33b)$$

$$\overset{0}{T}{}^{ab} \equiv \bar{p} \delta_{ab} . \quad (29.33c)$$

The unperturbed energy-momentum satisfies the conservation equations

$$D_m \overset{0}{T}{}^{m0} = \frac{1}{a} \left[\frac{\partial \bar{\rho}}{\partial \eta} + 3(\bar{\rho} + \bar{p}) \frac{\dot{a}}{a} \right] = 0 \quad (29.34a)$$

$$D_m \overset{0}{T}{}^{ma} = 0 . \quad (29.34b)$$

The perturbed components T^{mn} of the tetrad-frame energy-momentum are the energy density ρ , the energy flux f_a , and the pressure p_{ab} ,

$$T^{00} \equiv \rho = \bar{\rho} + \delta\rho , \quad (29.35a)$$

$$T^{0a} \equiv f_a , \quad (29.35b)$$

$$T^{ab} \equiv p_{ab} = \bar{p} \delta_{ab} + \delta p_{ab} . \quad (29.35c)$$

In perturbation theory, the perturbations $\delta\rho$, f_a , and δp_{ab} are treated as of linear order. The trace of the spatial energy-momentum defines the isotropic pressure p ,

$$\frac{1}{3} T^a_a = p = \bar{p} + \delta p . \quad (29.36)$$

The equations of conservation of energy and momentum are to linear order

$$D_m T^{m0} = \frac{1-\psi}{a} \left[\frac{\partial \rho}{\partial \eta} + \nabla_a f_a + (\rho + p) \left(3 \frac{\dot{a}}{a} - 3\dot{\phi} - \nabla^2 w \right) \right] = 0 , \quad (29.37a)$$

$$D_m T^{ma} = \frac{1}{a} \left[\frac{\partial f_a}{\partial \eta} + 4 \frac{\dot{a}}{a} f_a + \nabla_b p_{ab} + (\rho + p) \nabla_a \psi - \left(\frac{\partial p}{\partial \eta} + (\rho + p) \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) \right) (\nabla_a \tilde{w} + \tilde{w}_a) \right] = 0 . \quad (29.37b)$$

Any fluid component that conserves energy-momentum satisfies equations similar to (29.37). For a fluid component with equation of state $p/\rho = w = \text{constant}$, the unperturbed energy conservation equation (29.34a) recovers the usual result that $\bar{\rho} \propto a^{-3(1+w)}$.

29.5.5 Primordial curvature fluctuation

According to the rule established in §26.7, the change in any quantity A under an infinitesimal coordinate shift ϵ^μ equals minus its Lie derivative $\mathcal{L}_\epsilon A$. If A is a tetrad-frame tensor, then its Lie derivative reduces to $\epsilon^k \partial_k A$. Thus any tetrad-frame perturbation $\overset{1}{A}$ transforms under an infinitesimal coordinate shift as

$$\overset{1}{A} \rightarrow \overset{1}{A} - \mathcal{L}_\epsilon A = \overset{1}{A} - \epsilon^k \partial_k A . \quad (29.38)$$

The perturbation $\overset{1}{A}$ of any tetrad-frame tensor A that vanishes in the background is automatically coordinate gauge-invariant. If a tetrad-frame tensor A does not vanish in the background, then

$$\frac{\overset{1}{A}}{\mathcal{L}_\epsilon \bar{A}} \quad (29.39)$$

is coordinate gauge-invariant (because it equals the coordinate gauge-invariant quantity $1 + (\overset{1}{A} - \mathcal{L}_\epsilon \bar{A})/\mathcal{L}_\epsilon \bar{A}$).

In an FLRW universe, the only components of the energy-momentum tensor that do not vanish in the background are its diagonal components, the density ρ and pressure p . The only derivative that does not vanish in the background is the time derivative of the density $\bar{\rho}$, equations (29.37), which is proportional to $\bar{\rho} + \bar{p}$. A gauge-invariant density perturbation δ can be defined to be

$$\delta \equiv \frac{\delta \rho}{\bar{\rho} + \bar{p}} . \quad (29.40)$$

Outside the horizon, the density perturbation δ satisfies an equation that looks like conservation of entropy (provided that the density does not exchange entropy with other species; but total entropy is always conserved, as in electron-positron annihilation to photons, §10.15):

$$d\delta + 3d \ln a = 0 . \quad (29.41)$$

But outside the horizon, a density perturbation cannot be distinguished from a global rescaling of the cosmic scale factor a , equation (29.18a), for which $3d \ln a = -3\Phi$. A density perturbation that remains gauge-invariant both outside and inside the horizon is the **primordial curvature fluctuation** ζ defined by

$$\zeta \equiv \frac{1}{3}\delta - \Phi . \quad (29.42)$$

Both δ and Φ are gauge-invariant, but only the combination ζ remains invariant outside the horizon under global coordinate transformations.

Concept question 29.2 How can the primordial curvature fluctuation be constant outside the horizon, rather than redshifting to nothing?

29.6 Gauge choices

Since only the 6 tetrad and coordinate gauge-invariant potentials Ψ , Φ , W_a , and h_{ab} have physical significance, it is legitimate to choose a particular **gauge**, a set of conditions on the non-gauge-invariant potentials, arranged to simplify the equations, or to bring out some physical aspect.

This book for the most part uses the conformal Newtonian gauge, §29.8, which is constructed so as to retain only physical perturbations.

29.7 ADM gauge choices

The ADM (3+1) formalism, Chapter 17, chooses the tetrad time axis γ_0 to be orthogonal to hypersurfaces of constant time, $\eta = \text{constant}$, equivalent to requiring that the tetrad time axis be orthogonal to each of the spatial coordinate axes, $\gamma_0 \cdot e_a = 0$, equation (17.2). The ADM choice is equivalent to setting

$$\tilde{w} = \tilde{w}_a = 0 . \quad (29.43)$$

The ADM choice simplifies the tetrad-frame connections (29.24) and the time-space component G_{0a} of the tetrad-frame Einstein tensor, equation (29.30b). The ADM lapse α and shift β^α are

$$\alpha = a(1 + \psi) , \quad \beta^\alpha = \nabla_\alpha w + w_\alpha . \quad (29.44)$$

Another gauge choice that significantly simplifies the tetrad connections (29.24), though does not affect the Einstein tensor (29.30), is

$$\tilde{h} = \tilde{h}_a = 0 . \quad (29.45)$$

If the wavevector k is taken along the coordinate z -direction, then the gauge choice $\tilde{h}_a = 0$ is equivalent to choosing the tetrad 3-axis (z -axis) γ_3 to be orthogonal to the coordinate x and y -axes, $\gamma_3 \cdot e_x = \gamma_3 \cdot e_y = 0$. The gauge choice $\tilde{h} = 0$ is equivalent to rotating the tetrad axes about the 3-axis (z -axis) so that $\gamma_1 \cdot e_y = \gamma_2 \cdot e_x$.

29.8 Conformal Newtonian (Copernican) gauge

The most physical gauge is one in which the 6 perturbations retained coincide with the 6 coordinate and tetrad gauge-invariant perturbations (29.16). This gauge is called **conformal Newtonian gauge**, analogously to the Newtonian gauge of Minkowski space, §27.8. Because in conformal Newtonian gauge the perturbations are precisely the physical perturbations, if the perturbations are physically weak (small), then the perturbations in conformal Newtonian gauge will necessarily be small.

I think conformal Newtonian gauge should be called conformal Copernican gauge, for the same reason that Newtonian gauge should be called Copernican gauge, §27.8. Dynamically, collapsed systems such as galaxies or solar systems are highly nonlinear systems, but gravitationally they are weakly perturbed systems. Conformal Newtonian (Copernican) gauge keeps the coordinates aligned with the unperturbed FLRW comoving coordinates even in highly nonlinear systems. Conformal Newtonian gauge breaks down only in gravitationally nonlinear systems such as black holes.

Conformal Newtonian (Copernican) gauge in an FLRW background makes the same gauge choices as Newtonian gauge in a Minkowski background, equation (27.58),

$$w = \tilde{w} = \tilde{w}_a = h = \tilde{h} = h_a = \tilde{h}_a = 0 , \quad (29.46)$$

so that the retained perturbations are the 6 coordinate and tetrad gauge-invariant perturbations (29.16),

$$\begin{array}{c} \Psi \\ \text{scalar} \end{array} = \psi , \quad (29.47a)$$

$$\begin{array}{c} \Phi \\ \text{scalar} \end{array} = \phi , \quad (29.47b)$$

$$\begin{array}{c} W_a \\ \text{vector} \end{array} = w_a , \quad (29.47c)$$

$$\begin{array}{c} h_{ab} \\ \text{tensor} \end{array} . \quad (29.47d)$$

In a frame where the wavevector $\mathbf{k}$ is along the z -direction, the perturbations (29.47) are essentially the same as those (27.60) in a Minkowski background. In conformal Newtonian gauge, the quantity F defined by equation (29.25) becomes the coordinate and tetrad gauge-invariant quantity

$$F \equiv \frac{\dot{a}}{a} \Psi + \dot{\Phi} . \quad (29.48)$$

The conformal Newtonian metric is

$$ds^2 = a^2 \{ - (1 + 2\Psi) d\eta^2 - 2W_a d\eta dx^a + [\delta_{ab}(1 - 2\Phi) - 2h_{ab}] dx^a dx^b \} . \quad (29.49)$$

The tetrad-frame 4-velocity of a person at rest in the tetrad frame is by definition $u^m = \{1, 0, 0, 0\}$, and the corresponding coordinate 4-velocity u^μ is, in conformal Newtonian gauge,

$$u^\mu = e_0^\mu = \frac{1}{a} \{1 - \Psi, W_a\} . \quad (29.50)$$

Various tetrad-frame connections Γ_{kmn} , equations (29.24), define the acceleration $K_a \equiv \Gamma_{a00}$ and extrinsic curvature $K_{ab} \equiv \Gamma_{a0b} = -\Gamma_{0ab}$. The trace, antisymmetric, and traceless symmetric parts of the extrinsic curvature define the expansion, vorticity, and shear, equations (18.17), which play a key role in the Raychaudhuri equations, §18.2. Also relevant is the precession $\Gamma_{ab0} = -\Gamma_{ba0}$ (not to be confused with the vorticity). In conformal Newtonian gauge, the acceleration, expansion, vorticity, shear, and precession are

$$\text{acceleration } \kappa_a \equiv \Gamma_{a00} = \frac{1}{a} \nabla_a \Psi , \quad (29.51a)$$

$$\text{expansion } \vartheta \equiv -\Gamma_{0a}^a = \frac{1}{a} \left(\frac{\dot{a}}{a} - \frac{F}{\text{scalar}} \right) , \quad (29.51b)$$

$$\text{vorticity } \varpi_{ab} \equiv -\Gamma_{0[ab]} = 0 , \quad (29.51c)$$

$$\text{shear } \sigma_{ab} \equiv -\Gamma_{0(ab)} - \frac{1}{3} \vartheta \delta_{ab} = \frac{1}{a} \left(\frac{1}{2} (\nabla_a W_b + \nabla_b W_a) - \frac{\dot{h}_{ab}}{\text{tensor}} \right) , \quad (29.51d)$$

$$\text{precession } \equiv \Gamma_{ab0} = \frac{1}{2a} (\nabla_a W_b - \nabla_b W_a) . \quad (29.51e)$$

29.8.1 Conformal Newtonian gauge: energy-momentum conservation

In conformal Newtonian gauge, the equations of conservation of energy and momentum are to linear order, from equations (29.37),

$$D_m T^{m0} = \frac{1-\Psi}{a} \left[\frac{\partial \rho}{\partial \eta} + \nabla_a f_a + 3(\rho+p) \left(\frac{\dot{a}}{a} - \dot{\Phi} \right) \right] = 0 , \quad (29.52a)$$

$$D_m T^{ma} = \frac{1}{a} \left[\frac{\partial f_a}{\partial \eta} + 4 \frac{\dot{a}}{a} f_a + \nabla_b p_{ab} + (\rho+p) \nabla_a \Psi \right] = 0 . \quad (29.52b)$$

Notice that the energy-momentum conservation equations (29.52) involve only the scalar potentials Ψ and Φ , not the vector or tensor potentials W_a or h_{ab} . The energy equation (29.52a) has only a scalar component, while the momentum equation (29.52b) has both scalar and vector components, Exercise 29.3.

Exercise 29.3 **Scalar, vector, tensor components of energy-momentum conservation.** What are the scalar, vector, and tensor components of the energy-momentum conservation equations (29.52) in conformal Newtonian gauge?

Solution. The energy conservation equation (29.52a) contains only scalar components. The momentum conservation equation (29.52b) contains scalar and vector components, but no tensor component. The scalar component of the pressure is the sum of an isotropic part $p \delta_{ab}$, and a traceless quadrupole part which is discussed in §32.6. The vector component of the pressure takes the form

$$p_{ab}^{\text{vector}} = \nabla_a p_{\perp,b} + \nabla_b p_{\perp,a} , \quad (29.53)$$

where $p_{\perp,a}$ is transverse, $\nabla_a p_{\perp,a} = 0$. The vector part of the momentum conservation equation (29.52b) is

$$D_m T_{\text{vector}}^{ma} = \frac{1}{a} \left(\frac{\partial f_{\perp,a}}{\partial \eta} + 4 \frac{\dot{a}}{a} f_{\perp,a} + \nabla^2 p_{\perp,a} \right) = 0 . \quad (29.54)$$

If the vector pressure is negligible, $p_{\perp,a} = 0$, then the vector momentum conservation equation (29.54) implies that

$$f_{\perp,a} \propto a^{-4} . \quad (29.55)$$

The tensor component p_{ab}^T of the pressure is traceless and transverse. Being traceless, p_{ab}^T makes no contribution to the isotropic pressure p , and being transverse, it satisfies $\nabla_b p_{ab}^T = 0$. Consequently the energy-momentum conservation equations (29.52) contain no tensor component.

29.8.2 Conformal Newtonian gauge: scalar Einstein equations

In conformal Newtonian gauge, the scalar perturbations of the Einstein equations are, from the expressions (29.30) for the Einstein tensor, the energy density, energy flux, monopole pressure, and quadrupole

pressure equations,

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 \dot{T}^{00} , \quad (29.56a)$$

$$ikF = 4\pi G a^2 \hat{k}_a T^{0a} , \quad (29.56b)$$

$$\dot{F} + 2 \frac{\dot{a}}{a} F + \left(\frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \Psi - \frac{k^2}{3} (\Psi - \Phi) = \frac{4}{3} G \pi a^2 \delta_{ab} \dot{T}^{ab} , \quad (29.56c)$$

$$k^2 (\Psi - \Phi) = 8\pi G a^2 \left(\frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) T^{ab} . \quad (29.56d)$$

The perturbation overscript 1 has been omitted from the right hand sides of equations (29.56b) and (29.56d) since the unperturbed energy-momentum vanishes for these components. All 4 of the scalar Einstein equations (29.56) are expressed in terms of gauge-invariant variables, and are therefore fully gauge-invariant.

If the energy-momentum tensors of the various matter components are arranged so as to conserve overall energy-momentum, as they should, then 2 of the 4 equations (29.56a)–(29.56d) are redundant, since they serve simply to enforce conservation of energy and scalar momentum. Usually the 1st equation, the energy equation (29.56a), and the 4th equation, the quadrupole pressure equation (29.56d), are most convenient to retain. But sometimes the 2nd equation, the scalar momentum equation (29.56b), is more convenient in place of the energy equation (29.56a).

29.8.3 Conformal Newtonian gauge: vector Einstein equations

The vector (spin-1) Einstein equations in conformal Newtonian gauge are, from the expressions (29.30) for the Einstein tensor,

$$\nabla^2 W_a = -16\pi G a^2 T_{\text{vector}}^{0a} , \quad (29.57a)$$

$$\left(\frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) (\nabla_a W_b + \nabla_b W_a) = 16\pi G a^2 T_{\text{vector}}^{ab} . \quad (29.57b)$$

If the overall matter energy-momentum is conserved, as it must be, then either equation (29.57a) or equation (29.57b) can be discarded as redundant, since the two equations together serve to enforce conservation of (the vector components of) overall energy-momentum.

In the absence of a vector source of pressure, $T_{\text{vector } ab} = p_{\text{vector } ab} = 0$, the Einstein equation (29.57b) ensures that the vector perturbation redshifts as a^{-2} ,

$$W_a \propto a^{-2} \quad \text{if } T_{\text{vector}}^{ab} = 0 . \quad (29.58)$$

The same conclusion follows from the other vector Einstein equation (29.57a). If the vector pressure vanishes, then the vector momentum conservation equation (29.54) ensures that the vector energy flux $T_{\text{vector}}^{0a} = f_{\perp, a}$ redshifts as $f_{\perp, a} \propto a^{-4}$, which when plugged into the Einstein equation (29.57a) implies $W_a \propto a^{-2}$.

In practice, collisions in the early post-inflation Universe tend to isotropize particle distributions, driving

not only the pressure but also the bulk velocity to zero, as discussed in more detail in §35.11. If the bulk velocity vanishes, so $f_a = 0$, then the Einstein equation (29.57a) forces the vector potential to vanish, $W_a = 0$.

The tendency of vector perturbations to redshift away has the consequence that vector perturbations are usually negligible in standard cosmological models.

Concept question 29.4 Can the Universe at large rotate? Is it possible for a Universe to rotate globally? **Answer.** Yes, but only to second order in perturbation theory. Consider a line-element which is FLRW except for a time-dependent but spatially uniform rotation by angular velocity $w(\eta)$ in the azimuthal direction, and a time-dependent but spatially uniform vorticity term $\tilde{w}(\eta)$ in the same azimuthal direction,

$$ds^2 = a^2 \left\{ -(d\eta + x^2 \sin^2\theta \tilde{w} d\phi)^2 + \frac{dx^2}{1 - \kappa x^2} + x^2 [d\theta^2 + \sin^2\theta (d\phi - w d\eta)^2] \right\} . \quad (29.59)$$

To linear order, the two terms combine to yield the vector perturbation

$$2W_a dx^a = x^2 \sin^2\theta (w + \tilde{w}) d\phi . \quad (29.60)$$

If the vorticity term vanishes, $\tilde{w} = 0$, then a coordinate transformation $d\phi - w d\eta \rightarrow d\phi$ eliminates the w term. Thus a w term by itself does not encode a global rotation. If on the other hand the vorticity term does not vanish, $\tilde{w} \neq 0$, then the same coordinate transformation brings the time 1-form to

$$d\eta + x^2 \sin^2\theta \tilde{w} d\phi \rightarrow (1 + x^2 \sin^2\theta w \tilde{w}) d\eta + x^2 \sin^2\theta \tilde{w} d\phi . \quad (29.61)$$

The vorticity term $\tilde{w}$ cannot be removed by any coordinate transformation. The best that can be done is to remove, to linear order, the vector perturbation (29.60) by a coordinate transformation $d\phi - w d\eta \rightarrow d\phi + \tilde{w} d\eta$, which transforms the time 1-form to

$$d\eta + x^2 \sin^2\theta \tilde{w} d\phi \rightarrow (1 + x^2 \sin^2\theta \tilde{w} (w + \tilde{w})) d\eta + x^2 \sin^2\theta \tilde{w} d\phi . \quad (29.62)$$

The conclusion is that a global rotation is possible, but it is detectable only at second order in perturbation theory. The size of the vorticity is constrained by the physical requirement that there be no closed timelike curves. Closed timelike curves occur where the azimuthal metric component $g_{\phi\phi}$, which is gauge-invariant in an azimuthally symmetric spacetime as here, is negative. The azimuthal metric component $g_{\phi\phi}$ in the line-element (29.59) is

$$g_{\phi\phi} = a^2 x^2 \sin^2\theta (1 - x^2 \sin^2\theta \tilde{w}^2) , \quad (29.63)$$

which must be positive to avoid closed timelike curves. The strongest constraint is where $x \sin\theta$ is largest, which is at the comoving horizon distance $x = \eta$, and at the equator $\theta = \pi/2$ of the putative vortical rotation. Thus the vorticity $\tilde{w}_0$ at the present time must satisfy $|\tilde{w}_0| < 1/\eta_0$. In the absence of a mechanism to generate large-scale vector modes, vorticity declines with cosmic scale factor as $\tilde{w} \propto a^{-2}$, equation (29.58), so the constraint is more stringent at early times. The result is that the absence of closed timelike curves requires the vorticity $\tilde{w}_0$ today to satisfy

$$|\tilde{w}_0| < a^2/\eta , \quad (29.64)$$

where the right hand side is smallest at the earliest times. Note that the comoving horizon distance evolves as $\eta \propto a^{-1}$, a , and $a^{1/2}$ during vacuum-, radiation-, and matter-dominated regimes.

29.8.4 Conformal Newtonian gauge: tensor Einstein equations

The tensor (spin-2) Einstein equations in conformal Newtonian gauge are, from the expressions (29.30) for the Einstein tensor,

$$\left(\frac{\partial^2}{\partial \eta^2} + 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} - \nabla^2 \right) h_{ab} = -8\pi G a^2 T_{\text{tensor}}^{ab} . \quad (29.65)$$

Whereas vector perturbations necessarily redshift as $W_a \propto a^{-2}$ in the absence of a source, tensor perturbations h_{ab} at superhorizon wavelengths $k\eta \ll 1$ have a solution where they are constant,

$$h_{ab} = \text{constant} \quad \text{for } k\eta \ll 1 \quad \text{if } T_{\text{tensor}}^{ab} = 0 . \quad (29.66)$$

Inflation generates tensor modes, which describe gravitational waves. In contrast to vector modes, long wavelength gravitational waves generated during inflation can survive to the present time. Gravitational waves leave an observable imprint in the B -mode polarization of the cosmic microwave background. A detection of B -mode polarization was claimed by the BICEP2 collaboration (Ade et al., 2014), but the signal may have been from aligned galactic dust rather primordial (Planck Collaboration, 2015). The cosmic gravitational wave background could potentially be observed directly in the future.

29.9 Energy-momentum carried by gravitational waves

The calculation of the energy-momentum carried by gravitational waves in an FLRW background proceeds similarly to that in a Minkowski background, §27.14. With zero contribution to the energy-momentum on the right hand side of the tensor Einstein equation (29.65), it might seem that gravitational waves carry zero energy momentum. However, the linear order tensor equation (29.65) with zero source can be interpreted as a wave equation for the tensor perturbation h_{ab} , and then the second order Einstein tensor, which is quadratic in h_{ab} , can be interpreted as the energy-momentum in gravitational waves, when taken over to the right hand side of the Einstein equation. The parts of the second-order perturbation to the Einstein tensor that depend on the tensor perturbation h_{ab} are, similarly to equation (27.125) in a Minkowski background,

$$\bar{G}_{mn} = -(\nabla_m h_{ab})(\nabla_n h^{ab}) + \frac{1}{4} \eta_{mn} \square h^2 , \quad (29.67)$$

where $\square \equiv D_m D^m$ is covariant d'Alembertian operator, and

$$h^2 \equiv h_{ab} h^{ab} = 2(h_+^2 + h_\times^2) = 2h_{++}h_{--} . \quad (29.68)$$

In an FLRW background, $\square h^2$ is

$$\square h^2 = \frac{1}{a^2} \left(-\frac{\partial^2}{\partial \eta^2} - 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} + \nabla^2 \right) h^2 . \quad (29.69)$$

As in Minkowski space, the second order contribution $\bar{G}_{mn}^2$ to the Einstein tensor is taken over to the right hand side of the Einstein equations, and re-interpreted as the energy-momentum T_{mn}^{gw} in gravitational waves

$$T_{mn}^{\text{gw}} \equiv \frac{1}{8\pi} [(\nabla_m h_{ab})(\nabla_n h^{ab}) - \frac{1}{4}\eta_{mn}\Box h^2] . \quad (29.70)$$

If the perturbation h_{ab} oscillates rapidly compared to the timescale for the cosmic scale factor a to vary, then the $\Box h^2$ term tend to average out. However, the $\Box h^2$ term is still needed to ensure that the energy-momentum T_{mn}^{gw} in gravitational waves satisfies covariant conservation of energy-momentum,

$$D^m T_{mn}^{\text{gw}} = 0 . \quad (29.71)$$

Concept question 29.5 Energy-momentum in gravitational waves. Assigning energy-momentum to gravitational waves by taking terms in the Einstein tensor over to the right hand side of the Einstein equations is reminiscent of the device of interpreting the curvature terms $\propto \kappa/a^2$ in the Einstein tensor (10.29) for the FLRW metric as energy-momentum, equation (10.49). It is also reminiscent of Einstein's introduction of a cosmological constant Λ in the Hilbert action, which nowadays most cosmologists would interpret as some kind of vacuum energy. Is there some way to distinguish whether energy-momentum is or is not “real?” Discuss.

Exercise 29.6 Evolution of tensor perturbations (gravitational waves) in FLRW spacetimes. Show that equation (29.65) can be rewritten in Fourier space

$$\left(\frac{\partial^2}{\partial \eta^2} - \frac{\ddot{a}}{a} + k^2 \right) (ah_{ab}) = -8\pi G a^3 T_{\text{tensor}}^{ab} . \quad (29.72)$$

What is the solution of equation (29.72) if there is no tensor source, $T_{\text{tensor}}^{ab} = 0$, and the background energy-momentum is dominated by a species with equation of state $p/\rho = w = \text{constant}$? Plot the solution in the radiation-dominated regime, subject to the condition that h_{ab} is initially finite.

Solution. The conformal factor η is related to cosmic scale factor a by equation (10.84), which for a background energy-momentum dominated by a single species with $p/\rho = w = \text{constant}$ integrates to a power-law (10.85), with η running over $[0, \infty]$ if $w > -\frac{1}{3}$ (decelerating), or $[-\infty, 0]$ if $w < -\frac{1}{3}$ (accelerating), as a runs over $[0, \infty]$. The power law (10.85) may also be written

$$a \propto \eta^{n+\frac{1}{2}} , \quad (29.73)$$

where

$$n \equiv \frac{3(1-w)}{2(1+3w)} . \quad (29.74)$$

Then

$$\frac{\dot{a}}{a} = \frac{2}{(1+3w)\eta} = \frac{n+\frac{1}{2}}{\eta} , \quad \frac{\ddot{a}}{a} = \frac{2(1-3w)}{(1+3w)^2\eta^2} = \frac{n^2 - \frac{1}{4}}{\eta^2} . \quad (29.75)$$

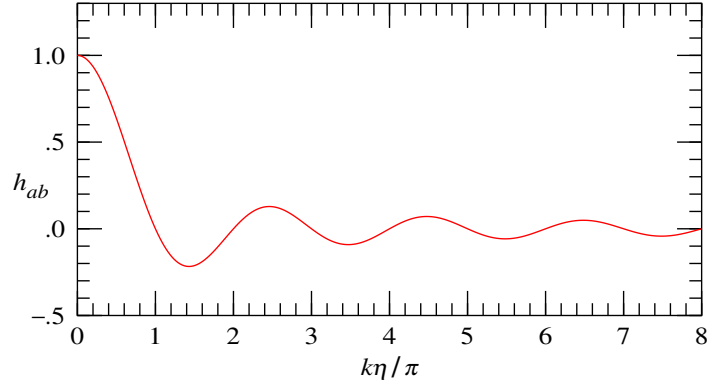


Figure 29.1 Evolution of the tensor potential h_{ab} in the radiation-dominated regime, where $w = 1/3$, $n = 1/2$, and $h_{ab} \propto \sin(k\eta)/(k\eta)$.

The tensor evolution equation (29.72) in the absence of sources becomes

$$\left[\frac{\partial^2}{\partial \eta^2} - \frac{n^2 - \frac{1}{4}}{\eta^2} + k^2 \right] (ah_{ab}) = 0 , \quad (29.76)$$

The solution of equation (29.76) is a linear combination of Bessel functions $J_{\pm n}(k|\eta|)$ of argument $\pm n$,

$$h_{ab} = (k|\eta|)^{-n} [A_+ J_n(k|\eta|) + A_- J_{-n}(k|\eta|)] . \quad (29.77)$$

Examples are

$$n = \begin{cases} \frac{1}{2} & w = \frac{1}{3} , \\ \frac{3}{2} & w = 0 , \\ -\frac{3}{2} & w = -1 , \end{cases} \quad (29.78)$$

in which case the solutions happen to be spherical Bessel functions, where n is half-integral. The solution that is finite at $|\eta| \rightarrow 0$ (corresponding to $a \rightarrow 0$ for $w > -\frac{1}{3}$, or $a \rightarrow \infty$ for $w < -\frac{1}{3}$) is, for n either positive or negative, the $J_n(k|\eta|)$ component, which normalized to 1 at $\eta = 0$ is

$$h_{ab} = \Gamma(n+1) \left(\frac{k|\eta|}{2} \right)^{-n} J_n(k|\eta|) \rightarrow \begin{cases} 1 & k|\eta| \ll 1 , \\ \frac{\Gamma(1+n)}{\sqrt{\pi}} \left(\frac{k|\eta|}{2} \right)^{-(n+\frac{1}{2})} \cos [k|\eta| - (n+\frac{1}{2})\pi/2] & k|\eta| \gg 1 . \end{cases} \quad (29.79)$$

The conformal time η is related to a by the proportionality (29.73), so the solution at large $k|\eta|$ goes as

$$h_{ab} \sim a^{-1} \cos(k|\eta| + \text{constant}) \quad k|\eta| \gg 1 . \quad (29.80)$$

Physically, the gravitational wave amplitude h_{ab} is constant well outside the horizon, $k|\eta| \ll 1$, while it

redshifts as $1/a$ well inside the horizon, $k|\eta| \gg 1$. Figure 29.1 illustrates the evolution of the tensor potential h_{ab} in the radiation-dominated regime.

29.10 Spatially flat gauge

One gauge that is advantageous in calculating scalar perturbations that develop during inflation driven by a single scalar field, §38.4, is **spatially flat gauge**, which sets to zero all spatial perturbations except tensor perturbations,

$$\phi = \tilde{w} = \tilde{w}_a = h = \tilde{h} = h_a = \tilde{h}_a = 0 . \quad (29.81)$$

In this gauge, the perturbed spacetime remains spatially homogeneous, which requires that the source of energy-momentum be spatially homogeneous. For inflation driven by a single scalar field φ , the time coordinate can be chosen to be a hypersurface of constant φ , in which case the scalar field is spatially uniform, a function $\varphi(\eta)$ only of time.

The gauge-invariant perturbations (29.16) in spatially flat gauge are

$$\Psi_{\text{scalar}} = \psi - \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) w , \quad (29.82a)$$

$$\Phi_{\text{scalar}} = \frac{\dot{a}}{a} w , \quad (29.82b)$$

$$W_{\text{vector}} = w_a , \quad (29.82c)$$

$$h_{\text{tensor}} = h_{ab} . \quad (29.82d)$$

29.11 Conformal synchronous gauge

One gauge that was important historically, and is still in common use in cosmology, but is not used here, is conformal synchronous gauge, discussed in the case of Minkowski background space in §27.9. The cosmological synchronous gauge choices are the same as for the Minkowski background, equations (27.65) and (27.66):

$$\psi = w = \tilde{w} = w_a = \tilde{w}_a = \tilde{h} = \tilde{h}_a = 0 . \quad (29.83)$$

The gauge-invariant perturbations (29.16) in synchronous gauge are

$$\Psi_{\text{scalar}} = \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) \dot{h}, \quad (29.84a)$$

$$\Phi_{\text{scalar}} = \phi - \frac{\dot{a}}{a} \dot{h}, \quad (29.84b)$$

$$W_a^{\text{vector}} = -\dot{h}_a, \quad (29.84c)$$

$$h_{ab}^{\text{tensor}}. \quad (29.84d)$$

Like synchronous gauge, §27.9, conformal synchronous gauge chooses a coordinate system and tetrad that is attached to the locally inertial frames of freely falling observers. Thus synchronous gauge follows the frame of cold collisionless matter (“dust”). To the extent that non-baryonic cold dark matter has always been cold and collisionless (not quite true, but an excellent approximation), synchronous frame is the frame of non-baryonic cold dark matter.

Conformal synchronous gauge fails at non-linear scales where collisionless matter has turned around and collapsed into galaxies and galaxy clusters. This contrasts with conformal Newtonian gauge, which holds as long as gravitational perturbations remain weak, including in highly non-linear collapsed systems such as galaxies and solar systems.

Concept question 29.7 What frame does the CMB define? Answer. The CMB frame is the frame where the CMB temperature is constant, and CMB photons have zero bulk velocity. That statement depends on the scale (in Fourier space, the wavenumber k) over which the CMB temperature or velocity is averaged. For adiabatic fluctuations at superhorizon scales, all particle species start with essentially the same overdensity and velocity. The initial frame that comoves with particles is, by construction, the synchronous frame. Once a scale comes inside the horizon, different components that are not kept coupled by collisions (non-baryonic dark matter, photons, neutrinos) evolve differently, as illustrated for example by Figure 33.1. At scales well inside the horizon, the bulk velocity of free-streaming relativistic particles in conformal Newtonian gauge tends to zero in oscillatory fashion, again as illustrated by Figure 33.1. Thus at subhorizon scales conformal Newtonian gauge provides a good approximation to the frame of CMB photons.

Concept question 29.8 Are congruences of comoving observers in cosmology hypersurface-orthogonal? Comoving observers are defined to be those at rest in the tetrad frame, $u^m = \{1, 0, 0, 0\}$. The worldlines of comoving observers define a timelike congruence. Are congruences of comoving observers hypersurface-orthogonal, §18.7? **Answer.** Common cosmological gauges, including conformal Newtonian or conformal synchronous, impose the ADM gauge condition that the time axis γ_0 is orthogonal to hypersurfaces of constant time t , §29.7. This ADM condition (coupled with the general relativistic assumption of vanishing torsion) implies that vorticity vanishes, which is one of the two conditions for a timelike congruence to be hypersurface-orthogonal, §18.7. The other condition for a timelike congruence to be hypersurface-orthogonal

is that the congruence be geodesic; this is true in the specific case of conformal synchronous gauge, but not for other gauges, such as conformal Newtonian gauge.

Cosmological perturbations: a simplest set of assumptions

The purpose of this Chapter is to set forward the simplest approximate model of the development of perturbations to matter and radiation in our Universe.

The model consists of two non-interacting perfect fluids, non-baryonic cold dark matter with a pressureless equation of state $p/\rho = 0$, and radiation with a relativistic equation of state $p/\rho = 1/3$. The model neglects baryons, since their energy density is sub-dominant, being $\Omega_b/\Omega_c \approx 1/5$ of the dark matter density. The model lumps neutrinos with photons, neutrinos being relativistic with energy density about two thirds that of photons, $\rho_\nu/\rho_\gamma = 6 \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}/2 \approx 0.68$. It would be wrong to lump baryonic perturbations with those of non-baryonic dark matter, since prior to recombination electron-photon scattering keeps the baryonic fluid tightly coupled to photons, preventing the baryons from clustering gravitationally like the non-baryonic cold dark matter. In the simple approximation, recombination occurs abruptly at a redshift $1 + z_{\text{rec}} \approx 1100$. After recombination, baryons can cluster gravitationally, forming galaxies, stars, and eventually people.

Well after recombination, a third energy component, dark energy, becomes important. It too can be treated as a perfect fluid, with equation of state $p/\rho = -1$.

The perfect fluid approximation keeps only the lowest momentum moments of the particle distributions, the energy density and the bulk velocity, along with an isotropic pressure p that is a given function of density ρ in the rest frame of the fluid. The evolution of a perfect fluid is determined entirely by the energy-momentum conservation equations that the fluid satisfies.

The model includes only scalar modes. The quadrupole pressure vanishes for perfect fluids, so the two scalar potentials are equal, $\Psi = \Phi$, equation (29.56d). However, the two scalar potentials will often be kept separate in this Chapter, to facilitate later reference. Tensor modes (gravitational waves) are neglected, since their energy-momentum is sub-dominant. Tensor modes leave a distinctive imprint on the polarization of the CMB, which is addressed in Chapter 36.

Section 30.7 summarizes the equations for the simplest set of assumptions.

30.1 Perturbed FLRW line-element

The perturbed FLRW line-element in conformal Newtonian gauge, equation (29.49), including only scalar perturbations, is

$$ds^2 = a^2 \left[-(1 + 2\Psi)d\eta^2 + \delta_{ab}(1 - 2\Phi)dx^a dx^b \right] , \quad (30.1)$$

where $a(\eta)$ is the cosmic scale factor, a function only of conformal time η .

30.2 Energy-momenta of perfect fluids

In the simplest approximation, each component of the cosmological energy-momentum, including matter, radiation, and dark energy, can be treated as a perfect fluid, that is, a fluid whose pressure is isotropic in the rest frame of the fluid. The tetrad-frame energy-momentum tensor of a perfect fluid with proper density ρ and isotropic pressure p in its own rest frame, moving with bulk 4-velocity $u^m \equiv dx^m/d\tau$ relative to the conformal Newtonian tetrad frame, is

$$T^{mn} = (\rho + p)u^m u^n + p\eta^{mn} . \quad (30.2)$$

It is a good approximation to assume further that the equation of state of the fluid is such that its proper pressure p is some prescribed function of its proper density ρ (such a fluid is called **barotropic**),

$$p = p(\rho) . \quad (30.3)$$

Define w to be the (not necessarily constant) derivative

$$w \equiv \frac{dp}{d\rho} , \quad (30.4)$$

which proves to be (at least for $w \geq 0$) the square of the sound speed of the fluid in units of the speed of light. In the simple model considered in this Chapter, each of the fluids considered, matter, radiation, and dark energy, has constant w , with $w = 0$, $\frac{1}{3}$, and -1 respectively. Chapter 32 considers the more realistic situation of a photon-baryon fluid with non-constant w .

Each fluid moves with nonrelativistic bulk velocity, including radiation, which is almost isotropic, and therefore has a small bulk velocity even though individual particles of radiation move at the speed of light. The bulk tetrad-frame 4-velocity u^m of the fluid is thus, to linear order

$$u^m = \{1, v_a\} , \quad (30.5)$$

where v_a is its nonrelativistic spatial bulk 3-velocity (the spatial tetrad metric is Euclidean, so $v^a = v_a$). The bulk velocity v_a is to be considered as of linear order, so its square vanishes to linear order.

The proper fluid density ρ can be written as a sum of an unperturbed density $\bar{\rho}$ and a linear order fluctuation $\delta\rho$,

$$\rho = \bar{\rho} + \delta\rho . \quad (30.6)$$

It proves advantageous, because it simplifies the resulting perturbation equations (30.13), to characterize the density fluctuation $\delta\rho$ in terms of the gauge-invariant fluctuation δ defined by equation (29.40),

$$\delta\rho = (\bar{\rho} + \bar{p})\delta , \quad (30.7)$$

where $\bar{p} = p(\bar{\rho})$ in the unperturbed pressure. As you will discover in Exercise 30.1, the fluctuation δ can be interpreted physically as the entropy fluctuation,

$$\delta \equiv \frac{\delta\rho}{\bar{\rho} + \bar{p}} = \frac{\delta s}{\bar{s}} . \quad (30.8)$$

For matter, where $p = 0$, the entropy fluctuation coincides with the density fluctuation, $\delta = \delta\rho/\bar{\rho}$. For dark energy, where $p = -\rho$, the density fluctuation is necessarily zero, $\delta\rho = 0$, reflecting the fact that vacuum energy cannot cluster. To linear order in the bulk velocity v_a , the tetrad-frame energy-momentum tensor (30.2) of the perfect fluid is then

$$T^{00} \equiv \rho = \bar{\rho} + \delta\rho , \quad (30.9a)$$

$$T^{0a} \equiv f_a = (\bar{\rho} + \bar{p})v_a , \quad (30.9b)$$

$$T^{ab} \equiv p\delta_{ab} = (\bar{p} + \delta p)\delta_{ab} , \quad (30.9c)$$

where the pressure fluctuation δp is, from equation (30.4),

$$\delta p = w\delta\rho . \quad (30.10)$$

If a species does not exchange energy or momentum with other species, then it satisfies the energy-momentum conservation equations (29.52) in conformal Newtonian gauge. Subtracting appropriate amounts of the unperturbed energy conservation equation (29.34a) from the perturbed energy-momentum conservation equations (29.52) yields equations for the entropy fluctuation δ and bulk velocity v_a of the fluid (recall that overdot denotes partial differentiation with respect to conformal time η , equation (29.5), so for example $\dot{\delta} = \partial\delta/\partial\eta$),

$$\dot{\delta} + \nabla_a v_a = 3\dot{\Phi} , \quad (30.11a)$$

$$\dot{v}_a + (1 - 3w)\frac{\dot{a}}{a}v_a + w\nabla_a\delta = -\nabla_a\Psi . \quad (30.11b)$$

Physically, equation (30.11a) represents conservation of entropy, while equation (30.11b) represents conservation of momentum.

Now decompose the bulk 3-velocity v_a into its scalar v and vector $v_{\perp,a}$ parts. Up to this point, the scalar part of a vector has been taken to be the gradient of a potential. But here it is advantageous to absorb a factor of k into the definition of the scalar part v of the velocity, so that instead of $v_a = -ik_a v + v_{\perp,a}$ in Fourier space, the velocity is given in Fourier space by

$$v_a = -i\hat{k}_a v + v_{\perp,a} . \quad (30.12)$$

The advantage of this choice is that v is dimensionless, as are δ and Ψ and Φ . Note that the comoving

wavenumber k (a constant for any given mode) has units of η^{-1} . The scalar parts of the perturbation equations (30.11) are then

$$\dot{\delta} - k v = 3\dot{\Phi} , \quad (30.13a)$$

$$\dot{v} + (1 - 3w)\frac{\dot{a}}{a}v + wk\delta = -k\Psi . \quad (30.13b)$$

The vector part of equations (30.11) is considered in Exercise 30.3.

Combining the two equations (30.13) for the scalar fluctuation δ and scalar bulk velocity v yields a second-order differential equation for $\delta - 3\Phi$,

$$\left[\frac{d^2}{d\eta^2} + (1 - 3w)\frac{\dot{a}}{a}\frac{d}{d\eta} + wk^2 \right] (\delta - 3\Phi) = -k^2(\Psi + 3w\Phi) . \quad (30.14)$$

Equation (30.14) holds for any perfect fluid that conserves energy-momentum and that has equation of state (30.4), with w not necessarily constant. For positive w , equation (30.14) is a wave equation for a damped, forced oscillator with sound speed $\sqrt{w}$. The resulting generic behaviour for the particular cases of matter ($w = 0$) and radiation ($w = \frac{1}{3}$) is considered in §30.5 and §30.6 below.

A more careful treatment, deferred to Chapter 33, accounts for the complete momentum distribution of radiation by expanding the temperature perturbation $\Theta \equiv \delta T/\bar{T}$ in multipole moments, equation (33.47). The radiation fluctuation δ_r and scalar bulk velocity v_r are related to the first two multipole moments of the temperature perturbation, the monopole Θ_0 and the dipole Θ_1 , by

$$\delta_r = 3\Theta_0 , \quad (30.15a)$$

$$v_r = 3\Theta_1 . \quad (30.15b)$$

The factor of 3 arises because the unperturbed radiation distribution is in thermodynamic equilibrium, for which the entropy density is $s \propto T^3$, so $\delta_r = 3\delta T/\bar{T}$.

Exercise 30.1 Entropy perturbation. The purpose of this exercise is to discover that the fluctuation δ defined by equation (30.6) can be interpreted as the entropy fluctuation. According to the first law of thermodynamics, the entropy density s of a fluid of energy density ρ , pressure p , and temperature T in a volume V satisfies

$$d(\rho V) + pdV = Td(sV) . \quad (30.16)$$

If the fluid is ideal, so that ρ , p , T , and s are independent of volume V , then integrating the first law (30.16) implies that

$$\rho V + pV = TsV . \quad (30.17)$$

This implies that the entropy density s is related to the other variables by

$$s = \frac{\rho + p}{T} . \quad (30.18)$$

Show that, for a perfect, barotropic fluid (one in which pressure is a prescribed function $p(\rho)$ of density ρ), small variations of the density and entropy are related by

$$\frac{\delta\rho}{\rho+p} = \frac{\delta s}{s} , \quad (30.19)$$

confirming equation (30.8). [Hint: Do not confuse what is being asked here with adiabatic expansion. The result (30.19) is a property of the fluid, independent of whether the fluid is changing adiabatically. For adiabatic expansion, the fluid satisfies the additional condition $sV = \text{constant}$.]

Solution. Use equation (30.18) to eliminate the temperature T from the first law (30.16), obtaining

$$\frac{d\rho}{\rho+p} = \frac{ds}{s} . \quad (30.20)$$

In the situation being considered, where pressure is a prescribed function $p(\rho)$ of density, equation (30.20) implies equation (30.19).

Concept question 30.2 Entropy perturbation when number is conserved. The derivation of the entropy perturbation (30.19) in Exercise 30.1 was based on the first law of thermodynamics (30.16) without any term μdN representing number conservation. Should not such a term be included? **Answer.** This question was addressed in Exercise 10.15. Each chemical potential μ is associated with a conserved species. Terms associated with number conservation can be dropped provided that the fluid contains all particles belonging to a conserved species. For example, electrons and positrons can annihilate with each other, so the numbers N_e and $N_{\bar{e}}$ of electrons e and positrons $\bar{e}$ in a comoving volume are not conserved, but their sum $N_e + N_{\bar{e}}$ is conserved. Electrons and positrons in thermodynamic equilibrium satisfy $\mu_{\bar{e}} = -\mu_e$, so the terms representing number conservation in the combined electron-positron fluid vanish,

$$\mu_e dN_e + \mu_{\bar{e}} dN_{\bar{e}} = \mu_e d(N_e - N_{\bar{e}}) = 0 . \quad (30.21)$$

Thus the entropy perturbation equation (30.19) does not hold individually for electrons and positrons, but it does hold for the combined electron-positron fluid.

Exercise 30.3 Vector fluctuation. What is the vector part of the perturbation equations (30.11)? Solve it.

Solution. The vector part of equations (30.11) is

$$\dot{v}_{\perp,a} + (1-3w)\frac{\dot{a}}{a}v_{\perp,a} = 0 . \quad (30.22)$$

If w is constant, the solution is

$$v_{\perp,a} \propto a^{-(1-3w)} . \quad (30.23)$$

Together with $\bar{\rho} \propto a^{-3(1+w)}$, equation (30.23) implies

$$f_{\perp,a} \equiv (\bar{\rho} + \bar{p})v_{\perp,a} \propto a^{-4} , \quad (30.24)$$

which agrees with the vector component of the momentum conservation equation (29.52b) for any combination of perfect fluids (which have vanishing vector component $p_{\perp,ab}$ of pressure).

30.3 Entropy conservation at superhorizon scales

At superhorizon scales, where $k\eta \ll 1$, the bulk velocity term in equation (30.13a) for the entropy fluctuation δ is negligible, and the equation reduces to

$$\dot{\delta} = 3\dot{\Phi} . \quad (30.25)$$

The quantity $\zeta \equiv \frac{1}{3}\delta - \Phi$, the primordial curvature fluctuation defined by equation (29.42), has the property that it is constant at large scales in any fluid component that does not exchange energy with other components,

$$\zeta = \text{constant} \quad \text{if } k\eta \ll 1 . \quad (30.26)$$

Physically, the constancy of ζ at superhorizon scales is associated with a conservation law that has the appearance of a law of conservation of entropy. Recall that in an FLRW universe, the Einstein equations enforce a conservation law (10.34) that looks like the first law of thermodynamics with conserved entropy. The constancy of ζ is a generalization of this law to superhorizon perturbations of an FLRW universe. An observer cannot distinguish a superhorizon perturbation from a strictly FLRW universe (such perturbations can be measured only by later observers after the superhorizon perturbation has entered their horizon). Specifically, an observer inside a horizon patch can perform a global transformation (29.18) of the cosmic scale factor a (and time coordinate η) so as to set the large-scale potential Φ (and Ψ) to zero in their patch. Then equation (30.25) becomes $\dot{\delta} = 0$, expressing the first law of thermodynamics (10.34) in the FLRW background of the patch.

The -3Φ part of the conserved fluctuation $\delta - 3\Phi$ is associated with the transformation between comoving and proper volumes, and the fact that the proper spatial volume element is $a^3(1 - 3\Phi)d^3x$ ¹²³ (which remains true when not only scalar but also vector and tensor fluctuations are included).

Exercise 30.4 Relation between entropy and the primordial curvature fluctuation ζ . Assume that the proper pressure $p(\rho)$ is a definite function of proper density ρ . Define entropy s per unit volume by (see Exercise 30.1)

$$\ln s \equiv \int \frac{d\rho}{\rho + p} . \quad (30.27)$$

Confirm that, if the bulk peculiar velocity can be neglected so that the energy flux is zero, $f_a = 0$, as is true at superhorizon scales, then the energy conservation equation (29.52a) in Newtonian gauge reduces to

$$d \ln s + 3 d \ln [a(1 - \Phi)] = 0 , \quad (30.28)$$

whose unperturbed part is

$$d \ln \bar{s} + 3 d \ln a = 0 . \quad (30.29)$$

Conclude that energy conservation implies the conservation of ζ defined by

$$\zeta = \frac{1}{3} \ln(s/\bar{s}) - \Phi . \quad (30.30)$$

Concept question 30.5 If the Friedmann equations enforce conservation of entropy, where does the entropy of the Universe come from? Friedmann's equations enforce conservation of entropy, equation (10.34). The constancy of ζ is a generalization of this law to evolution at superhorizon scales, Exercise 30.4. But the entropy of the vacuum as a mode exits the horizon is tiny, and the entropy of the matter-radiation fluid when a mode re-enters the horizon is large, yet no entropy has been created because ζ is constant. How can these viewpoints be reconciled? **Answer.** The first law (10.34) can be construed as an equation representing conservation of entropy only if the system is evolving through states of thermodynamic equilibrium. The expanding Universe is not a system in thermodynamic equilibrium, even when its geometry is precisely FLRW. For systems not in thermodynamic equilibrium, the first law of thermodynamics (10.34) enforced by the Friedmann equations simply represents conservation of energy in a general relativistic context. The proof in Exercise 30.4 that ζ represents a fluctuation in entropy depended on the proposition that the proper pressure $p(\rho)$ is a definite function of proper density ρ . But in a system that is not in thermodynamic equilibrium and that evolves irreversibly from one state to another, the pressure is not a definite function of density. Reheating, the transition between vacuum and particle energy that marks the end of inflation, represents an irreversible (explosive!) increase in entropy. If the expansion of the Universe were reversed, the collapsing Universe would not revert from particle energy to vacuum energy, since that would require a reduction of entropy, in violation of the second law of thermodynamics. Reheating is analogous to the situation of a fluid that passes through a shock front. The shock converts kinetic into heat energy, increasing the entropy of the fluid, while conserving its energy.

30.3.1 Adiabatic and isocurvature initial conditions

The conserved fluctuation in any particular species x that does not exchange energy with other species is denoted ζ_x with a subscript x . The conserved fluctuation over all species is denoted ζ with no subscript. A generic prediction of inflation is that the conserved fluctuation ζ_x is the same for all species x ,

$$\zeta_x = \zeta . \quad (30.31)$$

Fluctuations in which the fluctuation is the same for all species are said to be **adiabatic**.

There are also **isocurvature** fluctuations, in which the entropy fluctuations δ_x of different species oppose each other so as to make zero contribution to the curvature potential Φ . Among N species, there are 1 adiabatic and $N - 1$ isocurvature modes subject to the condition that the initial fluctuations are finite.

30.4 Unperturbed background

The evolution of the cosmic scale factor a as a function of conformal time η depends on the energy-momentum content of the unperturbed background FLRW geometry. Much of this Chapter is concerned with an epoch starting somewhat after electron-positron annihilation at a redshift $1+z \sim 10^9$, and ending somewhat after recombination at $1+z_{\text{rec}} \approx 1100$. During this time the Universe was dominated by matter ($w = 0$) and radiation ($w = 1/3$), transitioning from radiation- to matter-dominated at a redshift of $1+z_{\text{eq}} \approx 3400$.

In the unperturbed background, the unperturbed dark matter density $\bar{\rho}_c$ and radiation density $\bar{\rho}_r$ evolve with cosmic scale factor as

$$\bar{\rho}_c \propto a^{-3}, \quad \bar{\rho}_r \propto a^{-4}. \quad (30.32)$$

The Hubble parameter H is defined in the usual way to be

$$H \equiv \frac{1}{a} \frac{da}{dt} = \frac{\dot{a}}{a^2}, \quad (30.33)$$

in which overdot represents differentiation with respect to conformal time, $\dot{a} \equiv da/d\eta$. The Friedmann equations for the background imply that the Hubble parameter for a universe dominated by dark matter and radiation is

$$H^2 = \frac{8\pi G}{3}(\bar{\rho}_c + \bar{\rho}_r) = \frac{H_{\text{eq}}^2}{2} \left(\frac{a_{\text{eq}}^3}{a^3} + \frac{a_{\text{eq}}^4}{a^4} \right), \quad (30.34)$$

where a_{eq} and H_{eq} are the cosmic scale factor and the Hubble parameter at the time of matter-radiation equality, $\bar{\rho}_c = \bar{\rho}_r$.

The comoving horizon distance η is defined to be the comoving distance that light travels starting from zero expansion. This is

$$\eta = \int_0^a \frac{da}{a^2 H} = \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left(\sqrt{1 + \frac{a}{a_{\text{eq}}}} - 1 \right) = \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left(\frac{a/a_{\text{eq}}}{1 + \sqrt{1 + a/a_{\text{eq}}}} \right). \quad (30.35)$$

The horizon distance η_{eq} at matter-radiation equality $a = a_{\text{eq}}$ is

$$\eta_{\text{eq}} = \frac{2\sqrt{2}}{(1 + \sqrt{2})a_{\text{eq}} H_{\text{eq}}}. \quad (30.36)$$

Equation (30.35) inverts to give the cosmic factor a as a function of the horizon distance η ,

$$\frac{a}{a_{\text{eq}}} = \frac{\eta}{8\eta_{\text{eq}}} \left(\frac{\eta}{\eta_{\text{eq}}} + 4\sqrt{2} \right). \quad (30.37)$$

In the radiation- and matter-dominated epochs respectively, the comoving horizon distance η is

$$\eta = \begin{cases} \frac{\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left(\frac{a}{a_{\text{eq}}} \right) = \frac{(1 + \sqrt{2})\eta_{\text{eq}}}{2} \left(\frac{a}{a_{\text{eq}}} \right) & \propto a \quad \text{radiation-dominated,} \\ \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left(\frac{a}{a_{\text{eq}}} \right)^{1/2} = (1 + \sqrt{2})\eta_{\text{eq}} \left(\frac{a}{a_{\text{eq}}} \right)^{1/2} & \propto a^{1/2} \quad \text{matter-dominated.} \end{cases} \quad (30.38)$$

The ratio of the comoving horizon distance η to the comoving Hubble distance $1/(aH)$ is

$$\eta aH = \frac{2\sqrt{1+a/a_{\text{eq}}}}{1+\sqrt{1+a/a_{\text{eq}}}}, \quad (30.39)$$

which is evidently a number of order unity, varying between 1 in the radiation-dominated epoch $a \ll a_{\text{eq}}$, and 2 in the matter-dominated epoch $a \gg a_{\text{eq}}$.

Concept question 30.6 What is meant by the horizon in cosmology? See §10.21.

Exercise 30.7 Redshift of matter-radiation equality.

1. Argue that the redshift z_{eq} of matter-radiation equality is given by

$$1 + z_{\text{eq}} = \frac{a_0}{a_{\text{eq}}} = ? \Omega_{\text{m}} h^2, \quad (30.40)$$

where Ω_{m} is the matter density today relative to critical. What is the factor, and what is its numerical value? The factor depends on the energy-weighted effective number of relativistic species g_{ρ} , equation (10.153b). Should this g_{ρ} be that now, or that at matter-radiation equality?

2. Show that the ratio H_{eq}/H_0 of the Hubble parameter at matter-radiation equality to that today is

$$\frac{H_{\text{eq}}}{H_0} = \sqrt{2\Omega_{\text{m}}} (1 + z_{\text{eq}})^{3/2}. \quad (30.41)$$

Solution. The redshift z_{eq} of matter-radiation equality is given by

$$1 + z_{\text{eq}} = \frac{\Omega_{\text{m}}}{\Omega_{\text{r}}} = \frac{45c^5 \hbar^3 \Omega_{\text{m}} H_0^2}{4\pi^3 G g_{\rho} (kT_0)^4} = 8.093 \times 10^4 \frac{\Omega_{\text{m}} h^2}{g_{\rho}} = 3400 \left(\frac{g_{\rho}}{3.36} \right)^{-1} \left(\frac{\Omega_{\text{m}} h^2}{0.143} \right), \quad (30.42)$$

where $T_0 = 2.725 \text{ K}$ is the present-day CMB temperature, and $g_{\rho} = 2 + 6 \frac{7}{8} \left(\frac{4}{11} \right)^{4/3} = 3.36$ is the energy-weighted effective number of relativistic species at matter-radiation equality, equation (10.153b). The value $\Omega_{\text{m}} h^2 = 0.143 \pm 0.001$ is from Planck Collaboration (2018).

30.5 Generic behaviour of non-baryonic cold dark matter

Non-baryonic cold dark matter is pressureless, $w = 0$, and it conserves energy-momentum because it does not scatter off radiation or baryons. Equation (30.14), which expresses energy-momentum conservation of a fluid, reduces for $w = 0$ to

$$\left(\frac{d^2}{d\eta^2} + \frac{\dot{a}}{a} \frac{d}{d\eta} \right) (\delta_{\text{c}} - 3\Phi) = -k^2 \Psi. \quad (30.43)$$

If $\Psi = \Phi$, then the source on the right hand side is $-k^2 \Phi$.

In the absence of a driving potential, $\Psi = 0$, the dark matter velocity would redshift as $v_{\text{c}} \propto 1/a$, equation (30.13b), and the dark matter density would then evolve as $\dot{\delta}_{\text{c}} = k v_{\text{c}} \propto a^{-1}$, equation (30.13a).

In the radiation-dominated epoch, where $\eta \propto a$, this leads to a logarithmic growth in the overdensity δ_c , even though there is no driving potential, and the velocity is redshifting to a halt. In the matter-dominated epoch, where $\eta \propto a^{1/2}$, the dark matter overdensity δ_c would freeze out at a constant value, in the absence of a driving potential.

More generally, equation (30.43) is a linear differential equation for $\delta_c - 3\Phi$ driven by a potential Ψ . You will find the solution to this equation for a prescribed potential Ψ in Exercise 30.8.

Exercise 30.8 Generic behaviour of dark matter. Find the homogeneous solutions of equation (30.43) for $\delta_c - 3\Phi$ with horizon distance η related to cosmic scale factor a by equation (30.35). Hence find the retarded Green's function of the equation. Write down the general solution of equation (30.43) as an integral over the Green's function. Solve for the case of constant potential Ψ .

Solution. The general solution of equation (30.43) is, in units $a_{\text{eq}} = 1$,

$$\delta_c(a) - 3\Phi(a) = A_0 + A_1 \ln x + 2k^2 \int_0^x \Psi(a') \ln \left(\frac{x'}{x} \right) a'^2 \frac{dx'}{x'} , \quad (30.44)$$

where A_0 and A_1 are constants, and

$$x \equiv \exp \left(\frac{1}{\sqrt{2}} \int \frac{d\eta}{a} \right) = \frac{a}{(1 + \sqrt{1+a})^2} = \frac{\eta}{\eta + 4\sqrt{2}} , \quad (30.45)$$

which simplifies to $x \rightarrow a/4$ as $a \rightarrow 0$ and $x \rightarrow 1 - 2/\sqrt{a}$ as $a \rightarrow \infty$. In the radiation-dominated and matter-dominated regimes, equation (30.44) reduces to

$$\delta_c(a) - 3\Phi(a) \rightarrow \begin{cases} A_0 + A_1 \ln(a/4) + 2k^2 \int_0^a \Psi(a') \ln \left(\frac{a'}{a} \right) a' da' & (a \ll 1) , \\ B_0 - 2B_1 a^{-1/2} + 4k^2 \int_0^a \Psi(a') \left(1 - \sqrt{\frac{a'}{a}} \right) da' & (a \gg 1) , \end{cases} \quad (30.46)$$

where the constants B_0 and B_1 in the $a \gg 1$ expression will usually differ from A_0 and A_1 thanks to contributions to the integral at $a' \lesssim 1$ that are not given correctly by the $a' \gg 1$ approximation.

30.6 Generic behaviour of radiation

Before recombination, photons are tightly coupled to baryons through nonrelativistic electron-photon (Thomson) scattering. The photon-baryon fluid thus behaves as a single energy-momentum conserving fluid. In the simple limit of negligible baryon density, the photon-baryon fluid can be treated as a relativistic fluid with $w = 1/3$. Equation (30.14) then reduces to

$$\left(3 \frac{d^2}{d\eta^2} + k^2 \right) (\Theta_0 - \Phi) = -k^2 (\Psi + \Phi) . \quad (30.47)$$

If $\Psi = \Phi$, then the source on the right hand side is just $-2k^2\Phi$.

In the absence of a driving potential, $\Psi + \Phi = 0$, the radiation oscillates as $\Theta_0 \propto e^{\pm i\omega\eta}$ with frequency $\omega = k/\sqrt{3}$. In other words, the solutions are sound waves, moving at the sound speed

$$c_s = \frac{\omega}{k} = \sqrt{\frac{1}{3}}. \quad (30.48)$$

Define the sound horizon distance η_s by

$$\eta_s \equiv c_s \eta = \frac{\eta}{\sqrt{3}}. \quad (30.49)$$

In terms of the sound horizon distance η_s , the differential equation (30.47) becomes

$$\left(\frac{d^2}{d\eta_s^2} + k^2 \right) (\Theta_0 - \Phi) = -k^2 (\Psi + \Phi). \quad (30.50)$$

Equation (30.50) is a linear differential equation for $\Theta_0 - \Phi$ driven by a potential $\Psi + \Phi$. You will find the solution to this equation for a prescribed potential $\Psi + \Phi$ in Exercise 30.9.

Exercise 30.9 Generic behaviour of radiation. Find the homogeneous solutions of equation (30.50). Hence find the retarded Green's function of the equation. Write down the general solution of equation (30.50) as an integral over the Green's function. Convince yourself that $\Theta_0 - \Phi$ oscillates about $-(\Psi + \Phi)$.

Solution. The general solution of equation (30.50) is, with $y \equiv k\eta_s$,

$$\Theta_0(y) - \Phi(y) = A_0 \cos y + A_1 \sin y - \int_0^y [\Psi(y') + \Phi(y')] \sin(y - y') dy', \quad (30.51)$$

where A_0 and A_1 are constants.

Concept question 30.10 Can neutrinos be treated as a fluid? Since neutrinos stream collisionlessly, how can it be legitimate to treat neutrinos as a fluid? **Answer.** The complete momentum distribution of neutrinos is characterized by a full set of multipole moments, which can be solved using the hierarchy (33.91) of Boltzmann equations. A fluid approximation amounts to keeping the first three momentum moments, the energy, bulk velocity, and pressure, in the multipole expansion of the momentum distribution. The Einstein equations depend only on these moments. If an adequate approximation to the pressure can be made, then the Boltzmann hierarchy can be truncated. The perfect fluid approximation amounts to approximating the pressure as isotropic, and given as a prescribed function of energy. The perfect fluid approximation is adequate for photons, which are isotropized by collisions, but is poor for neutrinos. As a result of free streaming, neutrinos develop a quadrupole (anisotropic pressure), as well as higher multipoles. You will discover in Exercise 32.7 that, in contrast to photons which behave as a fluid with sound speed $\sqrt{1/3}$ times the speed of light, neutrinos more closely approximate a fluid with sound speed equal to the speed of light. Thus the simple approximation of the present Chapter is not really adequate for neutrinos.

30.7 Equations for the simplest set of assumptions

The equations for two perfect fluids consisting of matter ($w = 0$) and radiation ($w = 1/3$) in a perturbed FLRW universe comprise 5 equations as follows. The first two equations express conservation of energy and momentum for non-baryonic cold dark matter (subscript c):

$$\dot{\delta}_c - k v_c = 3 \dot{\Phi} , \quad (30.52a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c = -k \Psi . \quad (30.52b)$$

The next two equations express conservation of energy and momentum for radiation (subscript r), which includes both photons and neutrinos:

$$\dot{\Theta}_0 - k \Theta_1 = \dot{\Phi} , \quad (30.53a)$$

$$\dot{\Theta}_1 + \frac{k}{3} \Theta_0 = -\frac{k}{3} \Psi . \quad (30.53b)$$

The final equation is the Einstein energy equation (29.56a):

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 (\bar{\rho}_c \delta_c + 4\bar{\rho}_r \Theta_0) , \quad (30.54)$$

where F , equation (29.48), is

$$F \equiv \frac{\dot{a}}{a} \Psi + \dot{\Phi} . \quad (30.55)$$

In place of one of the equations (30.52)–(30.54) it is sometimes convenient to use the Einstein momentum equation (29.56b)

$$-kF = 4\pi G a^2 (\bar{\rho}_c v_c + 4\bar{\rho}_r \Theta_1) , \quad (30.56)$$

which, because the matter and radiation equations (30.52) and (30.53) already satisfy covariant energy-momentum conservation, is not an independent equation. In the simple approximation of perfect fluids considered here, the radiation quadrupole vanishes, and then the Einstein quadrupole pressure equation (29.56d) implies that the scalar potentials Ψ and Φ are equal,

$$\Psi = \Phi . \quad (30.57)$$

Exercise 30.11 Program the equations for the simplest set of cosmological assumptions.

Write computer code that integrates numerically the evolution equations (30.52)–(30.54). In Exercise 32.2 you will generalize this code to include more components and more processes, so you should write the code in a well-structured fashion that allows you to update it easily. It is theoretically and numerically advantageous to treat $\delta_c - 3\Phi$ and $\Theta_0 - \Phi$ as dependent variables, rather than δ_c and Θ_0 . I found it convenient to use $\ln a$ as the integration variable, and to work in units $a_{\text{eq}} = H_{\text{eq}} = 1$. Assume adiabatic initial conditions, $\zeta_c = \zeta_r$ (see §30.10), and without loss of generality normalize to unit initial amplitudes, $\zeta_c = \zeta_r = 1$. Do the computation for a selection of wavenumbers k . Plot $\Theta_0 - \Phi$ and -2Φ together to bring out the fact that the

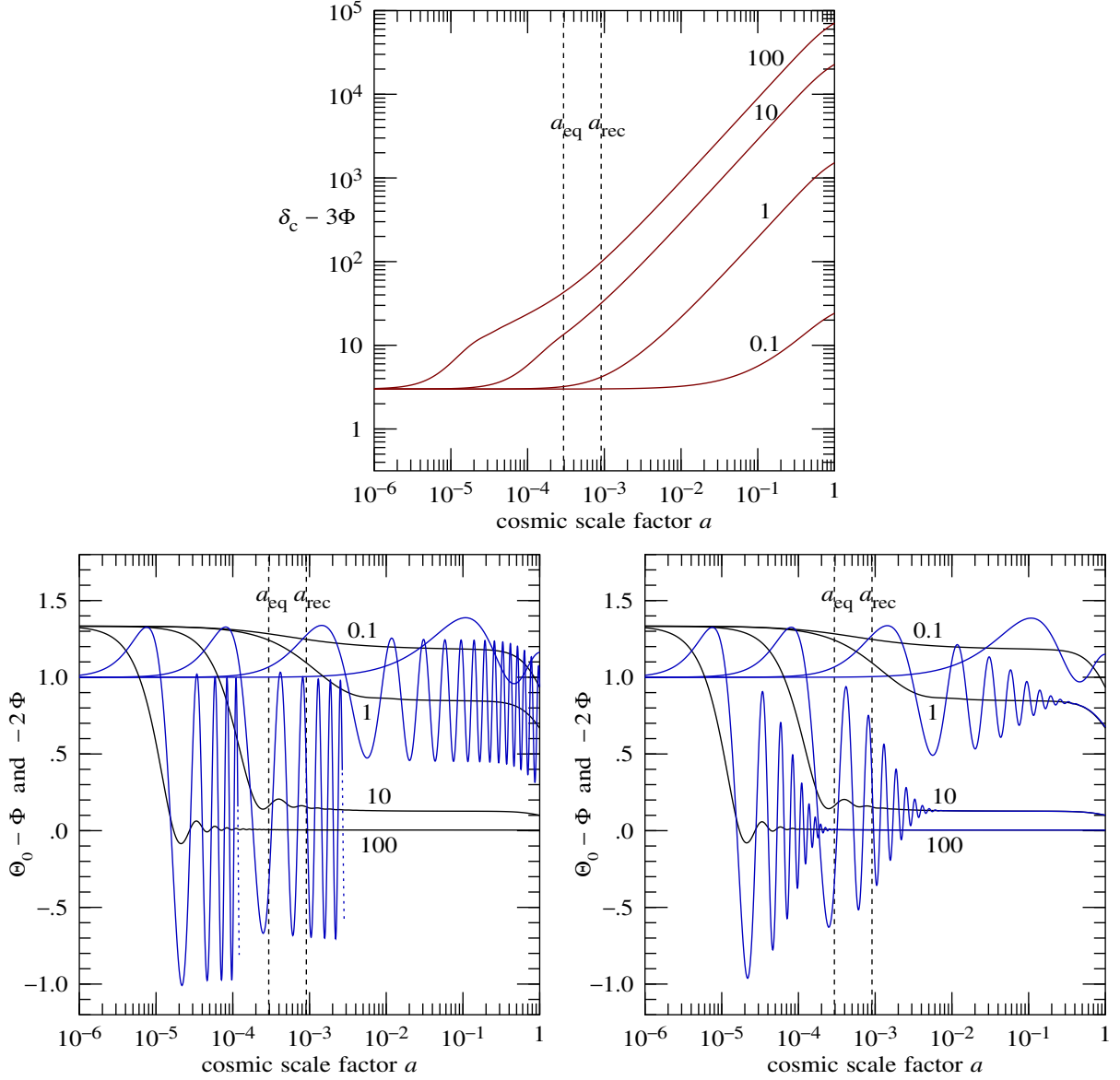


Figure 30.1 (Top) dark matter overdensity $\delta_c - 3\Phi$, (bottom left) radiation monopole $\Theta_0 - \Phi$, and (bottom right) radiation monopole $\Theta_0 - \Phi$ with artificial damping, as a function of cosmic scale factor a in units $a_0 = 1$, in the simple approximation, for several wavenumbers k . The cosmological model is a flat Λ CDM model with concordance parameters $\Omega_\Lambda = 0.69$ and $\Omega_m = 0.31$, and adiabatic initial conditions (see §32.3). The radiation monopole $\Theta_0 - \Phi$ (blue) is plotted along with minus twice the gravitational potential, -2Φ (black), to bring out the fact that the former oscillates about the latter, as expected from equation (30.47). Curves are labelled with the comoving wavenumber $k/(a_{\text{eq}} H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. For the larger wavenumbers, $k/(a_{\text{eq}} H_{\text{eq}}) = 10$ and 10^2 , the radiation monopole without damping is truncated (bottom left, dotted lines) to avoid confusing the plot. The radiation monopole shown here in the simple approximation may be compared to results in the hydrodynamic approximation, Figure 32.3, and using a Boltzmann computation, Figure 33.3.

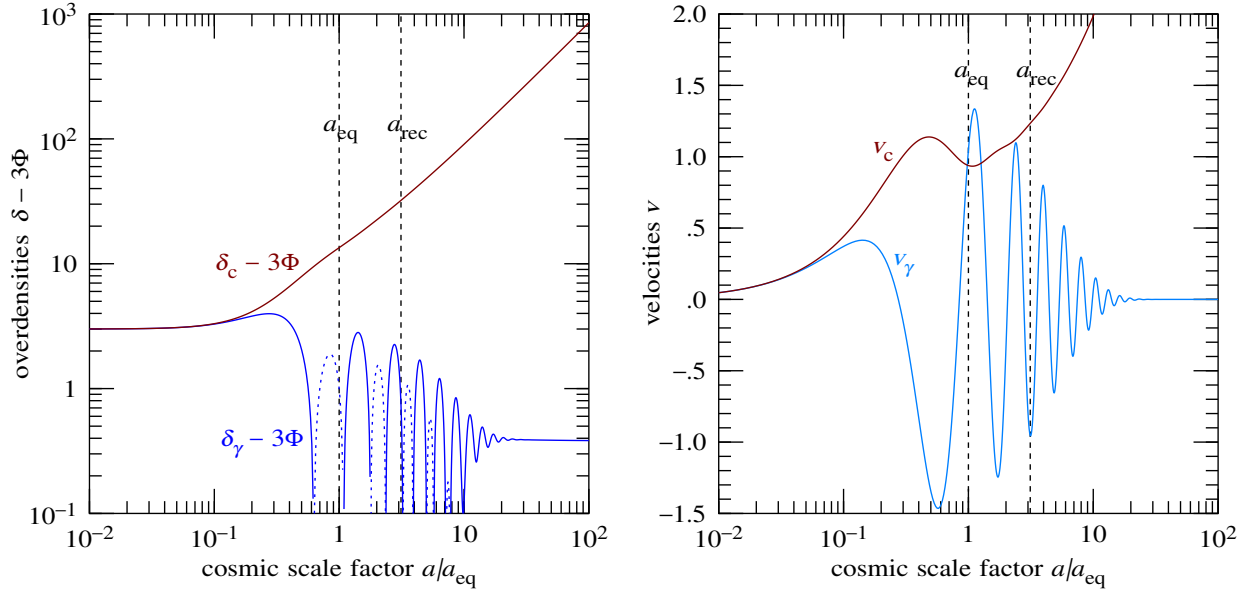


Figure 30.2 (Left) Overdensities $\delta - 3\Phi$, and (right) bulk velocities v in the simple approximation with artificial damping, as a function of cosmic scale factor a/a_{eq} , at wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$, for non-baryonic dark matter (c) and radiation (γ). The radiation overdensity and bulk velocity are related to their monopole and dipole moments by $\delta_\gamma - 3\Phi = 3(\Theta_0 - \Phi)$ and $v_\gamma = 3\Theta_1$, equations (30.15). The results may be compared to those from the hydrodynamic approximation, Figure 32.1, and a Boltzmann computation, Figure 33.1.

former oscillates about the latter, as expected from Exercise 30.9. A numerical issue you may encounter is that your integration routine may get stuck trying to integrate the oscillating radiation monopole and dipole once the mode is well inside the horizon, $k\eta \gg 1$. One strategy is to stop following the photon moments after a certain time. Another convenient strategy is to introduce an artificial damping term, by changing the radiation dipole equation (30.53b) to

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 + \Psi) = -2k\kappa\Theta_1, \quad (30.58)$$

where κ is a dimensionless damping coefficient that becomes large when the fluctuation is well inside the horizon, $k\eta \gg 1$,

$$\kappa = \epsilon k\eta, \quad (30.59)$$

with ϵ some suitably small number (I chose $\epsilon = 10^{-3}$). To see why the damping term works as claimed, combine the radiation monopole and dipole equations into a second order differential equation, and read §32.5. The introduction of damping anticipates, but is not an adequate substitute for, the physical processes of damping addressed in Chapter 32.

Solution. Figure 30.1 shows the dark matter overdensity, radiation monopole, and potential for a flat Λ CDM

model with $\Omega_\Lambda = 0.69$ and $\Omega_c = 0.31$, consistent with Planck parameters (Planck Collaboration, 2018). and adiabatic initial conditions. The radiation monopole is shown both without (bottom left panel) and with (bottom right panel) artificial damping. Figure 30.2 illustrates the overdensity and bulk velocity of each of matter and radiation for the same model at an illustrative wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$.

30.8 On the numerical computation of cosmological power spectra

See Seljak and Zaldarriaga (1996) for a discussion of the numerical computation of the CMB power spectrum. Modern codes that compute cosmological power spectra from linear perturbation theory, such as CAMB (google it), are impressively fast. With default settings, CAMB takes a few cpu seconds to compute a complete CMB power spectrum. CAMB is written in parallelized fortran 90.

To accomplish its task, CAMB:

1. reads cosmological parameters from an input file edited by the user;
2. calls RecFast (or other code) to compute recombination (Chapter 31);
3. uses a Boltzmann code (Chapter 33) to calculate the evolution of non-baryonic cold dark matter, baryonic matter, photons, and neutrinos at each of ~ 200 wavenumbers k ;
4. pre-calculates tables of spherical Bessel functions j_ℓ ;
5. computes CMB transfer functions $T_\ell(\eta_0, k)$, equation (34.20), by integrating source functions over Bessel functions at each of ~ 2000 wavenumbers k and ~ 100 harmonics ℓ (Chapter 34);
6. computes the CMB power spectrum $C_\ell(\eta_0)$ today by integrating the squared transfer functions over an almost scale-free primordial curvature spectrum, equation (34.35).

That is a lot of computation. The two most time-consuming steps are step 3, the Boltzmann computation, and step 5, the computation of CMB transfer functions. For step 3, CAMB uses the open-source ordinary differential equation solver dverk (Hull, Enright & Jackson 1976). Step 5 involves integration over highly oscillatory integrands. One could contemplate using some clever mathematical approach to integrate the highly oscillatory integrands, but CAMB simply uses a brute-force sum, interpolating pre-computed source functions in k -space, and splining over pre-computed spherical Bessel functions.

Most of the calculations of cosmological perturbations and power spectra reported in this book used Mathematica, a program that I use and value a lot. Sadly, high speed numerical calculations are not Mathematica's forte. One elementary issue is that Mathematica's inbuilt spherical Bessel functions j_ℓ are inexplicably slow for large ℓ , which is unacceptable given that many thousands of Bessel functions must be evaluated (on my 2015 laptop, a single evaluation of $j_\ell(\ell)$ takes approximately $(\ell/20,000)^2$ cpu seconds). Mathematica's biggest challenge is integrating the highly oscillatory functions in step 5. Mathematica's numerical integration routine NDSolve (or worse, NIntegrate, which treats each integrand in a list separately) evaluates its integrands far too often to be efficient. If you choose to program in Mathematica, good luck; but be warned that Mathematica assumes control over many details that basic languages like c and fortran leave up to you. Working with Mathematica is like trying to persuade a recalcitrant child to perform what seems to be a simple task; there is no guarantee who will win the contest of wills.

30.9 Analytic solutions in various regimes

Much of the remainder of this Chapter is concerned with obtaining approximate analytic solutions that describe the evolution of perturbations of the matter and radiation in various regimes. The aim is to gain some intuitive understanding of the solutions to the system of equations (30.52)–(30.57).

Figure 30.3 illustrates key features in the evolution of perturbations. Evolution is punctuated by the transition from radiation- to matter-dominated at $1+z_{\text{eq}} \approx 3400$, and by the transition from opaque to transparent at recombination, at $1+z_{\text{rec}} \approx 1100$. Meanwhile the comoving horizon distance η increases monotonically.

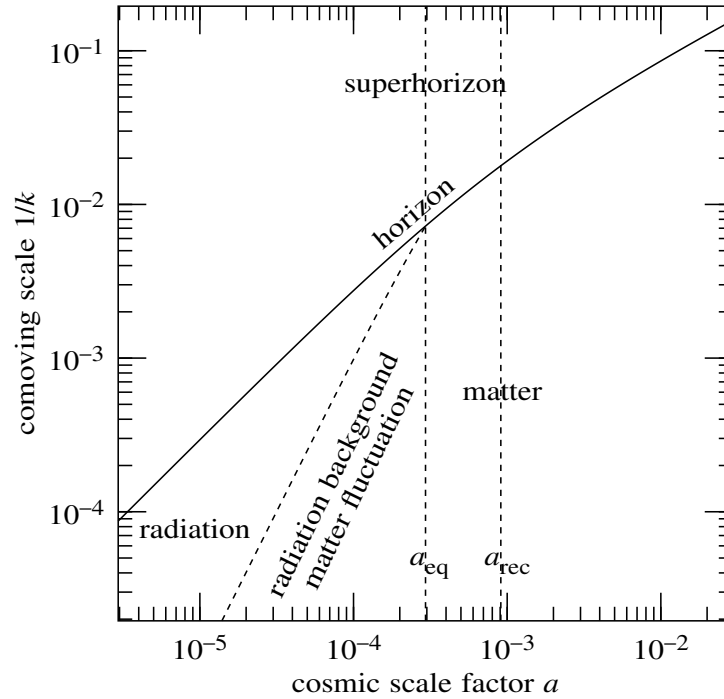


Figure 30.3 Various regimes in the evolution of fluctuations. The line increasing diagonally from bottom left to top right is the comoving horizon distance η . Above this line are superhorizon fluctuations, whose comoving wavelengths exceed the horizon distance, while below the line are subhorizon fluctuations, whose comoving wavelengths are less than the horizon distance. The dashed vertical line at cosmic scale factor $a_{\text{eq}} \approx a_0/3400$ marks the moment of matter-radiation equality. Before matter-radiation equality (to the left), the background mass-energy was dominated by radiation, while after matter-radiation equality (to the right), the background mass-energy was dominated by matter. Once a fluctuation enters the horizon, the non-baryonic matter fluctuation tends to grow, whereas the radiation fluctuation tends to decay, so there is an epoch prior to matter-radiation equality where gravitational perturbations are dominated by matter rather than radiation fluctuations, even though radiation dominates the background energy density. The dashed vertical line at $a_{\text{rec}} \approx a_0/1100$ marks recombination, where the temperature cooled to the point that baryons changed from being mostly ionized to mostly neutral, and the Universe changed from being opaque to transparent. The observed CMB comes from the time of recombination.

Small wavelength perturbations enter the horizon early, during the radiation-dominated regime, while long wavelength perturbations enter the horizon late, during the matter-dominated regime.

One regime not covered by the analytic approximations is perturbations that enter the horizon near the epoch of matter-radiation equality. The regime is important because the first few peaks, the most prominent peaks, in the CMB entered the horizon around or shortly after matter-radiation equality. Covering this regime satisfactorily requires solving numerically the full set of equations (30.52)–(30.54).

The regimes covered below are:

1. Superhorizon scales, §30.10.
2. Radiation-dominated:
 - a. adiabatic initial conditions, §30.11;
 - b. isocurvature initial conditions, §30.12.
3. Subhorizon scales, §30.13.
4. Fluctuations that enter the horizon in the matter-dominated epoch, §30.14.
5. Matter-dominated regime, §30.15.
6. Baryons post-recombination, §30.16.
7. Matter with dark energy, §30.17.
8. Matter with dark energy and curvature, §30.18.

30.10 Superhorizon scales

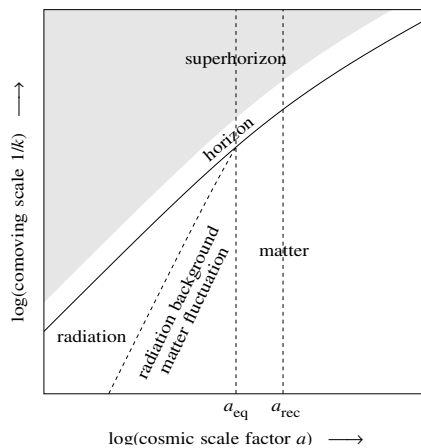


Figure 30.4 Superhorizon scales.

At sufficiently early times, any mode is outside the horizon, $k\eta < 1$. In the superhorizon limit $k\eta \ll 1$, the

evolution equations (30.52)–(30.54) reduce to

$$\dot{\delta}_c = 3\dot{\Phi} , \quad (30.60a)$$

$$\dot{\Theta}_0 = \dot{\Phi} , \quad (30.60b)$$

$$-3\frac{\dot{a}}{a}F = 4\pi Ga^2(\bar{\rho}_c\delta_c + 4\bar{\rho}_r\Theta_0) , \quad (30.60c)$$

with F defined by equation (30.55). In effect, the dark matter velocity v_c and radiation dipole Θ_1 can be treated as negligibly small at superhorizon scales,

$$v_c = \Theta_1 = 0 . \quad (30.61)$$

The first two of equations (30.60) imply that the dark matter overdensity δ_c and radiation monopole Θ_0 are related to the potential Φ by

$$\frac{1}{3}\delta_c - \Phi = \zeta_c , \quad (30.62a)$$

$$\Theta_0 - \Phi = \zeta_r , \quad (30.62b)$$

where ζ_c and ζ_r are constants set by initial conditions. Plugging the solutions (30.62) into the Einstein energy equation (30.60c), and replacing derivatives with respect to horizon distance η with derivatives with respect to cosmic scale factor a ,

$$\frac{\partial}{\partial\eta} = \dot{a}\frac{\partial}{\partial a} = a^2H\frac{\partial}{\partial a} , \quad (30.63)$$

with the Hubble parameter H from equation (30.34) gives the first order differential equation, in units $a_{\text{eq}} = 1$,

$$2a(1+a)\Phi' + (6+5a)\Phi + 4\zeta_r + 3\zeta_ca = 0 , \quad (30.64)$$

where prime ' denotes differentiation with respect to cosmic scale factor, d/da . The solution to equation (30.64) that is finite at $a = 0$ is

$$\Phi = -\frac{2}{3}\zeta_r + \left(\frac{2}{3}\zeta_r - \frac{3}{5}\zeta_c\right)f , \quad (30.65)$$

where $f(a)$ is the function

$$f \equiv 1 - \frac{2}{a} + \frac{8}{a^2} + \frac{16}{a^3} - \frac{16\sqrt{1+a}}{a^3} = \frac{a(6+a+4\sqrt{1+a})}{(1+\sqrt{1+a})^4} , \quad (30.66)$$

in which the rightmost expression is written in a form that is numerically well-behaved for all a . The function f varies from 0 at $a = 0$ to 1 at $a \rightarrow \infty$. The initial and final values of the potential $\Phi(a)$ are

$$\Phi(0) = -\frac{2}{3}\zeta_r , \quad \Phi(\text{late}) = -\frac{3}{5}\zeta_c . \quad (30.67)$$

The potential $\Phi(\text{late})$ is designated “late” because it holds in the matter-dominated regime well after recombination, but fails when dark energy (or possibly curvature) become important.

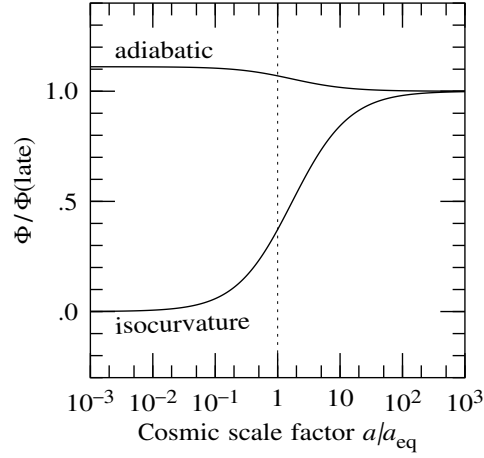


Figure 30.5 Evolution of the scalar potential Φ at superhorizon scales, equations (30.71), from radiation-dominated to matter-dominated. The scale for the potential is normalized to its value $\Phi(\text{late})$ at late times $a \gg a_{\text{eq}}$.

There are adiabatic and isocurvature initial conditions. Inflation generically produces adiabatic fluctuations, in which matter and radiation fluctuate together,

$$\zeta_c = \zeta_r \quad \text{adiabatic} , \quad (30.68)$$

so that

$$\delta_c(0) = 3\Theta_0(0) = -\frac{3}{2}\Phi(0) = \zeta_c \quad \text{adiabatic} . \quad (30.69)$$

Notice that a positive energy fluctuation corresponds to a negative potential, consistent with Newtonian intuition. Isocurvature initial conditions are defined by the vanishing of the initial potential, $\Phi(0) = 0$, requiring

$$\zeta_r = 0 \quad \text{isocurvature} . \quad (30.70)$$

The adiabatic and isocurvature solutions for the superhorizon potential Φ , equation (30.65), are

$$\Phi_{\text{ad}} = \zeta_c \left(-\frac{2}{3} + \frac{1}{15}f \right) , \quad (30.71a)$$

$$\Phi_{\text{iso}} = -\frac{3}{5}\zeta_c f , \quad (30.71b)$$

with f given by equation (30.66). Figure 30.5 shows the evolution of the potential Φ from equations (30.71), normalized to the value $\Phi(\text{late})$ at late times $a \gg a_{\text{eq}}$. For adiabatic fluctuations, the potential changes by a factor of 9/10 from initial to final value.

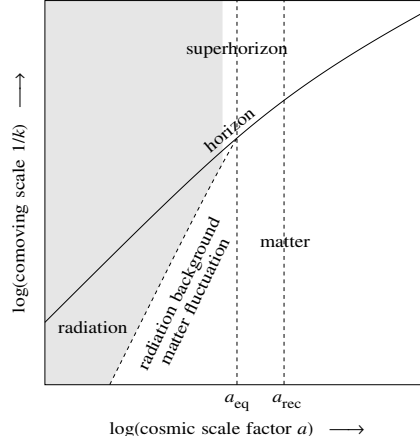


Figure 30.6 Radiation-dominated regime.

30.11 Radiation-dominated, adiabatic initial conditions

For adiabatic initial conditions, fluctuations that enter the horizon before matter-radiation equality, $k\eta_{\text{eq}} \gg 1$, are dominated by radiation. In the regime where radiation dominates both the unperturbed energy and its fluctuations, the relevant equations are, from equations (30.53), (30.54), and (30.56),

$$\dot{\Theta}_0 - k\Theta_1 = \dot{\Phi} , \quad (30.72a)$$

$$-3\frac{\dot{a}}{a}F - k^2\Phi = 16\pi Ga^2\bar{\rho}_r\Theta_0 , \quad (30.72b)$$

$$-kF = 16\pi Ga^2\bar{\rho}_r\Theta_1 , \quad (30.72c)$$

in which, because it simplifies the mathematics, the Einstein momentum equation is used as a substitute for the radiation dipole equation. In the radiation-dominated epoch, the horizon distance is proportional to the cosmic scale factor, $\eta \propto a$, equation (30.38). Inserting Θ_0 and Θ_1 from the Einstein energy and momentum equations (30.72b) and (30.72c) into the radiation monopole equation (30.72a) gives a second order differential equation for the potential Φ ,

$$\ddot{\Phi} + \frac{4}{\eta}\dot{\Phi} + \frac{k^2}{3}\Phi = 0 . \quad (30.73)$$

Equation (30.73) describes damped sound waves moving at sound speed $1/\sqrt{3}$ times the speed of light. The sound horizon, the comoving distance that sound can travel, is $\eta_s = \eta/\sqrt{3}$, the horizon distance η multiplied

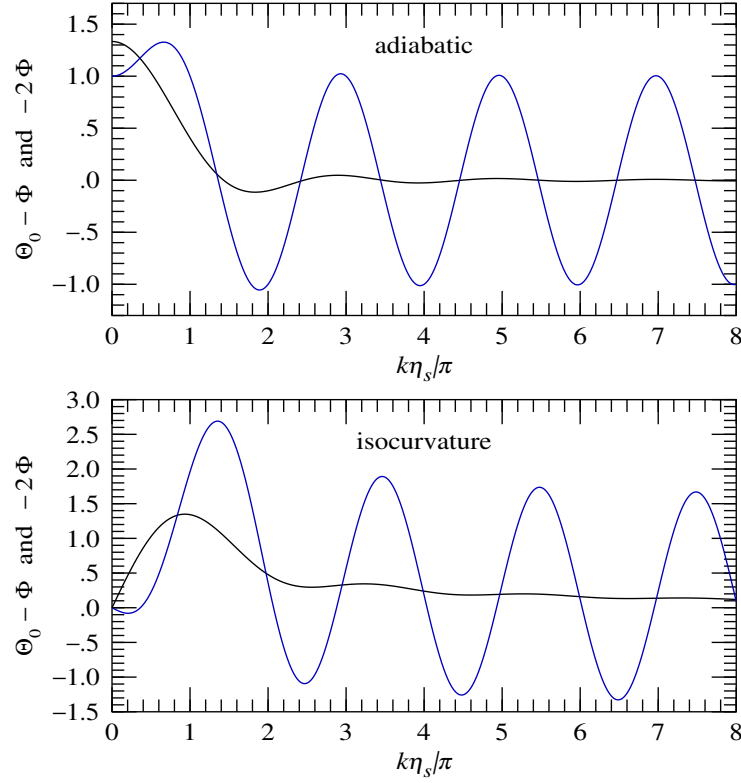


Figure 30.7 The potential Φ and radiation monopole Θ_0 for modes that enter the sound horizon $k\eta_s = 1$ during the radiation-dominated regime well before matter-radiation equality, for (top) adiabatic initial conditions, equations (30.76) and (30.77), and (bottom) isocurvature initial conditions, equations (30.83) and (30.84). The quantities shown are (blue) $\Theta_0 - \Phi$ and (black) -2Φ , to illustrate that the former oscillates about the latter as expected from equation (30.47). The difference, $(\Theta_0 - \Phi) - (-2\Phi) = \Theta_0 + \Phi$, which is the temperature Θ_0 redshifted by the potential Φ , is (for $\Psi = \Phi$) the monopole contribution to the temperature fluctuation of the CMB, equation (34.17). The units of Φ and Θ_0 are such that $\zeta_r = 1$ for adiabatic fluctuations, and $\zeta_c = 1$ for isocurvature fluctuations.

by the sound speed. The growing and decaying solutions to equation (30.73) are

$$\Phi_{\text{grow}} = \frac{3j_1(y)}{y} = \frac{3(\sin y - y \cos y)}{y^3}, \quad (30.74a)$$

$$\Phi_{\text{decay}} = -\frac{j_{-2}(y)}{y} = \frac{\cos y + y \sin y}{y^3}, \quad (30.74b)$$

where the dimensionless parameter y is the wavenumber k multiplied by the sound horizon distance η_s ,

$$y \equiv k\eta_s = \frac{k\eta}{\sqrt{3}} = \sqrt{\frac{2}{3}} \frac{k}{a_{\text{eq}} H_{\text{eq}}} \frac{a}{a_{\text{eq}}}, \quad (30.75)$$

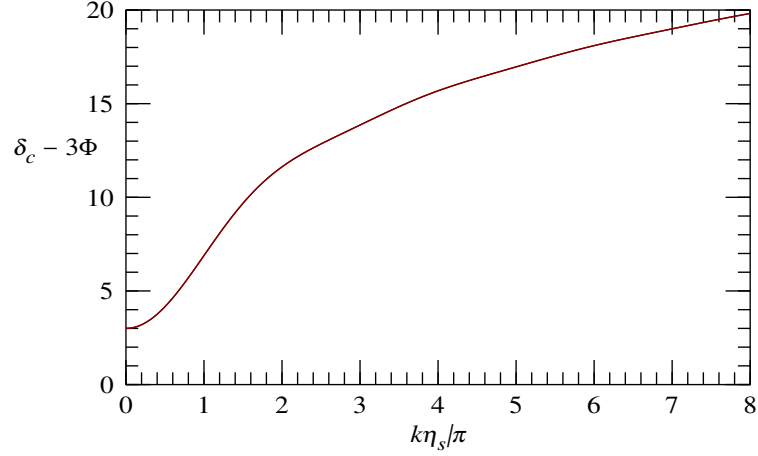


Figure 30.8 Evolution of the dark matter overdensity $\delta_c - 3\Phi$ for a mode that enters the horizon during the radiation-dominated regime, for adiabatic initial conditions. Like the radiation fluctuation Θ_0 illustrated in the top panel of Figure 30.7, the matter fluctuation δ_c is constant outside the sound horizon, $k\eta_s \ll 1$, and gets a boost as the fluctuation enters the sound horizon. But whereas the radiation fluctuation subsequently oscillates, the dark matter fluctuation grows monotonically, with logarithmic growth well inside the sound horizon, $k\eta_s \gg 1$. The units are such that $\zeta_r = 1$.

and $j_\ell(y) \equiv \sqrt{\pi/(2y)} J_{\ell+\frac{1}{2}}(y)$ are spherical Bessel functions. The physically relevant solution that satisfies adiabatic initial conditions, remaining finite as $y \rightarrow 0$, is the growing solution

$$\Phi = \Phi(0) \Phi_{\text{grow}} . \quad (30.76)$$

The growing solution (30.74a) shows that, after a mode enters the sound horizon the scalar potential Φ oscillates with an envelope that decays as y^{-2} . Physically, relativistically propagating sound waves tend to suppress the gravitational potential Φ . The suppression of the potential is responsible for the turnover in the observed power spectrum of matter fluctuations today from large to small scales evident in Figure 30.15.

The radiation monopole Θ_0 can be inferred either from the Einstein equation (30.72b) with the solutions (30.74) for the potential Φ , or from the Green's function solution (30.51) in the radiation-dominated regime. Either way, the difference $\Theta_0 - \Phi$ between the radiation monopole and the potential corresponding to the growing mode potential (30.76) is

$$\Theta_0 - \Phi = \zeta_r \frac{(2 \sin y - y \cos y)}{y} . \quad (30.77)$$

The top panel of Figure 30.7 shows the growing mode potential Φ , equation (30.76), and the radiation monopole Θ_0 , equation (30.77). The Figure plots these two quantities in the form -2Φ and $\Theta_0 - \Phi$ in order to bring out the fact that $\Theta_0 - \Phi$ oscillates about -2Φ , in accordance with equation (30.47). After a mode

is well inside the sound horizon, $y \gg 1$, the radiation monopole oscillates with constant amplitude,

$$\Theta_0 = -\zeta_r \cos y \quad \text{for } y \gg 1. \quad (30.78)$$

Fluctuations in the dark matter are driven by the gravitational potential of the radiation. The radiation-dominated Green's function solution (30.46) for the dark matter fluctuation δ_c driven by the growing mode potential (30.76) and satisfying adiabatic initial conditions (30.69) is

$$\delta_c - 3\Phi = 6\zeta_r \left(\frac{\sin y}{y} - \frac{1}{2} + \text{Cin } y \right), \quad (30.79)$$

where $\text{Cin } y \equiv \int_0^y (1 - \cos x) dx/x$ is the cosine integral. Figure 30.8 shows the density fluctuation (30.79). Once the mode is well inside the sound horizon, $y \gg 1$, the dark matter density δ_c , equation (30.79), evolves as, from the asymptotic behaviour $\text{Cin } y \sim \ln y + \gamma$ with $\gamma \equiv 0.5772\dots$ Euler's constant,

$$\delta_c - 3\Phi = 6\zeta_r \left(\ln y + \gamma - \frac{1}{2} \right) \quad \text{for } y \gg 1, \quad (30.80)$$

which grows logarithmically. This logarithmic growth translates into a logarithmic increase in the amplitude of matter fluctuations at small scales, and is a characteristic signature of non-baryonic cold dark matter.

Exercise 30.12 Radiation-dominated fluctuations.

1. Confirm equation (30.73). You might like to start by seeking a solution using the monopole and dipole radiation equations (30.53) along with the Einstein energy equation (30.54) including only radiation, namely equation (30.72b). Then try the solution advocated in the text, namely use the Einstein momentum equation (30.72c) in place of the radiation dipole equation. This is an example of a situation where, even though two sets of equations are equivalent, it is easier to find solutions from one set than the other.
2. Confirm that the homogeneous solutions of equation (30.73) are as given in the text, equations (30.74).
3. The initial condition for the temperature monopole is determined by equation (30.62b), $\Theta_0(0) - \Phi(0) = \zeta_r$, where ζ_r is some constant, the initial radiation entropy fluctuation set up during inflation. Find the initial conditions for the scalar potentials Ψ and Φ from the Einstein energy and quadrupole pressure equations at $\eta \rightarrow 0$ (in the present simple model, the Einstein quadrupole pressure equation simply sets $\Psi = \Phi$).
4. Confirm that the Green's function solution (30.51) for $\Theta_0 - \Phi$ satisfying the requisite boundary conditions is equation (30.77). Plot the solution for $\Theta_0 - \Phi$, along with -2Φ . Confirm that $\Theta_0 - \Phi$ oscillates around -2Φ .
5. Comment on the behaviour. How do the gravitational potential and temperature monopole evolve once a mode is inside the horizon? Can you come up with a physical explanation of what is going on?

30.12 Radiation-dominated, isocurvature initial conditions

For isocurvature initial conditions, the matter fluctuation contributes from the outset, $|\bar{\rho}_c \delta_c| > |4\bar{\rho}_r \Theta_0|$ even while radiation dominates the background density, $\bar{\rho}_c \ll \bar{\rho}_r$.

To develop an approximation adequate for isocurvature fluctuations entering the horizon well before matter-radiation equality, $k\eta_{\text{eq}} \gg 1$, regard the Einstein energy equation (30.54) as giving the radiation monopole Θ_0 , and the Einstein momentum equation (30.56) as giving the radiation dipole Θ_1 . Insert these into the radiation monopole equation (30.53a), and eliminate the $\dot{\delta}_c$ terms using the dark matter density equation (30.52a). The result is, in units $a_{\text{eq}} = 1$,

$$2a(1+a)\Phi'' + (8+9a)\Phi' + 2\left(1 + \frac{2k^2a}{3}\right)\Phi + \delta_c = 0, \quad (30.81)$$

where prime ' denotes differentiation with respect to cosmic scale factor a . Equation (30.81) is valid in all regimes, for any combination of matter and radiation.

For isocurvature initial conditions, the radiation monopole and potential vanish initially, $\Theta_0(0) = \Phi(0) = 0$, whereas the dark matter overdensity is finite, $\delta_c(0) = 3\zeta_c \neq 0$. For small scales that enter the horizon well before matter-radiation equality, $k\eta_{\text{eq}} \gg 1$, the potential Φ is small compared to δ_c , while δ_c has some approximately constant non-zero value up to and through the time when the mode enters the sound horizon, $k\eta_s = \sqrt{2/3}ka \approx 1$. In the radiation-dominated epoch, $a \ll 1$, but k large and $ka \sim 1$ so $k^2a \gg 1$, equation (30.81) simplifies to

$$2a\Phi'' + 8\Phi' + \frac{4k^2a}{3}\Phi + \delta_c = 0. \quad (30.82)$$

For constant $\delta_c = \delta_c(0) = 3\zeta_c$, the solution of equation (30.82) vanishing at $a = 0$ is, with y given by equation (30.75),

$$\Phi = -\frac{3\sqrt{3}\zeta_c}{\sqrt{2}k} \frac{1 - \cos y - y \sin y + \frac{1}{2}y^2}{y^3}. \quad (30.83)$$

With units restored, k is $k/(a_{\text{eq}}H_{\text{eq}})$. The Green's function solution (30.51) for the difference $\Theta_0 - \Phi$ between the radiation monopole and potential driven by the potential (30.83) is

$$\Theta_0 - \Phi = \frac{3\sqrt{3}\zeta_c}{\sqrt{2}k} \frac{(1 - \cos y - \frac{1}{2}y \sin y)}{y}. \quad (30.84)$$

Equations (30.83) and (30.84) are the solution for small scale modes with isocurvature initial conditions that enter the horizon well before matter-radiation equality. After a mode is well inside the sound horizon, $y \gg 1$, the radiation monopole (30.84) oscillates with constant amplitude,

$$\Theta_0 = -\frac{3\sqrt{3}\zeta_c}{2\sqrt{2}k} \sin y \quad y \gg 1. \quad (30.85)$$

The lower panel of Figure 30.7 shows the potential Φ , equation (30.83), and the radiation monopole Θ_0 from equation (30.84), again plotted as $\Theta_0 - \Phi$ and -2Φ to bring out the fact that $\Theta_0 - \Phi$ oscillates about -2Φ . Whereas for adiabatic initial conditions the radiation monopole oscillated as $\cos y$ well inside the sound

horizon, equation (30.78), for isocurvature initial conditions it oscillates as $\sin y$ well inside the sound horizon, equation (30.85).

The solution (30.83) for the potential Φ was derived from equation (30.82) on the assumption of constant δ_c . The accuracy of the approximation may be checked by calculating the radiation-dominated Green's function solution (30.46) for δ_c driven by this potential, which is

$$\delta_c - 3\Phi = 3\zeta_c \left(1 + a \frac{3 - 3 \cos y - 3y \operatorname{Si} y + \frac{3}{2}y^2}{y^2} \right), \quad (30.86)$$

where $\operatorname{Si} y \equiv \int_0^y \sin x \, dx/x$ is the sine integral. Equation (30.86) shows that $\delta_c - 3\Phi$ is approximately equal to $\delta_c(0) \equiv 3\zeta_c$ in the radiation-dominated regime $a \ll 1$ for all y . The dark matter overdensity δ_c itself is not constant, because Φ varies. However, Φ from equation (30.83) is of order $a\delta_c(0)$ for any y , and the small order a correction to δ_c leads to corrections of next order a^2 to Φ , $\Theta_0 - \Phi$, and $\delta_c - 3\Phi$, and can be neglected.

30.13 Subhorizon scales

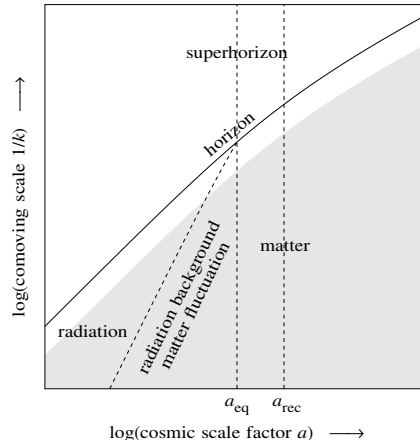


Figure 30.9 Subhorizon scales.

After a mode enters the horizon, the radiation fluctuation Θ_0 oscillates, but the non-baryonic cold dark matter fluctuation δ_c grows monotonically. In due course, the dark matter density fluctuation $\bar{\rho}_c \delta_c$ dominates the radiation density fluctuation $\bar{\rho}_r \Theta_0$, and this necessarily occurs before matter-radiation equality; that is, $|\bar{\rho}_c \delta_c| > |4\bar{\rho}_r \Theta_0|$ even though $\bar{\rho}_c < \bar{\rho}_r$. This is true for both adiabatic and isocurvature initial conditions; as noted in §30.12, for isocurvature initial conditions, the dark matter density fluctuation dominates from the outset. Even before the dark matter density fluctuation dominates, the cumulative contribution of the dark matter to the potential Φ begins to be more important than that of the radiation, because the potential sourced by the radiation oscillates, with an effect that tends to cancel when averaged over an oscillation.

Regard the Einstein energy equation (30.54) as giving the dark matter overdensity δ_c , and the Einstein momentum equation (30.56) as giving the dark matter velocity v_c . Insert these into the dark matter density equation (30.52a) and eliminate the $\dot{\Theta}_0$ terms using the radiation monopole equation (30.53a). The result is, in units $a_{\text{eq}} = 1$,

$$2a^2(1+a)\Phi'' + a(6+7a)\Phi' - 2\Phi - 4\Theta_0 = 0 , \quad (30.87)$$

where prime ' denotes differentiation with respect to cosmic scale factor a . Equation (30.87) is valid in all regimes, for any combination of matter and radiation.

Once the mode is well inside the horizon, $k\eta \gg 1$, the radiation monopole Θ_0 oscillates about an average value of $-\Phi$ (since $\Theta_0 - \Phi$ oscillates about -2Φ , as noted in §30.6):

$$\langle \Theta_0 \rangle = -\Phi . \quad (30.88)$$

Inserting this cycle-averaged value of Θ_0 into equation (30.87) gives the Meszaros differential equation

$$2(1+a)a^2\Phi'' + (6+7a)a\Phi' + 2\Phi = 0 . \quad (30.89)$$

The solutions of Meszaros' differential equation (30.89) are a linear combination of growing and decaying solutions

$$\Phi_{\text{grow}} = -\frac{3}{4k^2a} \left(1 + \frac{3a}{2} \right) , \quad (30.90a)$$

$$\Phi_{\text{decay}} = -\frac{3}{4k^2a} \left\{ \left(1 + \frac{3a}{2} \right) \ln \left[\frac{(\sqrt{1+a}+1)^2}{a} \right] - 3\sqrt{1+a} \right\} . \quad (30.90b)$$

A constant factor of $-3/(4k^2)$ has been included in the potential to simplify the overall factor in the resulting solution for the dark matter overdensity δ_c , equations (30.92). The solutions for δ_c driven by the growing and decaying potentials (30.90) are, from the Green's function solution (30.44), in units $a_{\text{eq}} = 1$,

$$\delta_c - 3\Phi = -\frac{4k^2a}{3}\Phi , \quad (30.91)$$

which holds for both growing and decaying modes. The solutions (30.91) omit possible additional contributions from the homogeneous solutions in equation (30.44), but the regime of interest is modes well inside the horizon, $ka \gg 1$, and the omitted contributions become dominated by the solutions (30.91) as the cosmic scale factor a increases. Explicitly, the growing and decaying modes for $\delta_c - 3\Phi$ are

$$(\delta_c - 3\Phi)_{\text{grow}} = 1 + \frac{3}{2}a , \quad (30.92a)$$

$$(\delta_c - 3\Phi)_{\text{decay}} = \left(1 + \frac{3}{2}a \right) \ln \left[\frac{(\sqrt{1+a}+1)^2}{a} \right] - 3\sqrt{1+a} . \quad (30.92b)$$

The desired solution for the dark matter overdensity δ_c is a linear combination of growing and decaying modes,

$$\delta_c - 3\Phi = C_{\text{grow}}(\delta_c - 3\Phi)_{\text{grow}} + C_{\text{decay}}(\delta_c - 3\Phi)_{\text{decay}} . \quad (30.93)$$

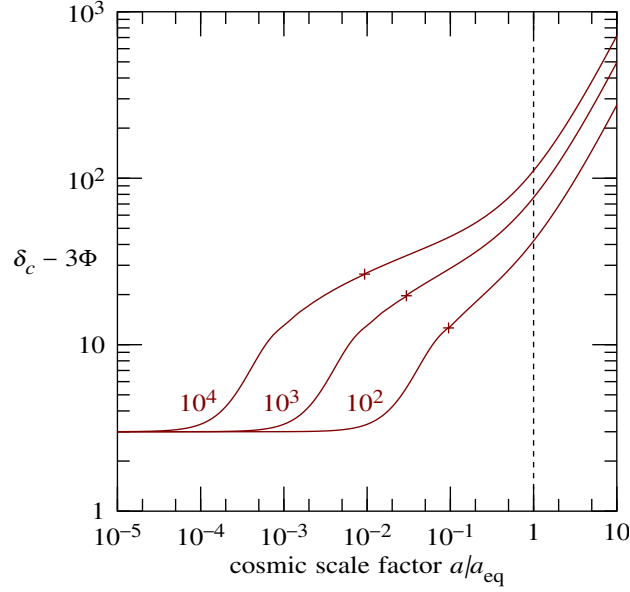


Figure 30.10 After an initial boost on entering the sound horizon, the dark matter overdensity δ_c grows logarithmically with cosmic scale factor a during the radiation-dominated regime, but then linearly with a after matter-radiation equality $a = a_{\text{eq}}$. The curves are labelled with the comoving wavenumber $k/(a_{\text{eq}} H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. The evolution is approximated by the radiation-dominated solution (30.79) at small a , and by the Meszaros solution (30.93) at larger a , with crosses marking the transition between the two approximations, at the geometric mean of the horizon distance at horizon crossing and matter-radiation equality $\eta \approx \sqrt{\eta_{\text{hor}} \eta_{\text{eq}}}$.

The coefficients C_{grow} and C_{decay} follow from matching to the earlier solutions for $\delta_c - 3\Phi$ obtained in the radiation-dominated regime. For modes that enter the horizon well before matter-radiation equality, $a \ll 1$, the growing and decaying modes (30.92) simplify to

$$(\delta_c - 3\Phi)_{\text{grow}} = 1, \quad (\delta_c - 3\Phi)_{\text{decay}} = -\ln(a/4) - 3 \quad \text{for } a \ll 1. \quad (30.94)$$

It was found in §30.11 that the potential Φ in the radiation-dominated regime oscillated with an envelope that decayed as $\sim a^{-2}$, equation (30.74a), driving a dark matter overdensity that grew as a combination of linear and logarithmic parts, equation (30.80). The result (30.94) demonstrates that a potential that is a sum of parts proportional to $1/a$ and $\ln a/a$, albeit reduced in amplitude by a factor of $1/k^2$, leads to the same behaviour of the dark matter overdensity.

For adiabatic initial conditions, the solution for the dark matter overdensity δ_c is the one that matches smoothly on to the logarithmically growing solution given by equation (30.80). Matching to the adiabatic solution (30.80) for $\delta_c - 3\Phi$ well inside the horizon determines the constants

$$C_{\text{grow}} = 6\zeta_r \left[\gamma - \frac{7}{2} + \ln \left(4\sqrt{\frac{2}{3}} k \right) \right], \quad C_{\text{decay}} = -6\zeta_r \quad \text{adiabatic}. \quad (30.95)$$

For isocurvature initial conditions, $\delta_c - 3\Phi$ is sensibly constant in the radiation-dominated regime $a \ll 1$, equation (30.86), and only the growing mode is present,

$$C_{\text{grow}} = 3\zeta_c, \quad C_{\text{decay}} = 0 \quad \text{isocurvature} . \quad (30.96)$$

At late times well into the matter-dominated epoch, $a \gg 1$, the growing mode of the Meszaros solution dominates,

$$(\delta_c - 3\Phi)_{\text{grow}} = \frac{3}{2}a, \quad (\delta_c - 3\Phi)_{\text{decay}} = \frac{4}{15}a^{-3/2} \quad \text{for } a \gg 1, \quad (30.97)$$

so that the dark matter overdensity δ_c at late times is

$$\delta_c - 3\Phi = \frac{3}{2}C_{\text{grow}}a \quad \text{for } a \gg 1. \quad (30.98)$$

The potential Φ , equation (30.90a), at late times is constant,

$$\Phi = -\frac{9}{8k^2}C_{\text{grow}} \quad \text{for } a \gg 1. \quad (30.99)$$

The constancy of the potential Φ , and the linear growth of the dark matter density δ_c , is characteristic of the matter-dominated regime.

Figure 30.10 shows the dark matter overdensity δ_c calculated for adiabatic conditions from the radiation-dominated solution (30.79) at small a , and the Meszaros solution (30.93) at larger a . The overdensity δ_c is constant before horizon crossing, receives a boost of growth during horizon-crossing, grows logarithmically with cosmic scale factor a during before matter-radiation equality, then grows linearly with a after matter-radiation equality.

For modes that enter the horizon well before matter-radiation equality, the radiation monopole Θ_0 at late times $a \gg 1$ is, with $y \equiv k\eta/\sqrt{3}$,

$$\Theta_0 = -\Phi - \zeta_r \cos y \quad \text{adiabatic} , \quad (30.100a)$$

$$\Theta_0 = -\Phi - \frac{3\sqrt{3}\zeta_c}{2\sqrt{2}k} \sin y \quad \text{isocurvature} . \quad (30.100b)$$

30.14 Fluctuations that enter the horizon during the matter-dominated epoch

For fluctuations that enter the horizon well after matter-radiation equality, $k\eta_{\text{eq}} \ll 1$, the potential Φ before entering the horizon is given by the superhorizon solution (30.65), while after entering the horizon the evolution of the potential is dominated by the dark matter density fluctuation. A satisfactory solution for the potential Φ valid both before and after entering the horizon is obtained by setting the radiation monopole equal to its superhorizon solution, $\Theta_0 = \Phi + \zeta_r$, equation (30.62b), and inserting this value into the differential equation (30.87). This solution remains an adequate approximation inside the horizon because after horizon crossing the radiation fluctuation Θ_0 makes a subdominant contribution to the Einstein energy equation, so its behaviour ceases to influence the evolution of the potential. Mathematically, once $a \gg a_{\text{eq}}$, the derivative terms Φ'' and Φ' dominate the Φ and Θ_0 terms in equation (30.87).

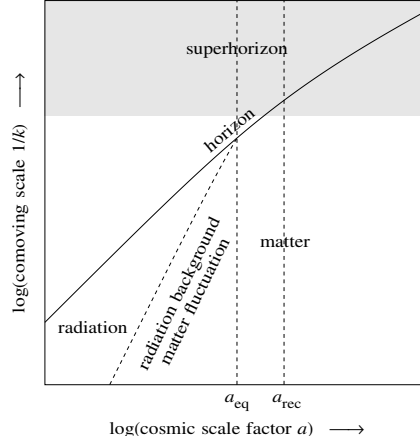


Figure 30.11 Fluctuations that enter the horizon in the matter-dominated regime.

Inserting the superhorizon solution $\Theta_0 = \Phi + \zeta_r$ into the differential equation (30.87) recovers the superhorizon solution (30.65) for Φ , which therefore remains a satisfactory approximation not only outside but also inside the horizon. But inside the horizon, the dark matter overdensity δ_c and radiation monopole Θ_0 driven by this potential are no longer their superhorizon solutions (30.62). Rather, the solution for the dark matter overdensity δ_c driven by the superhorizon potential, subject to the initial condition $\frac{1}{3}\delta_c - \Phi = \zeta_c$ is, from the Green's function solution (30.44),

$$\delta_c - 3\Phi = 3\zeta_c + k^2 \left\{ -2a^2 \left(\frac{3}{5}\zeta_c + \Phi \right) + \left(\frac{8}{3}\zeta_r - \frac{16}{5}\zeta_c \right) \left[4 \ln \left(\frac{1 + \sqrt{1+a}}{2} \right) - a \right] \right\}. \quad (30.101)$$

Well after matter-radiation equality, $a \gg 1$, the dark matter overdensity (30.101) is (note that for large scale modes $k^2 a$ can be small even when $a \gg 1$)

$$\delta_c - 3\Phi = 3\zeta_c \left(1 + \frac{4}{15}k^2 a \right) = 3\zeta_c \left(1 + \frac{1}{30}(k\eta)^2 \right) \quad \text{for } a \gg 1. \quad (30.102)$$

Since $\Phi_{\text{super}}(\text{late}) = -\frac{3}{5}\zeta_c$ for both adiabatic and isocurvature modes, equation (30.67), the overdensity δ_c from equation (30.102) is

$$\delta_c = \frac{6}{5}\zeta_c \left(1 + \frac{2}{3}k^2 a \right) = \frac{6}{5}\zeta_c \left(1 + \frac{1}{12}(k\eta)^2 \right) \quad \text{for } a \gg 1. \quad (30.103)$$

The solution for $\Theta_0 - \Phi$ driven by the superhorizon potential is, from the Green's function solution (30.51),

$$\begin{aligned} \Theta_0 - \Phi = \frac{1}{3}\zeta_r(4 - \cos y) + \left(\frac{1}{3}\zeta_r - \frac{3}{10}\zeta_c \right) \left\{ -4 + (4 - y_k^2) \cos y + 3y_k \sin y + y_k^2 \left[\frac{y_k}{y + y_k} \right. \right. \\ \left. \left. - y_k f(y_k + y) + 2g(y_k + y) + (y_k \cos y - 2 \sin y) f(y_k) - (2 \cos y + y_k \sin y) g(y_k) \right] \right\}, \end{aligned} \quad (30.104)$$

where $y \equiv k\eta_s$ is the wavenumber times the sound horizon distance, $y_k \equiv 4\sqrt{2/3}k$ is a constant proportional

to the wavenumber k , and $f(y)$ and $g(y)$ are the auxiliary sin/cosine integrals, related to the sin and cosine integrals $\text{Si } y \equiv \int_0^y \sin x \, dx/x$ and $\text{Ci } y \equiv \int_\infty^y \cos x \, dx/x$ by

$$f(y) \equiv (\pi/2 - \text{Si } y) \cos y + \text{Ci } y \sin y , \quad (30.105a)$$

$$g(y) \equiv (\pi/2 - \text{Si } y) \sin y - \text{Ci } y \cos y . \quad (30.105b)$$

The mode enters the horizon $y = 1$ at a cosmic scale factor of $a/a_{\text{eq}} = 4(1 + y_k)/y_k^2$. Figure 30.13 shows $\Theta_0 - \Phi$ and -2Φ from equations (30.104) and (30.65), for a mode with $k = \sqrt{3/8} = 0.61$, corresponding to $y_k = 2$. This mode enters the horizon at $a/a_{\text{eq}} = 3$, at approximately the epoch of recombination.

Concept question 30.13 Does the radiation monopole oscillate after recombination? Before recombination, photons and baryons are tightly coupled by electron scattering, and behave as a single fluid. After recombination, photons stream freely. Does the radiation monopole $\Theta_0 - \Phi$ keep oscillating after recombination, as in Figure 30.13, or does it stop oscillating, or does it do something else? **Answer.** The radiation monopole keeps oscillating, but differently. Two key differences in the free-streaming regime are, firstly, that the effective sound speed increases to the speed of light, and secondly, that the oscillations damp adiabatically. See Exercise 32.7 for an approximate treatment of a relativistic fluid — neutrinos — in the free-streaming regime. A full treatment of radiation in the free-streaming regime requires the radiative transfer equation, §34.1.

30.15 Matter-dominated regime

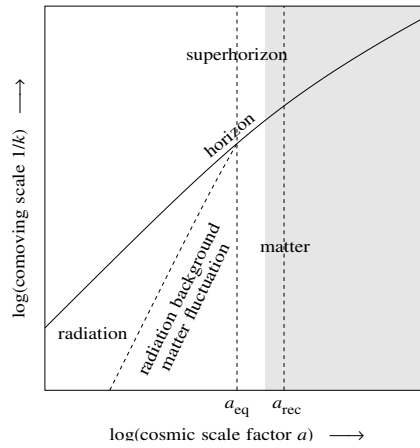


Figure 30.12 Matter-dominated regime.

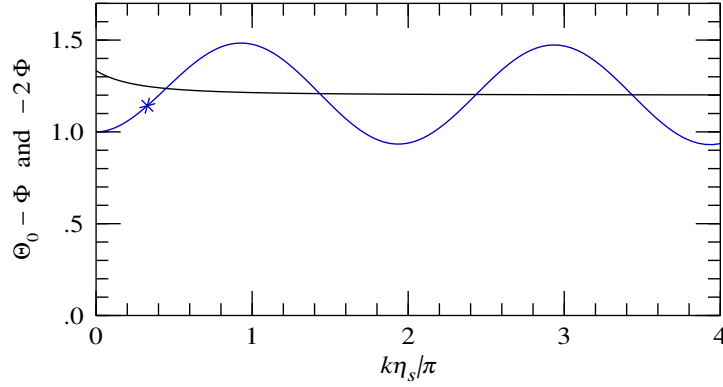


Figure 30.13 Similar to Figure 30.7, but for a mode that enters the sound horizon $k\eta_s = 1$ during the matter-dominated regime, for adiabatic initial conditions. The mode shown has $k = \sqrt{3/8} = 0.61$, which enters the horizon at $a = 3$, approximately the time of recombination, marked by a star.

After matter-radiation equality, but before curvature or dark energy become important, nonrelativistic matter dominates the mass-energy density of the Universe.

In the matter-dominated epoch, the relevant equations are, from equations (30.52), (30.54), and (30.56),

$$\dot{\delta}_c - k v_c = 3 \dot{\Phi} , \quad (30.106a)$$

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 \bar{\rho}_c \delta_c , \quad (30.106b)$$

$$-kF = 4\pi G a^2 \bar{\rho}_c v_c , \quad (30.106c)$$

in which, because it simplifies the mathematics, the Einstein momentum equation is used as a substitute for the matter velocity equation. In the matter-dominated epoch, the horizon is proportional to the square root of the cosmic scale factor, $\eta \propto a^{1/2}$, equation (30.38). Inserting δ_c and v_c from the Einstein energy and momentum equations (30.106b) and (30.106c) into the matter density equation (30.106a) yields a second order differential equation for the potential Φ

$$\ddot{\Phi} + \frac{6}{\eta} \dot{\Phi} = 0 . \quad (30.107)$$

The general solution of equation (30.107) is a linear combination

$$\Phi = C_{\text{grow}} \Phi_{\text{grow}} + C_{\text{decay}} \Phi_{\text{decay}} \quad (30.108)$$

of growing and decaying solutions

$$\Phi_{\text{grow}} = 1 , \quad \Phi_{\text{decay}} = y^{-5} , \quad (30.109)$$

where the dimensionless parameter y is, as previously, equation (30.75), the wavenumber k multiplied by the

sound horizon distance $\eta/\sqrt{3}$. In the matter-dominated regime y is, in units $a_{\text{eq}} = H_{\text{eq}} = 1$,

$$y \equiv \frac{k\eta}{\sqrt{3}} = 2\sqrt{\frac{2}{3}} ka^{1/2} . \quad (30.110)$$

The constants C_{grow} and C_{decay} in the solution (30.108) depend on conditions established before the matter-dominated epoch. The corresponding growing and decaying modes for the dark matter overdensity δ_c are, from the Einstein energy equation (30.106b),

$$(\delta_c - 3\Phi)_{\text{grow}} = -\left(5 + \frac{1}{2}y^2\right) \Phi_{\text{grow}} = -\left(5 + \frac{4}{3}k^2a\right) \Phi_{\text{grow}} , \quad (30.111a)$$

$$(\delta_c - 3\Phi)_{\text{decay}} = -\frac{1}{2}y^2 \Phi_{\text{decay}} = -\frac{4}{3}k^2a \Phi_{\text{decay}} . \quad (30.111b)$$

The behaviour of the growing and decaying modes (30.111) agrees with both the subhorizon Meszaros solution (30.97) and the superhorizon solution (30.102) well after matter-radiation equality $a \gg 1$, as they should. Any admixture of the decaying solution tends quickly to decay away, leaving the growing solution.

30.16 Baryons post-recombination

Recombination frees baryons and photons from each other's grasp. Starting at recombination, the freed baryons behaved as pressureless matter, like non-baryonic dark matter. In Exercise 30.14 you will figure out the behaviour of baryons and non-baryonic cold matter in the approximation that the Universe is matter-dominated.

Exercise 30.14 Growth of baryon fluctuations after recombination.

1. **Growing and decaying modes.** Assume that the Universe was matter-dominated at and after recombination. What are the growing and decaying solutions for the matter fluctuations δ_m ?
2. **Green's function for matter fluctuations.** Find the Green's function for any matter component subject to the initial conditions that the overdensity and its derivative with respect to cosmic scale factor a are $\delta_m(\text{rec})$ and $\delta'_m(\text{rec})$ at recombination.
3. **Initial conditions for dark matter and baryon fluctuations at recombination.** What are appropriate initial conditions at recombination for fluctuations in each of the two matter components, non-baryonic dark matter and baryons? Consider separately small-scale modes that entered the horizon well before matter-radiation equality, and large-scale modes that entered the horizon well after matter-radiation equality.
4. **Growth of dark matter and baryon fluctuations.** The matter density fluctuation is a sum of non-baryonic dark matter and baryonic contributions,

$$\delta_m = f_c \delta_c + f_b \delta_b , \quad (30.112)$$

where the constants f_c and f_b are the dark matter and baryon fractions

$$f_c \equiv \frac{\rho_c}{\rho_m} = 1 - f_b , \quad f_b \equiv \frac{\rho_b}{\rho_m} . \quad (30.113)$$

Use the Green's function with the chosen initial conditions to derive solutions for the dark matter and baryon overdensities δ_c and δ_b after recombination. Sketch the solutions for the matter, dark matter, and baryon overdensities through recombination.

5. **Comment.** A common statement is “Following recombination, baryons fall into the dark matter potential wells.” Comment, in the light of your solutions.
-

30.17 Matter with dark energy

Some time after recombination, dark energy becomes important. Observational evidence suggests that the dominant energy-momentum component of the Universe today is dark energy, with an equation of state consistent with that of a cosmological constant, $p_\Lambda = -\rho_\Lambda$. In what follows, dark energy is taken to have constant density, and therefore to be synonymous with a cosmological constant. Since dark energy has a constant energy density whereas matter density declines as a^{-3} , dark energy becomes important only well after recombination.

Dark energy does not cluster gravitationally, so the Einstein equations for the perturbed energy-momentum depend only on the matter fluctuation. However, dark energy does affect the evolution of the cosmic scale factor a . In fact, if matter is taken to be the only source of perturbation, then covariant energy-momentum conservation, as enforced by the Einstein equations, implies that the only addition that can be made to the unperturbed background is dark energy, with constant energy density. To see this, consider the equations (30.52) governing the matter overdensity δ_m and scalar velocity v_m (now subscripted m, since post-recombination matter includes baryons as well as non-baryonic cold dark matter), together with the Einstein energy and momentum equations (30.54) and (30.56) sourced only by matter,

$$\dot{\delta}_m - k v_m = 3\dot{\Phi} , \quad (30.114a)$$

$$\dot{v}_m + \frac{\dot{a}}{a} v_m = -k\Phi , \quad (30.114b)$$

$$-3\frac{\dot{a}}{a}F - k^2\Phi = 4\pi Ga^2\bar{\rho}_m\delta_m , \quad (30.114c)$$

$$-kF = 4\pi Ga^2\bar{\rho}_m v_m . \quad (30.114d)$$

The factor $4\pi Ga^2\bar{\rho}_m$ on the right hand side of the two Einstein equations can be written

$$4\pi Ga^2\bar{\rho}_m = \frac{3a_0^3 H_0^2 \Omega_m}{2a} , \quad (30.115)$$

where a_0 and H_0 are the present-day cosmic scale factor and Hubble parameter, and Ω_m is the present-day matter density (a constant). Allow the Hubble parameter $H(a) \equiv \dot{a}/a$ to be an arbitrary function of cosmic scale factor a . Inserting δ_m and velocity v_m from the Einstein energy and momentum equations (30.114c) and (30.114d) into the matter equations (30.114a) and (30.114b), and taking k times the overdensity equa-

tion (30.114a) minus $3\dot{a}/a$ times the velocity equation (30.114b), yields the condition

$$a^4 \frac{dH^2}{da} + 3a_0^3 H_0^2 \Omega_m = 0 , \quad (30.116)$$

whose solution is

$$\frac{H^2}{H_0^2} = \frac{\Omega_m}{(a/a_0)^3} + \Omega_\Lambda \quad (30.117)$$

for some constant Ω_Λ . This shows that, as claimed, if only matter perturbations are present, then the unperturbed background can contain, besides matter, only dark energy with constant density $\bar{\rho}_\Lambda = H_0^2 \Omega_\Lambda / (\frac{8}{3}\pi G)$. The result is a consequence of the fact that the Einstein equations enforce covariant conservation of energy-momentum.

With the Hubble parameter given by equation (30.117), the matter and Einstein equations (30.114) yield a second order differential equation with respect to cosmic scale factor a for the potential Φ , in units $a_0 = 1$:

$$2a(\Omega_m + a^3 \Omega_\Lambda) \Phi'' + (7\Omega_m + 10a^3 \Omega_\Lambda) \Phi' + 6a^2 \Omega_\Lambda \Phi = 0 . \quad (30.118)$$

The growing and decaying solutions to equation (30.118) are, in units $a_0 = 1$,

$$\Phi_{\text{grow}} = \frac{5\Omega_m H_0^2}{2} \frac{H(a)}{a} \int_0^a \frac{da'}{a'^3 H(a')^3} , \quad (30.119a)$$

$$\Phi_{\text{decay}} = \frac{H}{a} . \quad (30.119b)$$

The factor $\frac{5}{2}\Omega_m H_0^2$ in the growing solution is chosen so that $\Phi_{\text{grow}} \rightarrow 1$ as $a \rightarrow 0$. The growing solution Φ_{grow} can be expressed as an elliptic integral. The corresponding growing and decaying solutions for the matter overdensity δ_m are, again in units $a_0 = 1$,

$$(\delta_m - 3\Phi)_{\text{grow}} = -\frac{2k^2 a}{3\Omega_m H_0^2} \Phi_{\text{grow}} - 5 , \quad (\delta_m - 3\Phi)_{\text{decay}} = -\frac{2k^2 a}{3\Omega_m H_0^2} \Phi_{\text{decay}} . \quad (30.120)$$

For modes well inside the horizon, $k\eta \sim ka^{1/2}/H_0 \gg 1$, the relation (30.120) agrees with that (30.126) below.

30.18 Matter with dark energy and curvature

Curvature may also play a role after recombination. Since 2000, when the angular scale of the first peak in the CMB was resolved by the Boomerang balloon-based experiment (de Bernardis et al., 2000), observational evidence has been stubbornly consistent with the Universe having zero curvature. But it is possible that there may be some small curvature. If the curvature is significantly non-zero today (larger than treatable in perturbation theory), then by definition the curvature scale is less than the horizon size today. Scales larger than the curvature scale should strictly be treated using an unperturbed FLRW metric with curvature. However, a flat background FLRW metric remains a good approximation for modes whose scales are small compared to the curvature.

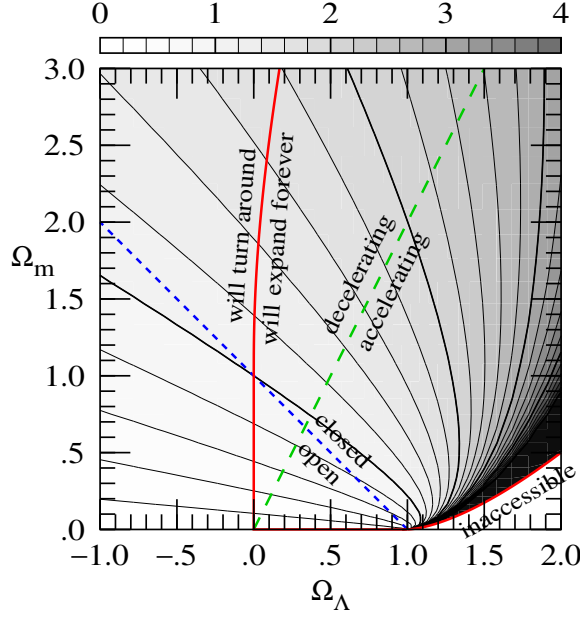


Figure 30.14 Contour plot of the growth factor $g(a)$ in a universe containing matter, curvature, and a cosmological constant. If the Universe is flat, $\Omega_k = 0$, then the Universe evolves from matter-dominated ($\Omega_m = 1$, $\Omega_\Lambda = 0$) to Λ -dominated ($\Omega_m = 0$, $\Omega_\Lambda = 1$) along the (blue) dashed line.

Concept question 30.15 Curvature scale. What is meant by the curvature scale? Is the curvature scale constant in comoving coordinates?

For modes much smaller than the horizon distance today, the time derivative of the potential can be neglected compared to its spatial gradient, $|\dot{\Phi}| \ll |k\Phi|$. If only matter, curvature, and dark energy are present, then only matter fluctuations contribute to the energy-momentum. At scales much less than the curvature scale, equations (30.114) then go over to the Newtonian limit,

$$\dot{\delta}_m - k v_m = 0, \quad (30.121a)$$

$$\dot{v}_m + \frac{\dot{a}}{a} v_m = -k\Phi, \quad (30.121b)$$

$$-k^2\Phi = 4\pi G a^2 \bar{\rho}_m \delta_m. \quad (30.121c)$$

The factor $4\pi G a^2 \bar{\rho}_m$ in the Einstein equation can be written as equation (30.115). The matter and Einstein equations (30.121) yield a second order equation for the matter overdensity δ_m , in units $a_0 = 1$:

$$\ddot{\delta}_m + \frac{\dot{a}}{a} \dot{\delta}_m - \frac{3\Omega_m H_0^2}{2} \frac{\delta_m}{a} = 0. \quad (30.122)$$

Equation (30.122) can be recast as a differential equation with respect to cosmic scale factor a :

$$\delta_m'' + \left(\frac{H'}{H} + \frac{3}{a} \right) \delta_m' - \frac{3\Omega_m H_0^2}{2} \frac{\delta_m}{a^5 H^2} = 0 , \quad (30.123)$$

where $H \equiv \dot{a}/a$ is the Hubble parameter, and prime ' denotes differentiation with respect to a . In the case of matter plus curvature plus dark energy, the Hubble parameter H satisfies, again in units $a_0 = 1$,

$$\frac{H^2}{H_0^2} = \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda , \quad (30.124)$$

where Ω_m , Ω_k , and Ω_Λ are the (constant) present-day values of the matter, curvature, and dark energy densities. The growing and decaying solutions to equation (30.123) are

$$\delta_{m,\text{grow}} \equiv a g(a) = \frac{5\Omega_m H_0^2}{2} H(a) \int_0^a \frac{da'}{a'^3 H(a')^3} , \quad \delta_{m,\text{decay}} = \frac{H}{H_0} . \quad (30.125)$$

The potential Φ is related to the matter overdensity δ_m by, again in units $a_0 = 1$, equation (30.121c),

$$\Phi = -\frac{3\Omega_m H_0^2}{2k^2} \frac{\delta_m}{a} . \quad (30.126)$$

The observationally relevant solution is the growing mode. The growing mode is conventionally given a special notation, the growth factor $g(a)$, because of its importance to relating the amplitude of clustering at various times, from recombination up to the present. For the growing mode,

$$\delta \propto a g(a) , \quad \Phi \propto g(a) . \quad (30.127)$$

The normalization factor $\frac{5}{2}\Omega_m H_0^2$ in equation (30.125) is chosen so that in the matter-dominated phase after recombination but before dark energy or curvature become important, the growth factor $g(a)$ is unity,

$$g(a) = 1 \quad (a_{\text{rec}} \ll a \ll 1) . \quad (30.128)$$

Thus as long as the Universe remains matter-dominated, the potential Φ remains constant. Curvature or dark energy causes the potential Φ to decrease. Figure 30.14 illustrates the growth factor $g(a)$ as a function of Ω_m and Ω_Λ .

It should be emphasized that the growing and decaying solutions (30.125) are valid *only* for the case of matter plus curvature plus constant density dark energy, where the Hubble parameter takes the form (30.124). If another kind of mass-energy is considered, such as dark energy with non-constant density, then equations governing perturbations of the other kind must be adjoined, and the Einstein equations modified accordingly.

The growth factor $g(a)$ may be expressed analytically as an elliptic function. A good analytic approximation is (Carroll, Press, and Turner, 1992)

$$g \approx \frac{5\Omega_m}{2 \left[\Omega_m^{4/7} - \Omega_\Lambda + \left(1 + \frac{1}{2}\Omega_m \right) \left(1 + \frac{1}{70}\Omega_\Lambda \right) \right]} , \quad (30.129)$$

where Ω_x are densities at the epoch being considered (such as the present, $a = a_0$).

30.19 Primordial power spectrum

Initial conditions from inflation are conveniently characterized in terms of the gauge-invariant fluctuation ζ defined by equation (29.42), which has the property that it remains constant during evolution at super-horizon scales. The fluctuation ζ is commonly called the primordial curvature fluctuation. According to the inflationary paradigm, fluctuations in ζ are generated by quantum fluctuations in the inflaton field that drives inflation. The amplitude ζ of a mode freezes as the mode exits the horizon during inflation, and remains constant until the mode subsequently re-enters the horizon after inflation has ended.

Generically, inflation predicts that primordial curvature fluctuations ζ generated by vacuum fluctuations during inflation have a spectrum that is (1) Gaussian, and (2) scale-free. Inflation also predicts generically that the fluctuations are adiabatic, meaning that the curvature fluctuation is the same for all species, $\zeta_x = \zeta$ for all species x .

Gaussian distributions, §30.22.6, are ubiquitous in statistics as a consequence of the Central Limit Theorem (CLT), §30.22.5. The CLT states that the distribution of a random variable that is a sum of independent random increments is asymptotically Gaussian in the limit of a large number of increments. A Gaussian distribution is characterized entirely by its mean and variance, all higher irreducible moments vanishing.

A scale-free spectrum of fluctuations is one in which the spatial variance ξ_ζ of the dimensionless fluctuation ζ is the same on all scales,

$$\langle \zeta(\mathbf{x}') \zeta(\mathbf{x}) \rangle \equiv \xi_\zeta(|\mathbf{x}' - \mathbf{x}|) = \text{constant} , \quad (30.130)$$

independent of spatial separation $|\mathbf{x}' - \mathbf{x}|$. Inflation predicts a scale-free spectrum because the vacuum energy that drives inflation is constant in time, and quantum fluctuations in the vacuum remain statistically the same as time goes by. Thus the characteristic amplitude of fluctuations ζ flying over the horizon remains the same as time goes by.

The power spectrum $P_\zeta(k)$ of fluctuations in ζ is defined by

$$\langle \zeta(\mathbf{k}') \zeta(\mathbf{k}) \rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) P_\zeta(k) . \quad (30.131)$$

The “momentum-conserving” Dirac delta function $(2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k})$ in equation (30.131) is a consequence of the assumed statistical spatial translation symmetry of fluctuations in the spatially homogeneous FLRW background. The power spectrum $P_\zeta(k)$ is related to the correlation function $\xi_\zeta(x)$ by (with the standard convention in cosmology for the choice of signs and factors of 2π)

$$P_\zeta(k) = \int e^{i\mathbf{k} \cdot \mathbf{x}} \xi_\zeta(x) d^3x , \quad \xi_\zeta(x) = \int e^{-i\mathbf{k} \cdot \mathbf{x}} P_\zeta(k) \frac{d^3k}{(2\pi)^3} . \quad (30.132)$$

Whereas the correlation function $\xi_\zeta(x)$ is dimensionless, the power spectrum $P(k)$ has units of comoving length cubed. The scale-free character means that the dimensionless power spectrum $\Delta_\zeta^2(k)$ defined by

$$\Delta_\zeta^2(k) \equiv P_\zeta(k) \frac{4\pi k^3}{(2\pi)^3} \quad (30.133)$$

is constant.

Actually, the power spectrum generated by inflation is not precisely scale-free, because inflation comes to

an end, which breaks scale-invariance. The departure from scale-invariance is conventionally characterized by a scalar spectral index, the tilt n_s , such that

$$\Delta_\zeta^2(k) \propto k^{n_s-1} . \quad (30.134)$$

Thus a scale-invariant power spectrum has

$$n_s = 1 \quad (\text{scale-invariant}) . \quad (30.135)$$

The value $n_s = 1$ is a historical hangover. It was originally proposed as a natural initial condition, before the idea of inflation was conceived, by Harrison (1970) and Zeldovich (1972) who pointed out that density fluctuations satisfying $\nabla^2 \Phi \propto \text{constant}$ would have $P_\Phi(k) \propto k^{-4}$, hence $n_s = 1$. The tilt predicted by inflationary models depends on the shape of the scalar potential $V(\varphi)$ of the scalar field φ that drove inflation, and is generically predicted to be close to but slightly less than 1, equation (38.79).

A common practice is to report the value of the dimensionless primordial power spectrum $\Delta_\zeta^2(k)$ at some pivot scale k_p ,

$$\Delta_\zeta^2(k) = \Delta_\zeta^2(k_p) \left(\frac{k}{k_p} \right)^{n_s-1} . \quad (30.136)$$

The Planck Collaboration (2018) report

$$\Delta_\zeta^2(k_p = 0.05 \text{ Mpc}^{-1}) = (2.14 \pm 0.05) \times 10^{-9} , \quad n_s = 0.965 \pm 0.004 . \quad (30.137)$$

The pivot scale k_p was chosen in this case so that the error in the amplitude $\Delta_\zeta^2(k_p)$ was uncorrelated with the error in the tilt n_s .

30.20 Matter power spectrum

The **matter power spectrum** $P_m(\eta, k)$ at time η is defined by

$$\langle \delta_m(\eta, \mathbf{k}') \delta_m(\eta, \mathbf{k}) \rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) P_m(\eta, k) , \quad (30.138)$$

the Dirac delta function being as before a consequence of the assumption of statistical spatial homogeneity. The assumption of statistical isotropy implies that the power spectrum $P_m(\eta, k)$ is a function only of the magnitude k of the wavevector $\mathbf{k}$. The matter power spectrum $P_m(\eta, k)$ is related to the primordial power spectrum $P_\zeta(k)$ by

$$P_m(\eta, k) = T_m(\eta, k)^2 P_\zeta(k) = T_m(\eta, k)^2 \frac{(2\pi)^3}{4\pi k^3} \Delta_\zeta^2(k) , \quad (30.139)$$

where $T_m(\eta, k)$ is the **matter transfer function** defined by

$$T_m(\eta, k) \equiv \frac{\delta_m(\eta, \mathbf{k})}{\zeta(\mathbf{k})} . \quad (30.140)$$

The transfer function $T_m(\eta, k)$ for any given cosmological model may be calculated by the methods expounded in the bulk of this Chapter, Exercise 30.16.

The predictions of cosmological models of the matter power spectrum may be compared to measurements of the power spectrum of objects, such as galaxies, that may trace the matter distribution. Galaxy surveys probe the matter distribution well after recombination, and at scales much less than the horizon distance today. Under those circumstances, the matter transfer function $T_m(\eta, k)$ factors into a product of three factors: (1) a factor relating the matter overdensity δ_m to the potential Φ , which in the Newtonian regime at subhorizon scales well after recombination is given in units $a_0 = 1$ by equation (30.126); (2) a growth factor $g(a)$, equation (30.125), relating the potential $\Phi(\eta)$ at recent times η to the post-recombination matter-dominated potential $\Phi(\text{late})$; (3) a transfer function $T_{\Phi(\text{late})}(k)$ relating the matter-dominated potential $\Phi(\text{late})$ to the primordial fluctuation ζ :

$$\begin{aligned} T_m(\eta, k) &= \frac{\delta_m(\eta, \mathbf{k})}{\Phi(\eta, \mathbf{k})} \times \frac{\Phi(\eta, \mathbf{k})}{\Phi(\text{late}, \mathbf{k})} \times \frac{\Phi(\text{late}, \mathbf{k})}{\zeta(\mathbf{k})} \\ &= - \left(\frac{2ak^2}{3\Omega_m H_0^2} \right) \times g(a) \times T_{\Phi(\text{late})}(k) , \end{aligned} \quad (30.141)$$

where

$$T_{\Phi(\text{late})}(k) \equiv \frac{\Phi(\text{late}, \mathbf{k})}{\zeta(\mathbf{k})} . \quad (30.142)$$

The potential transfer function $T_{\Phi(\text{late})}(k)$ is independent of time η because the potential $\Phi(\text{late})$ is constant in the matter-dominated regime before dark energy (or curvature) becomes important. The factorization (30.141) of the matter transfer function $T_m(\eta, k)$ separates the dependence on time η (or equivalently cosmic scale factor a) and wavenumber k . The first factor δ_m/Φ is proportional to ak^2 ; the second is a function $g(a)$ only of cosmic scale factor a ; and the third is a function $T_{\Phi(\text{late})}(k)$ only of wavenumber k .

The factorization (30.141) of the matter transfer function $T_m(\eta, k)$ implies that the matter power spectrum $P_m(\eta, k)$, equation (30.139), is related to the primordial power spectrum $P_\zeta(\eta, k)$ by

$$P_m(\eta, k) = \left(\frac{2ag(a)}{3\Omega_m H_0^2} \right)^2 k^4 T_{\Phi(\text{late})}(k)^2 P_\zeta(k) = \left(\frac{2ag(a)}{3\Omega_m H_0^2} \right)^2 k T_{\Phi(\text{late})}(k)^2 \frac{(2\pi)^3}{4\pi} \Delta_\zeta^2(k) . \quad (30.143)$$

For a power-law primordial spectrum (30.134), the matter power spectrum at the largest scales, where the potential transfer function $T_{\Phi(\text{late})}(k)$ is a constant independent of k , goes as

$$P_m(\eta, k) \propto k^n . \quad (30.144)$$

The proportionality (30.144) explains the origin of the scalar index n_s .

Exercise 30.16 Power spectrum of matter fluctuations: simple approximation. Use the code you wrote in Exercise 30.11 to compute the matter transfer function $T_m(\eta, k)$, equation (30.140). Deduce the matter power spectrum $P_m(\eta_0, k)$, equation (30.139), at the present time, $\eta = \eta_0$. Use the normalization and tilt of primordial power measured from Planck, equation (30.137). Compute power spectra for a concordance Λ CDM model, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, and for a flat matter-dominated Universe, $\Omega_m = 1$. Compare

your matter power spectrum to data from Gil-Marín et al. (2020), downloadable from https://svn.sdss.org/public/data/eboss/DR16cosmo/tags/v1_0_1/dataveccov/lrg_elg_qso/LRG_Pk/. The best data sets are the “post-reconstruction” sets. The “reconstruction” involves undoing at least some of the effects of nonlinear evolution by moving galaxies around. Note the units of the data: wavenumber k in $h \text{ Mpc}^{-1}$ and power $P(k)$ in $(h^{-1} \text{ Mpc})^3$, with $h \equiv H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$.

As in Exercise 30.11, you may find that your integration routine gets stuck trying to integrate the oscillating radiation monopole and dipole once the mode is well inside the horizon, $k\eta \gg 1$. The strategy suggested in Exercise 30.11 was to modify the radiation dipole equation (30.53b) by introducing an artificial damping term, equation (30.58), that damps radiation once it is well inside the horizon. Since the radiation fluctuation ceases to influence the gravitational potential or the matter fluctuation once the radiation has oscillated many times, the artificial damping has little effect on the model power spectrum.

A second problem you will encounter is that of power at superhorizon scales. Astronomers on Earth cannot measure power at scales larger than our horizon because they cannot distinguish a superhorizon fluctuation from a change in the mean density of the background FLRW geometry. To eliminate the unmeasurable superhorizon power, calculate power from the overdensity $\delta_k - \delta_0$ with a large-scale constant δ_0 subtracted.

A third problem is that galaxies do not necessarily trace the distribution of matter. A simple model is to suppose a linear relation between galaxy overdensity δ_g and matter overdensity δ_m (in Fourier space),

$$\delta_g = b\delta_m , \quad (30.145)$$

where b is the **bias** parameter. Linear bias was introduced by Kaiser (1984), who showed that regions of a Gaussian field (§30.22.3) above a high threshold density are linearly biased.

Solution. See Figure 30.15. One of the trickier issues is getting the units right. The SDSS IV data are given in units where the length scale is such that the comoving Hubble distance at the present time is

$$\frac{c}{a_0 H_0} = \frac{299,792.458 \text{ km s}^{-1}}{100 h \text{ km s}^{-1} \text{ Mpc}} = 2,997.92458 h^{-1} \text{ Mpc}^{-1} . \quad (30.146)$$

My code worked in units where $c = a_{\text{eq}} = H_{\text{eq}} = 1$. With Ω_x representing values at the present time, the Hubble parameter now H_0 and at matter-radiation equality H_{eq} are related by

$$\frac{H_{\text{eq}}}{H_0} = \sqrt{\Omega_r(a_{\text{eq}}/a_0)^{-4} + \Omega_m(a_{\text{eq}}/a_0)^{-3} + \Omega_k(a_{\text{eq}}/a_0)^{-2} + \Omega_\Lambda} . \quad (30.147)$$

I chose $a_0/a_{\text{eq}} = 3400$, and present-day densities of $\Omega_m = 0.29$, $\Omega_r = \Omega_m/3400$, $\Omega_k = 0$, $\Omega_\Lambda = 1 - \Omega_r - \Omega_m - \Omega_k$. The code gave $H_0/H_{\text{eq}} = 6.4 \times 10^{-6}$, and so

$$\frac{c}{a_0 H_0} = \frac{1}{3400 \times (6.4 \times 10^{-6})} = 46.0 \text{ program units} . \quad (30.148)$$

The conversion factor between $h^{-1} \text{ Mpc}$ and program length units was therefore

$$1 \text{ program length unit} = \frac{2,997.92458 h^{-1} \text{ Mpc}}{46.0} = 65.1 h^{-1} \text{ Mpc} . \quad (30.149)$$

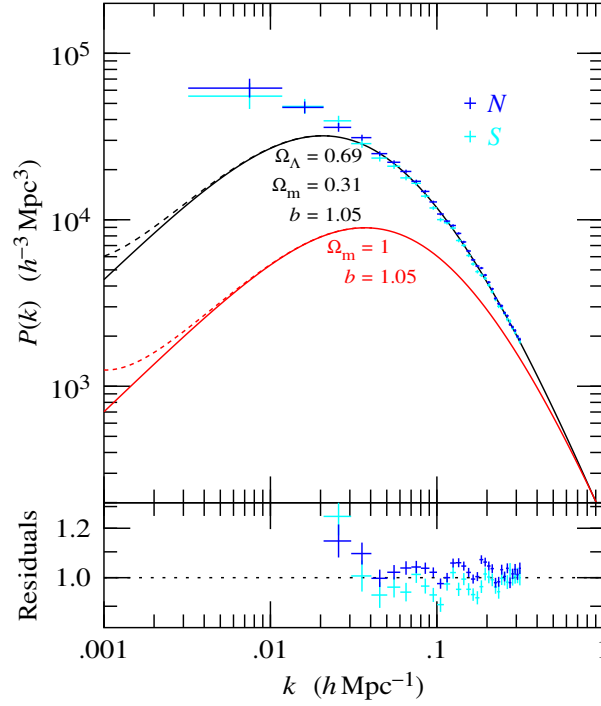


Figure 30.15 Model matter power spectra computed in the simple approximation, compared to observations from the North (N) and South (S) Galactic Caps of the Sloan Digital Sky Survey IV (Gil-Marín et al., 2020). The data comprise 377,458 luminous red galaxies covering approximately 18% of the sky over redshifts $z = 0.6$ – 1 . Two models are shown, a flat Λ CDM model with concordance parameters $\Omega_\Lambda = 0.69$ and $\Omega_m = 0.31$, and a flat matter-only CDM model, $\Omega_m = 1$. The dashed lines on the models show power calculated from $|\delta_k^2|$, which includes unmeasurable superhorizon power; the solid lines are calculated from $|\delta_k - \delta_0|^2$, which excludes the unmeasurable superhorizon power by subtracting a constant δ_0 from the overdensity. The Λ CDM model is normalized to the amplitude (30.137) measured by Planck (Planck Collaboration, 2018), multiplied by a bias squared factor of $b^2 = 1.05^2$. The Λ CDM power spectrum calculated here in the simple approximation may be compared to the corresponding power spectra in the hydrodynamic approximation, Figure 32.4, and from a Boltzmann computation, Figure 33.5.

For the wavenumber, this meant that the conversion between $h \text{ Mpc}^{-1}$ and program units was

$$k_{h \text{ Mpc}^{-1}} = \frac{k_{\text{prog}}}{65.1 h^{-1} \text{ Mpc}} . \quad (30.150)$$

The model power spectrum in Figure 30.15 has been multiplied, arbitrarily, by a squared bias factor of $b^2 = 1.05^2$, to give a better fit to the observed power spectrum. The residual difference between observed and model power shows wiggles. These are baryon acoustic oscillations (BAO), the presence of which is predicted when baryons are included, Figure 32.4.

30.21 Nonlinear evolution of the matter power spectrum

This Chapter has assumed throughout that linear perturbation theory holds, which requires that fluctuations be small, $\delta \ll 1$. This assumption fails for matter fluctuations at small scales, which in due course collapse into galaxies, with matter densities much greater than the mean, $\delta_m \gg 1$. Evolution in this regime is nonlinear, and must usually be followed with large computer simulations.

Since gravity remains weak, $\Phi \ll 1$ and matter moves nonrelativistically even in the nonlinear regime, gravity remains well described by the Newtonian limit, equation (30.121c). The equations of conservation of mass and momentum still hold for the matter, but these equations are no longer linear. To the extent that the matter streams collisionlessly, as is the case for nonbaryonic dark matter, its nonlinear evolution is straightforward, if computationally intensive, to follow. However, the collisional dynamics of baryons leads to interesting and complicated phenomena, including stars, planets, and black holes, and people to worry about them.

30.22 Statistics of random fields

30.22.1 Random field

A basic proposition of modern cosmology, so far well-supported by observational evidence, is that fluctuations in the Universe originated from some random process that operated in the same fashion from place to place. In the inflationary paradigm, fluctuations originated as quantum fluctuations in the inflaton field that drove inflation. According to this proposition, the fluctuating density $\rho(\mathbf{x})$ of any measurable quantity (such as matter density, or radiation temperature) in our Universe constitutes a **random field**. The density $\rho(\mathbf{x})$ at a randomly chosen position $\mathbf{x}$ constitutes a **random variable** with some probability distribution $P(\rho)$ of finding the density to lie in an interval $d\rho$. By definition, the probability distribution $P(\rho)$ is positive, and normalized to unit total probability,

$$\int P(\rho) d\rho = 1 . \quad (30.151)$$

In a random field, the densities $\rho(\mathbf{x}_1)$ and $\rho(\mathbf{x}_2)$ at two different points are in general not independent, so the 1-point probability (30.151) is not sufficient to determine completely the statistical properties of the field. For example, since gravity causes matter to cluster, the densities at two nearby points are correlated, not independent. For brevity, denote the density at spatial position $\mathbf{x}_i$ by ρ_i ,

$$\rho_i \equiv \rho(\mathbf{x}_i) . \quad (30.152)$$

The properties of the random field $\rho(\mathbf{x})$ are determined by an infinite set of **N -point probability distributions** $P(\rho_1, \dots, \rho_N)$ of finding the densities ρ_i at N positions $\mathbf{x}_i$ to lie in an interval $d\rho_1 \dots d\rho_N$. By definition, the joint N -point probability distribution is positive, and normalized to unit total probability,

$$\int P(\rho_1, \dots, \rho_N) d\rho_1 \dots d\rho_N = 1 . \quad (30.153)$$

By homogeneity, the N -point probability is a function only of the relative spatial positions $\mathbf{x}_i$, not of their absolute positions.

The limitations of observational accessibility and accuracy mean that the true N -point probability distributions $P(\rho_1, \dots, \rho_N)$ are not known exactly. It is then necessary to make hypotheses about the form of the probability, and to test those hypotheses against the available sampling of data.

30.22.2 Random fields in Fourier space

Any linear combination $\tilde{\rho}_i$ of random fields ρ_j is a random field,

$$\tilde{\rho}_i = \sum_j a_{ij} \rho_j, \quad (30.154)$$

where a_{ij} are constants, and the sum over j could represent an integral over a continuum. In particular, the Fourier transform $\rho(\mathbf{k})$ of a random field $\rho(\mathbf{x})$,

$$\rho(\mathbf{k}) \equiv \int \rho(\mathbf{x}) e^{i\mathbf{k} \cdot \mathbf{x}} d^3x, \quad \rho(\mathbf{x}) \equiv \int \rho(\mathbf{k}) e^{-i\mathbf{k} \cdot \mathbf{x}} \frac{d^3k}{(2\pi)^3}, \quad (30.155)$$

is a random field.

The Fourier modes $\rho(\mathbf{k})$ of a random field are of special importance when the field is statistically homogeneous, because Fourier modes are eigenmodes of the translation operator ∇ , and the statistical properties of a statistically homogeneous random field commute with the translation operator.

30.22.3 Gaussian random fields

A generic prediction of inflation is that the primordial distribution of fluctuations was Gaussian, as a result of their origin as quantum fluctuations. Whenever the values $\rho(\mathbf{x})$ at each point $\mathbf{x}$ of a random field are generated as a sum of a large number of independent random increments, then the resulting field will be Gaussian, as a consequence of the Central Limit Theorem. The CLT is proved for the simple case of a single random variable ρ in §30.22.5.

A Gaussian random field $\rho(\mathbf{x})$ is defined by the vanishing of all irreducible moments other than the first two, the mean $\bar{\rho}$, and the variance C_{ij} ,

$$C_{ij} \equiv \langle \Delta\rho_i \Delta\rho_j \rangle, \quad (30.156)$$

where $\Delta\rho_i \equiv \rho_i - \bar{\rho}$ is the deviation of ρ_i from the mean. The mean $\bar{\rho}$ is a single number. The assumption of statistical homogeneity and isotropy implies that the variance is a function $C_{ij} = C(x_{ij})$ only of the separation $x_{ij} \equiv |\mathbf{x}_i - \mathbf{x}_j|$ of the points. The covariance C_{ij} defined by equation (30.156) has dimensions of $\bar{\rho}^2$. Commonly, a dimensionless version ξ_{ij} of the covariance is defined by dividing by $\bar{\rho}^2$.

The N -point probability distribution of a Gaussian random field is, generalizing the 1-point probability (30.175) derived below,

$$P(\rho_1, \dots, \rho_N) d\rho_1 \dots d\rho_N = \frac{1}{\sqrt{(2\pi)^N |C_{ij}|}} \exp\left(-\frac{1}{2} C_{ij}^{-1} \Delta\rho_i \Delta\rho_j\right) d\rho_1 \dots d\rho_N, \quad (30.157)$$

where $|C_{ij}|$ is the determinant of the covariance matrix.

Any linear combination $\sum_j a_{ij}\rho_j$ of Gaussian random fields ρ_j is also a Gaussian random field. In particular, the Fourier transform $\rho(\mathbf{k})$ of a Gaussian random field $\rho(\mathbf{x})$ is a Gaussian random field.

30.22.4 Moment-generating functions

The proof of the Central Limit Theorem, §30.22.5, goes via moment-generating functions. For simplicity, moment-generating functions are defined in this section for a single random variable ρ , but the results generalize straightforwardly to a random field $\rho(\mathbf{x})$. The validity of the steps below requires that various integrals over the probability distribution $P(\rho)$ converge; any required convergence properties are tacitly assumed.

The random variable ρ has a positive probability distribution $P(\rho)$ normalized to unit total probability. The **moment-generating function** of the probability distribution $P(\rho)$ is defined to be

$$M(\mu) \equiv \int e^{\mu\rho} P(\rho) d\rho . \quad (30.158)$$

Expanding the exponential in the integrand as a power series in μ implies that the moment-generating function is

$$M(\mu) = 1 + \langle \rho \rangle \mu + \langle \rho^2 \rangle \frac{\mu^2}{2} + \langle \rho^3 \rangle \frac{\mu^3}{3!} + \dots , \quad (30.159)$$

where $\langle \rho^n \rangle$ is the n th moment of the probability distribution,

$$\langle \rho^n \rangle \equiv \int \rho^n P(\rho) d\rho . \quad (30.160)$$

Equation (30.159) accounts for the name moment-generating function.

Suppose that the measurement of ρ is repeated N times, and suppose that each measurement is independent of the others, meaning that the probability of measuring successive values $\rho_{(1)}, \dots, \rho_{(N)}$ is the product of probabilities (the subscripts are parenthesized to distinguish the i th observation $\rho_{(i)}$ from the i th position ρ_i)

$$P(\rho_{(1)}, \dots, \rho_{(N)}) = P(\rho_{(1)}) \dots P(\rho_{(N)}) . \quad (30.161)$$

The moment-generating function $M_N(\mu)$ of the sum $\sum_{i=1}^N \rho_{(i)}$ of N independent measurements $\rho_{(i)}$ is then the N th power of the moment-generating function $M(\mu)$,

$$\begin{aligned} M_N(\mu) &\equiv \int e^{\mu(\rho_{(1)} + \dots + \rho_{(N)})} P(\rho_{(1)}, \dots, \rho_{(N)}) d\rho_{(1)} \dots d\rho_{(N)} \\ &= \int e^{\mu\rho_{(1)}} P(\rho_{(1)}) d\rho_{(1)} \dots \int e^{\mu\rho_{(N)}} P(\rho_{(N)}) d\rho_{(N)} \\ &= M(\mu)^N . \end{aligned} \quad (30.162)$$

Thus the moment-generating function of a sum of independent measurements is multiplicative. The **irreducible-moment-generating function** $Z(\mu)$ is defined to be the logarithm of the moment-generating function,

$$Z(\mu) \equiv \ln [M(\mu)] . \quad (30.163)$$

In statistical mechanics, the irreducible-moment-generating function $Z(\mu)$ is called the partition function. Since the moment-generating function is multiplicative, the irreducible-moment-generating function $Z_N(\mu)$ of a sum $\sum_{i=1}^N \rho_{(i)}$ of N independent measurements $\rho_{(i)}$ is additive,

$$Z_N(\mu) = N Z(\mu) . \quad (30.164)$$

The coefficients of the series expansion of $Z(\mu)$ in μ define the irreducible moments κ_n ,

$$Z(\mu) = \mu \kappa_1 + \frac{\mu^2}{2!} \kappa_2 + \frac{\mu^3}{3!} \kappa_3 + \dots . \quad (30.165)$$

Unlike the moments $\langle \rho^n \rangle$, the irreducible moments κ_n have the important property of being additive over sums $\sum_{i=1}^N \rho_{(i)}$ of independent variables. The defining relation (30.163) between the irreducible $Z(\mu)$ and standard $M(\mu)$ moment-generating functions yields the relation between the irreducible moments κ_n and moments $\langle \rho^n \rangle$. The relations for the first few moments are, with $\Delta \rho \equiv \rho - \bar{\rho}$,

$$\kappa_1 = \bar{\rho} , \quad (30.166a)$$

$$\kappa_2 = \langle \Delta \rho^2 \rangle , \quad (30.166b)$$

$$\kappa_3 = \langle \Delta \rho^3 \rangle , \quad (30.166c)$$

$$\kappa_4 = \langle \Delta \rho^4 \rangle - 3 \langle \Delta \rho^2 \rangle^2 . \quad (30.166d)$$

The low order irreducible moments have names: the first, second, third, and fourth irreducible moments are called respectively the mean, variance, skewness, and kurtosis. Some works define skewness and kurtosis as the dimensionless combinations $\kappa_3/\kappa_2^{3/2}$ and κ_4/κ_2^2 .

More generally, the irreducible-moment-generating function $Z(\mu_i)$ of a random field $\rho(\mathbf{x})$ is

$$Z(\mu_i) = \mu_1 \kappa_1 + \frac{\mu_1 \mu_2}{2!} \kappa_{12} + \frac{\mu_1 \mu_2 \mu_3}{3!} \kappa_{123} + \dots , \quad (30.167)$$

where $\kappa_{1\dots n} \equiv \kappa(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is the n -point irreducible moment, also called the n -point correlation function. The first few correlation functions $\kappa_{1\dots n}$ are related to the moments $\langle \Delta \rho_1 \dots \Delta \rho_n \rangle$ of the distribution by

$$\kappa_1 = \bar{\rho} , \quad (30.168a)$$

$$\kappa_{12} = \langle \Delta \rho_1 \Delta \rho_2 \rangle , \quad (30.168b)$$

$$\kappa_{123} = \langle \Delta \rho_1 \Delta \rho_2 \Delta \rho_3 \rangle , \quad (30.168c)$$

$$\kappa_{1234} = \langle \Delta \rho_1 \Delta \rho_2 \Delta \rho_3 \Delta \rho_4 \rangle - \langle \Delta \rho_1 \Delta \rho_2 \rangle \langle \Delta \rho_3 \Delta \rho_4 \rangle - \langle \Delta \rho_1 \Delta \rho_3 \rangle \langle \Delta \rho_2 \Delta \rho_4 \rangle - \langle \Delta \rho_1 \Delta \rho_4 \rangle \langle \Delta \rho_2 \Delta \rho_3 \rangle . \quad (30.168d)$$

30.22.5 Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of averages of N independent measurements of a random variable is Gaussian in the limit of large N . The CLT generalizes to a random field, but for simplicity this section confines itself to the case of a single random variable ρ .

As shown in §30.22.4, irreducible moments are additive over sums of independent random variables. Thus the irreducible moment κ_n of a sum $\sum_{i=1}^N \rho_{(i)}$ of N independent variables $\rho_{(i)}$ goes as

$$\kappa_n \propto N . \quad (30.169)$$

The n th irreducible moment κ_n has units of $\bar{\rho}^n$. The shape of the probability distribution $P(\rho)$ can be characterized by dimensionless combinations of the irreducible moments. For example, the standard deviation σ , defined to be the square root of the variance, $\sigma \equiv \sqrt{\kappa_2} = \sqrt{\langle \Delta \rho^2 \rangle}$, has the dimension of $\bar{\rho}$. The standard deviation increases with the number N of independent measurements as $\sqrt{N}$, but the dimensionless ratio $\sigma/\bar{\rho}$ of the standard deviation to the mean decreases as $1/\sqrt{N}$,

$$\sigma \equiv \sqrt{\kappa_2} \propto \sqrt{N} , \quad \frac{\sigma}{\bar{\rho}} \equiv \frac{\sqrt{\kappa_2}}{\kappa_1} \propto \frac{1}{\sqrt{N}} . \quad (30.170)$$

This recovers the familiar result that the difference between the average $N^{-1} \sum_i \rho_{(i)}$ of a set of independent measurements and the true mean $\bar{\rho}$ decreases as $1/\sqrt{N}$ as the number N of measurements increases.

The shape of the probability distribution beyond its first and second irreducible moments can be characterized by the dimensionless ratio $\kappa_n^{1/n}/\kappa_2^{1/2}$ of the n th to 2nd irreducible moments. This ratio becomes small as the number N of independent measurements increases,

$$\frac{\kappa_n^{1/n}}{\kappa_2^{1/2}} \propto N^{1/n-1/2} \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \text{for } n \geq 3 . \quad (30.171)$$

The asymptotic behaviour (30.171) is the CLT: it says that higher order irreducible moments become negligible in the limit of large N .

30.22.6 Gaussian distribution

A Gaussian distribution is defined by the property that its only non-vanishing irreducible moments are the first two, the mean κ_1 and variance κ_2 . The third and higher irreducible moments of a Gaussian distribution vanish,

$$\kappa_n = 0 \quad (n \geq 3) \quad \text{Gaussian distribution} . \quad (30.172)$$

The irreducible-moment-generating function $Z(\mu)$ of a Gaussian distribution is, from equation (30.165),

$$Z(\mu) = \mu \bar{\rho} + \frac{\mu^2}{2} \langle \Delta \rho^2 \rangle . \quad (30.173)$$

Accordingly, the moment-generating function $M(\mu)$ of a Gaussian is

$$M(\mu) = \exp \left(\mu \bar{\rho} + \frac{\mu^2}{2} \langle \Delta \rho^2 \rangle \right) . \quad (30.174)$$

The probability distribution $P(\rho)$ that, when integrated in accordance with the definition (30.158), yields the Gaussian moment-generating function (30.174), is

$$P(\rho) d\rho = \frac{1}{\sqrt{2\pi\langle\Delta\rho^2\rangle}} \exp\left[-\frac{(\rho - \bar{\rho})^2}{2\langle\Delta\rho^2\rangle}\right] d\rho . \quad (30.175)$$

The 1-point Gaussian probability distribution (30.175) generalizes to the N -point Gaussian probability distribution (30.157).

Non-equilibrium processes in the FLRW background

The subject of cosmological perturbations will be resumed in the next Chapter 32. The present Chapter is concerned principally with an essential ingredient in the calculation of the power spectrum of CMB fluctuations, namely recombination in the unperturbed FLRW background. Recombination presents an opportunity to introduce the collisional Boltzmann equation, §31.5, which allows to follow the evolution of number densities of species out of thermodynamic equilibrium, and which will be invoked again in Chapter 33 to follow the evolution of the photon distribution of the CMB.

In the early Universe, density and temperature were high enough that collisional processes were fast enough to drive particles into mutual thermodynamic equilibrium. But as the Universe expanded, density and temperature decreased to the point that some processes fell out of equilibrium and froze out. Recombination, and its inverse photoionization, constitute one example of such a process. At times well before the epoch of recombination, the two-body process of recombination and its inverse process photoionization drove the ionization state of the gas into thermodynamic equilibrium. But as recombination approached, recombination rates could no longer keep up, slightly delaying the epoch of recombination, and leaving a residual level of ionization. The residual ionization later catalyzed the formation of molecular hydrogen, leading to the first generation of stars.

Besides recombination, there are some other processes of freeze-out in the expanding Universe that are associated with well-understood physics. (1) The weak interactions froze out after electron-positron annihilation, so that protons and neutrons could no longer interconvert, causing the neutron-to-proton ratio to freeze out. The frozen neutron-to-proton ratio subsequently determined the primordial abundance of helium to hydrogen. (2) Nuclear reactions froze out, causing primordial nucleosynthesis to cease at the light elements H, D ($\equiv {}^2\text{H}$), ${}^3\text{He}$, ${}^4\text{He}$, and Li, rather than proceeding all the way to the most tightly bound nucleus, iron. This is well and good, since if nucleosynthesis had proceeded to completion, there would be no stars, and no people.

Yet other processes of freeze-out probably occurred, but their physics is poorly understood, so only guesses and estimates can be made. (1) Our Universe shows an excess of matter (protons, neutrons, electrons) over antimatter (antiprotons, antineutrons, positrons). For this asymmetry to occur, there must have been some T -violating process that preferred the creation of matter over antimatter, and that process must have frozen out. (2) A leading candidate for the non-baryonic cold dark matter is a weakly-interacting massive particle

(WIMP). In order that the mass density of WIMPs be as observed today, their number density must be much less than that of relativistic particles (photons). If WIMPs were initially in thermodynamic equilibrium at some relativistic temperature, then the WIMPs must have annihilated with their antiparticles as they became nonrelativistic; moreover that annihilation must have frozen-out so as to leave the remnant density observed today. To achieve this outcome, the WIMP annihilation cross-section must be comparable to a weak-interaction cross-section, which explains the popularity of the WIMP proposal. As of writing (2015), laboratory attempts to detect WIMPs experimentally have led only to upper limits.

31.1 Conditions around the epoch of recombination

Two key quantities around the time of recombination were the photon temperature T and the baryon number density n_b . Because the baryon-to-photon ratio $n_b/n_\gamma \sim 10^{-9}$, equation (10.104), was so small, the photon distribution was essentially unaffected by the baryons. Photons remained in thermodynamic equilibrium at a temperature T that evolved with cosmic scale factor a (normalized to $a_0 = 1$) as

$$T = \frac{T_0}{a} , \quad (31.1)$$

where $T_0 = 2.725$ K is the CMB temperature today. Equation (31.1) held from after electron-positron annihilation at $T \sim 1$ MeV down to the present time. The baryon number density n_b was (again normalized to $a_0 = 1$)

$$n_b = \frac{3\Omega_b H_0^2}{8\pi G m_b a^3} , \quad (31.2)$$

where $m_b = 939$ MeV was the approximate mean mass per baryon.

The electron fraction X_e may be defined to be the ratio of the electron density n_e to the nuclear proton density n_+ , including all protons in all nuclei,

$$X_e \equiv \frac{n_e}{n_+} . \quad (31.3)$$

The definition (31.3) is chosen so that $X_e = 1$ when the plasma is fully ionized. The nuclear proton density n_+ is

$$n_+ = f_+ n_b , \quad (31.4)$$

where $f_+ \equiv n_+/n_b$ is the proton fraction. To a good approximation, baryons comprised H and ^4He nuclei, and $f_+ = 0.875$, Exercise 31.1.

Exercise 31.1 Proton and neutron fractions. Define the proton and neutron fractions f_+ and f_n by the proton- and neutron-to-baryon ratios

$$f_+ \equiv \frac{n_+}{n_b} = 1 - f_n , \quad f_n \equiv \frac{n_n}{n_b} . \quad (31.5)$$

Here n_+ and n_n are the number densities of protons and neutrons in all nuclei. The baryon number density is their sum $n_b = n_+ + n_n$. For a H plus ${}^4\text{He}$ composition, the nuclear proton and neutron number densities are

$$n_+ = n_{\text{H}} + 2n_{{}^4\text{He}} , \quad n_n = 2n_{{}^4\text{He}} . \quad (31.6)$$

Show that the primordial ${}^4\text{He}$ mass fraction defined by $Y_{{}^4\text{He}} \equiv \rho_{{}^4\text{He}}/(\rho_{\text{H}} + \rho_{{}^4\text{He}})$ satisfies

$$Y_{{}^4\text{He}} = 2f_n . \quad (31.7)$$

The observed primordial ${}^4\text{He}$ abundance is $Y_{{}^4\text{He}} = 0.245 \pm 0.004$ (Cyburt et al., 2016), implying

$$f_n = 0.1225 , \quad f_+ = 1 - f_n = 0.8775 . \quad (31.8)$$

31.2 Overview of recombination

The classic paper on cosmological recombination is Peebles (1968).

The ionization state of the Universe around the time of recombination was determined largely by hydrogen, the most abundant element. Recombination of hydrogen is a two-body process whose inverse is photoionization,



Helium, the next most abundant element, was largely neutral by the time of recombination; its effect on recombination was quite small.

At times well before recombination, the ionization state of the baryonic gas was close to thermodynamic equilibrium. At the temperatures of relevance, electrons and nuclei were nonrelativistic, and their occupation numbers f , given in thermodynamic equilibrium by equations (10.125), were much less than 1. The occupation numbers were small in part because the asymmetry between matter (protons, neutrons, electrons) and antimatter (antiprotons, antineutrons, positrons) is quite small, about 10^{-9} baryons per CMB photon, equation (10.104). Early in the Universe when the temperature exceeded their rest-mass energy, particles and antiparticles in thermodynamic equilibrium had number densities comparable to photons (with a factor of $\frac{3}{4}$ in the number density of fermions relative to bosons, equation (10.141)). Because of the small matter-antimatter asymmetry, the number density of particles and antiparticles were almost equal, so their chemical potentials were almost zero, Exercise 10.18. Relativistic fermions in thermodynamic equilibrium had occupation numbers of order unity for energies less than of order the temperature, $f = 1/(e^{E/T} + 1) \sim 1$ for $E \lesssim T$. As matter particles annihilated with their antiparticles, their occupation number fell to $\sim 10^{-9}$, Figure 10.17. As the Universe continued to expand, the occupation number of the now nonrelativistic particles, still in thermodynamic equilibrium with photons, fell further as $f \sim nT^{-3/2} \propto T^{3/2}$, equation (31.15).

Thus the occupation numbers of nonrelativistic electrons and nuclei was

$$f \sim 10^{-9} \left(\frac{T}{m} \right)^{3/2} \ll 1 \quad (31.10)$$

for particle kinetic energies $p^2/(2m)$ less than of order the temperature T .

Because of the low occupation number, hydrogen remained ionized down to a much lower temperature, $T \sim 0.3 \text{ eV} \sim 3,000 \text{ K}$, than the ionization energy 13.6 eV of hydrogen.

The temperature $T \sim 0.3 \text{ eV}$ of recombination was much lower than the difference $E_1 - E_2 \sim 10.2 \text{ eV}$ between the ground $n = 1$ and first excited $n = 2$ energy levels of hydrogen. Consequently the Boltzmann factor strongly favoured the ground state, so that near recombination almost all the hydrogen atoms were in their ground states, equation (31.20). The recombination temperature $T \sim 0.3 \text{ eV}$ was also significantly lower than the difference $E_2 - E_3 \sim 1.9 \text{ eV}$ between first $n = 2$ and second $n = 3$ excited energy levels of hydrogen, so the population of $n = 2$ substantially outnumbered higher excited states, equation (31.20). To a good approximation, recombination involved only the first two energy levels $n = 1$ and 2 of hydrogen.

As the density and temperature decreased because of adiabatic expansion, recombination could no longer keep up. The large density of hydrogen atoms in the ground state meant that Lyman transitions, transitions between the ground state and other states, were optically thick. Any radiative decay to the ground state produced a Lyman line or continuum photon that was quickly absorbed by a nearby hydrogen atom. Recombination to the ground state was inhibited. The bottleneck caused the $n = 2$ energy level to become overpopulated relative to the ground state, compared to thermodynamic equilibrium.

Recombination nevertheless proceeded via two slow processes, one from the $2p$ level, the other from the $2s$ level of hydrogen. The first process is that, as the Universe expands, the Lyman α $2p - 1s$ transition redshifts, and there is a finite probability for the photon to redshift out of the line without being absorbed. The second process is that the $2s$ level can decay by a forbidden 2-photon transition. A possible third process, collisional deexcitation of excited levels to the ground state, was slower than either of the first two.

31.3 Energy levels and ionization state in thermodynamic equilibrium

Electrons and nuclei near recombination were nonrelativistic, and their occupation numbers were small, and therefore well described by Boltzmann statistics, with occupation number f given by equation (10.126).

31.3.1 Number density of nonrelativistic Boltzmann species in thermodynamic equilibrium

The energy E of a nonrelativistic particle of mass m is related to its momentum p by $E = m + p^2/(2m)$. For a hydrogen atom in energy level n , the rest mass m is less than the rest mass m_p of a proton by the binding energy E_n of the atom, $m = m_p - E_n$. In thermodynamic equilibrium, the number density n of a

nonrelativistic Boltzmann species is

$$n = \int f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = e^{(\mu-m)/T} \int e^{-p^2/(2mT)} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} . \quad (31.11)$$

The integral on the right hand side of equation (31.11) is

$$\int e^{-p^2/(2mT)} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = g \left(\frac{mT}{2\pi\hbar^2} \right)^{3/2} , \quad (31.12)$$

so the number density in thermodynamic equilibrium is

$$n = g \left(\frac{mT}{2\pi\hbar^2} \right)^{3/2} e^{(\mu-m)/T} . \quad (31.13)$$

The factor $(mT/(2\pi\hbar^2))^{3/2}$ defines a length scale λ_T which is a characteristic thermal Compton wavelength of the particles,

$$\lambda_T \equiv \left(\frac{mT}{2\pi\hbar^2} \right)^{-1/2} . \quad (31.14)$$

In terms of their number density n , the occupation number $f = e^{(\mu-E)/T}$ of a Boltzmann species is

$$f = \frac{n}{g} \left(\frac{mT}{2\pi\hbar^2} \right)^{-3/2} e^{-p^2/(2mT)} = \frac{n\lambda_T^3}{g} e^{-p^2/(2mT)} . \quad (31.15)$$

The condition for the validity of the Boltzmann approximation of small occupation numbers is that there be few particles per Compton volume, $n\lambda_T^3 \ll 1$.

31.3.2 Level populations of hydrogen in thermodynamic equilibrium

Bound eigenstates of hydrogen are characterized by quantum numbers n , l , and m associated with their energy, total angular momentum, and projection of the angular momentum along an arbitrary direction. Ignoring the small corrections to energy levels arising from relativistic and spin effects, the energies of the bound eigenstates of hydrogen are

$$-E_n = -13.6 \text{ eV} / n^2 , \quad (31.16)$$

with $n = 1, \dots, \infty$ an integer running from the ground state 1 to the continuum ∞ . Within each energy level n , the total angular momentum l runs over n integers $l = 0, \dots, n-1$. Within each angular momentum level l the “magnetic” quantum number m runs over $2l+1$ integers $m = -l, \dots, l$. Altogether, each hydrogenic energy level n contains $4n^2$ individual states, comprising 2 spin states of the nuclear proton, 2 spin states of the electron, and $\sum_{l=0}^{n-1} (2l+1) = n^2$ states of orbital angular momentum.

In thermodynamic equilibrium, the number density n_{nl} in level nl of hydrogen relative to the number density n_{1s} in the ground level $1s$ is, from equation (31.13),

$$\frac{n_{nl}}{n_{1s}} = (2l+1) e^{(E_n - E_1)/T} . \quad (31.17)$$

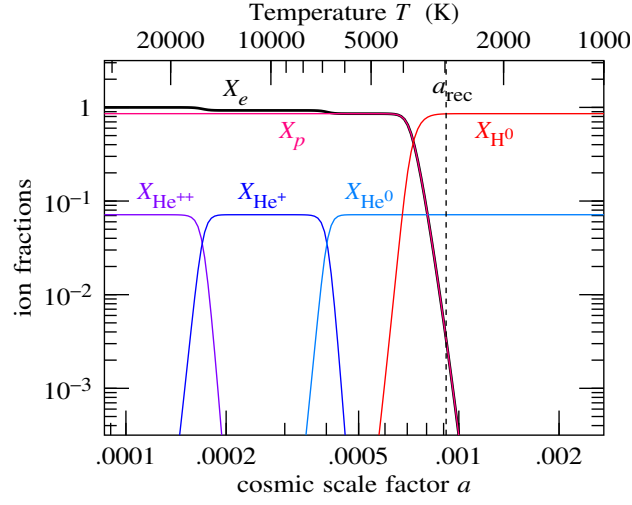


Figure 31.1 Hydrogen and helium ion fractions in thermodynamic equilibrium as a function of cosmic scale factor a scaled to $a_0 = 1$. The total hydrogen and helium fractions are $X_H = 1 - f_n/f_+ = 0.86$ and $X_{He} = \frac{1}{2}f_n/f_+ = 0.07$ where $f_p \equiv 1 - f_n = 0.875$ and $f_n \equiv n_n/n_b = 0.125$ are the neutron- and proton-to-baryon ratios, Exercise 31.1. The dashed vertical line indicates where recombination actually occurs (where the Thomson scattering optical depth is unity), somewhat later than predicted by equilibrium.

31.3.3 Ionization state in thermodynamic equilibrium

In thermodynamic equilibrium, the chemical potentials of protons, electrons, and neutral hydrogen atoms are related by $\mu_p + \mu_e = \mu_H$, equation (10.128). Inserting this equilibrium condition into equation (31.13), valid for nonrelativistic Boltzmann species, implies the relation between the number densities n_p , n_e , and n_{nl} of protons, electrons, and hydrogen atoms in level nl ,

$$\frac{n_p n_e}{n_{nl}} = \frac{g_p g_e}{g_{nl}} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{3/2} e^{-E_n/T}. \quad (31.18)$$

Equation (31.18) is the **Saha equation** for hydrogen. The m_e on the right hand side of equation (31.18) is strictly $m_p m_e / m_{nl}$ where m_{nl} is the mass of the hydrogen atom in level nl , but $m_p \approx m_{nl}$ to a good approximation.

More generally, the Saha equation relating the number densities of an ion X to the next-ionized ion X^+ is

$$\frac{n_{X^+} n_e}{n_X} = \frac{g_{X^+} g_e}{g_X} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{3/2} e^{-E_X/T}. \quad (31.19)$$

Figure 31.1 illustrates the ionization fractions of H and ^4He in thermodynamic equilibrium at the photon temperature T and baryon density n_b given by equations (31.1) and (31.2).

Exercise 31.2 Level populations of hydrogen near recombination. Use the approximation of thermodynamic equilibrium to estimate the relative number densities of states of hydrogen near recombination, where $T \sim 0.3 \text{ eV}$.

Solution. From equation (31.17), the ratio of excited $n = 2$ to ground $n = 1$ levels in thermodynamic equilibrium is

$$n_{2p} = 3 n_{2s} \sim 3 e^{\left(\frac{1}{4}-1\right)13.6 \text{ eV}/0.3 \text{ eV}} n_{1s} \sim 10^{-14} n_{1s} , \quad (31.20)$$

which is tiny. The equilibrium ratio depends steeply on the temperature, which is one reason why recombination cannot keep up as the temperature falls. Similarly, the ratio of the population of the second $n = 3$ to first $n = 2$ excited states is

$$n_3 \sim \frac{9}{4} e^{\left(\frac{1}{9}-\frac{1}{4}\right)13.6 \text{ eV}/0.3 \text{ eV}} n_2 \sim 4 \times 10^{-3} n_2 , \quad (31.21)$$

which is also small. Thus the ground state $n = 1$ dominates the level population, followed by the first excited states $n = 2$,

$$n_1 \gg n_2 \gg n_{n \geq 3} . \quad (31.22)$$

Exercise 31.3 Ionization state of hydrogen near recombination. Use the approximation of thermodynamic equilibrium to estimate the temperature at which hydrogen recombines.

Solution. Almost all the hydrogen atoms are in their ground states. In the approximation that all hydrogen atoms are in their ground state $1s$, the Saha equation (31.18) implies

$$\frac{n_p n_e}{n_{1s}} = \left(\frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_1/T} , \quad (31.23)$$

the statistical weight factor cancelling, $g_{1s} = g_p g_e = 4$. In the approximation of a pure hydrogen gas, in which case the nuclear proton density equals the baryon density, $n_+ = n_b$, the Saha equation (31.23) is

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_+} \left(\frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_1/T} = \frac{2^{3/2} G m_b T_0^3}{3\pi^{1/2} \Omega_b H_0^2} \left(\frac{m_e}{T} \right)^{3/2} e^{-E_1/T} , \quad (31.24)$$

where X_e is the electron fraction, equation (31.3), and n_b the baryon density, equation (31.2). Recombination occurs at $X_e \approx \frac{1}{2}$. Equation (31.24) is then an implicit equation for the temperature T . It can be solved iteratively by guessing an initial T , and calculating an improved value from

$$\frac{E_1}{T} = \ln \left[\frac{2^{3/2} G m_b T_0^3}{3\pi^{1/2} \Omega_b H_0^2} \left(\frac{m_e}{T} \right)^{3/2} \right] . \quad (31.25)$$

Guessing $T = 10^4 \text{ K}$ yields

$$\frac{E_1}{T} \approx 40 , \quad (31.26)$$

which gives the estimated recombination temperature of $T \approx 4,000 \text{ K}$. Iterating a second time gives

$$T \approx 3,800 \text{ K} . \quad (31.27)$$

Concept question 31.4 Atomic structure notation. An eigenstate with $l = 0$ is denoted s , while one with $l = 1$ is denoted p . Why? **Answer.** For historical reasons. In atomic spectroscopy, angular momentum levels $l = 0, 1, 2, 3, 4, \dots$ are conventionally denoted $s, p, d, f, g, \dots$, the first 4 letters standing for sharp, principal, diffuse, and fundamental. After fundamental f , the labelling is alphabetical.

31.4 Occupation numbers

Occupation number was discussed previously in §10.26.

Each species of energy-momentum is described by a dimensionless occupation number, or phase-space probability distribution, a function $f(t, \mathbf{x}, \mathbf{p})$ of time t , comoving position $\mathbf{x}$, and tetrad-frame momentum $\mathbf{p}$, which describes the number dN of particles in a tetrad-frame element $d^3r d^3p / (2\pi\hbar)^3$ of phase-space,

$$dN(t, \mathbf{x}, \mathbf{p}) = f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3}, \quad (31.28)$$

with g being the number of spin states of the particle. The tetrad-frame phase-space element $d^3r d^3p / (2\pi\hbar)^3$ is dimensionless and Lorentz-invariant, and the occupation number f is likewise dimensionless and Lorentz-invariant. The tetrad-frame energy-momentum 4-vector p^m of a particle is

$$p^m \equiv e^m{}_\mu \frac{dx^\mu}{d\lambda} = \{E, \mathbf{p}\} = \{E, p^a\}, \quad (31.29)$$

where λ is the affine parameter, related to proper time τ along the worldline of the particle by $d\lambda \equiv d\tau/m$, which remains well-defined in the limit of massless particles, $m = 0$. The tetrad-frame energy E and momentum $p \equiv |\mathbf{p}|$ for a particle of rest mass m are related by

$$E^2 - p^2 = m^2. \quad (31.30)$$

31.5 Boltzmann equation

The detailed evolution of the abundance of any species can be followed using the **Boltzmann equation**. The Boltzmann equation splits the evolution of the occupation number f of a species into a collisionless part in which each particle evolves as a test particle in the background geometry, and a collisional part in which particles are destroyed or created as a result of collisions with other particles.

Collisionless evolution is described by the single-particle distribution function, the occupation number f . Because phase-space volume is conserved as the system evolves, §4.22.1, conservation of particle number along the paths of particles, $dN/d\lambda = 0$, is equivalent to conservation of the occupation number f defined by equation (31.28),

$$\frac{df}{d\lambda} = 0. \quad (31.31)$$

Equation (31.31) is the **collisionless Boltzmann equation**. The derivative with respect to affine parameter

λ on the left hand side of the Boltzmann equation (31.31) is a Lagrangian derivative along the (timelike or lightlike) worldline of a particle in the fluid.

The collisionless Boltzmann equation holds without modification for particles that do not collide, such as neutrinos or non-baryonic dark matter particles, but it fails for particles whose trajectories are substantially modified by collisions with other particles, such as photons or baryons. Collisions are both a sink and a source of particles, destroying particles of momentum $\mathbf{p}$ and creating others of momentum $\mathbf{p}'$ in the single-particle distribution f . The effect of collisions is modelled by introducing a **collision term**, schematically written $C[f]$, containing both sinks and sources,

$$\frac{df}{d\lambda} = C[f] . \quad (31.32)$$

Equation (31.32) is the **collisional Boltzmann equation**. Since f is dimensionless while the affine parameter $d\lambda \equiv d\tau/m$ has units of time/mass, the units of the collision term $C[f]$ are mass/time.

31.5.1 Boltzmann equation in the FLRW geometry

In the FLRW geometry, homogeneity and isotropy imply that the occupation number is a function $f(t, p)$ only of cosmic time t and of the magnitude p of the proper momentum. The collisional Boltzmann equation (31.32) is then

$$\frac{df}{d\lambda} = \frac{dt}{d\lambda} \frac{\partial f}{\partial t} + \frac{dp}{d\lambda} \frac{\partial f}{\partial p} = C[f] . \quad (31.33)$$

To follow lots of particles simultaneously, switch the integration variable from the affine parameter λ , which is particle-dependent, to cosmic time t , which is the same for all. With cosmic time t as the integration variable, the only non-vanishing vierbein coefficient that depends on t in the background FLRW geometry is $e_0^t = 1$. The relation between cosmic time t and affine parameter λ is

$$\frac{dt}{d\lambda} = p^t = e_0^t p^0 = E , \quad (31.34)$$

where $E = p^0$ is the proper energy of the particle in the tetrad rest-frame. It would be equally possible to use conformal time η as the integration variable, as will be done later in §33.2, in which case $e_0^\eta = 1/a$ and $d\eta/d\lambda = E/a$; for the present purpose however, cosmic time t is slightly more convenient. As found in Exercise 10.6, the proper momentum of a particle, massless or massive, redshifts as $p \propto 1/a$, so $d \ln p / dt = -d \ln a / dt$. Thus the Boltzmann equation (31.33) is

$$\frac{df}{dt} = \frac{\partial f}{\partial t} - \frac{d \ln a}{dt} \frac{\partial f}{\partial \ln p} = \frac{1}{E} C[f] . \quad (31.35)$$

The proper number density n is an integral (10.121) of the occupation number f over momenta. Integrating the left hand side of the Boltzmann equation (31.35) gives

$$\begin{aligned} \int \frac{df}{dt} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} &= \int \frac{\partial f}{\partial t} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} - \frac{d \ln a}{dt} \int \frac{\partial f}{\partial \ln p} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} \\ &= \frac{\partial}{\partial t} \int f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} - \frac{d \ln a}{dt} \left\{ \left[f \frac{g 4\pi p^3}{(2\pi\hbar)^3} \right] - \int 3f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} \right\} \\ &= \frac{dn}{dt} + 3 \frac{d \ln a}{dt} n = \frac{1}{a^3} \frac{dna^3}{dt} . \end{aligned} \quad (31.36)$$

Integrated over momenta, the collisional Boltzmann equation (31.35) is thus

$$\frac{1}{a^3} \frac{dna^3}{dt} = \int C[f] \frac{g 4\pi p^2 dp}{E(2\pi\hbar)^3} . \quad (31.37)$$

Equation (31.37) holds for both massive and massless particles. In the absence of collisions, $C[f] = 0$, the integrated Boltzmann equation (31.37) shows that proper number density n decreases as a^{-3} ,

$$n \propto a^{-3} . \quad (31.38)$$

Equation (31.38) says that the number na^3 of particles in a comoving volume remains constant in the absence of collisions that destroy or create particles.

31.6 Collisions

For a 2-body collision of the form

$$1 + 2 \leftrightarrow 3 + 4 , \quad (31.39)$$

the rate per unit time and volume at which particles of type 1 leave and enter an interval d^3p_1 of momentum space is, in units $c = \hbar = 1$,

$$\begin{aligned} C[f_1] \frac{g_1 d^3p_1}{E_1(2\pi)^3} &= \int \langle |\mathcal{M}|^2 \rangle [-f_1 f_2 (1 \mp f_3)(1 \mp f_4) + f_3 f_4 (1 \mp f_1)(1 \mp f_2)] \\ &\quad (2\pi)^4 \delta_D^4(p_1 + p_2 - p_3 - p_4) \frac{g_1 d^3p_1}{2E_1(2\pi)^3} \frac{g_2 d^3p_2}{2E_2(2\pi)^3} \frac{d^3p_3}{2E_3(2\pi)^3} \frac{d^3p_4}{2E_4(2\pi)^3} . \end{aligned} \quad (31.40)$$

All factors in equation (31.40) are Lorentz scalars. On the left hand side, the collision term $C[f_1]$ and the momentum 3-volume element d^3p_1/E_1 are both Lorentz scalars. On the right hand side, the mean amplitude squared $\langle |\mathcal{M}|^2 \rangle$, the various occupation numbers f_i , the energy-momentum conserving 4-dimensional Dirac delta function $\delta_D^4(p_1 + p_2 - p_3 - p_4)$, and each of the four momentum 3-volume elements $d^3p_i/(2E_i)$, are all Lorentz scalars. The factor of $1/2$ in each momentum element $d^3p_i/(2E_i)$ has its roots in quantum field theory, where it serves to normalize propagators of quanta correctly, equation (42.102).

The first ingredient in the integrand on the right hand side of the expression (31.40) is the Lorentz-invariant

mean scattering amplitude squared $\langle |\mathcal{M}|^2 \rangle$, calculated using quantum field theory, §47.4. By convention (see for example equation (47.48) in §47.7), the mean amplitude squared $\langle |\mathcal{M}|^2 \rangle$ represents a rate averaged over initial spin states and summed over final spin states,

$$\langle |\mathcal{M}|^2 \rangle \equiv \frac{1}{g_1 g_2} \sum_{\text{spins } i} |\mathcal{M}|_i^2, \quad (31.41)$$

so that the mean amplitude squared represents a rate per incoming spin state. To convert the mean amplitude squared to a net rate per unit time and volume, it is necessary to sum over particles in the initial states, which explains why equation (31.40) includes spin factors g_1 and g_2 in the integral over initial momenta. The average-over-incoming spins factor $1/(g_1 g_2)$ in the mean amplitude squared cancels the sum-over-incoming spins factor $g_1 g_2$ in the integral (31.40). The convention to average over initial states when in the end they must be summed over may seem strange, but then so are many conventions. For a process involving 4 particles such as (31.39), the mean amplitude squared $\langle |\mathcal{M}|^2 \rangle$ is dimensionless (in units $c = \hbar = 1$), but it is not dimensionless in general, equation (47.17).

The second ingredient in the integrand on the right hand side of expression (31.40) is the combination of rate factors

$$\text{rate}(1 + 2 \rightarrow 3 + 4) \propto f_1 f_2 (1 \mp f_3)(1 \mp f_4), \quad (31.42a)$$

$$\text{rate}(1 + 2 \leftarrow 3 + 4) \propto f_3 f_4 (1 \mp f_1)(1 \mp f_2), \quad (31.42b)$$

where the $1 \mp f$ factors are blocking or stimulation factors, the choice of $\mp$ sign depending on whether the species in question is fermionic or bosonic:

$$1 - f = \text{Fermi-Dirac blocking factor}, \quad (31.43a)$$

$$1 + f = \text{Bose-Einstein stimulation factor}. \quad (31.43b)$$

The first rate factor (31.42a) expresses the fact that the rate to lose particles from $1 + 2 \rightarrow 3 + 4$ collisions is proportional to the occupancy $f_1 f_2$ of the initial states, modulated by the blocking/stimulation factors $(1 \mp f_3)(1 \mp f_4)$ of the final states. Likewise the second rate factor (31.42b) expresses the fact that the rate to gain particles from $1 + 2 \leftarrow 3 + 4$ collisions is proportional to the occupancy $f_3 f_4$ of the initial states, modulated by the blocking/stimulation factors $(1 \mp f_1)(1 \mp f_2)$ of the final states. In thermodynamic equilibrium, the rates (31.42) balance, Exercise 31.5, a property that is called detailed balance, or microscopic reversibility. Microscopic reversibility is a consequence of time reversal symmetry.

The final ingredient in the integrand on the right hand side of expression (31.40) is the 4-dimensional Dirac delta function, which imposes energy-momentum conservation on the process $1 + 2 \leftrightarrow 3 + 4$. The 4-dimensional delta function is a product of a 1-dimensional delta function expressing energy conservation, and a 3-dimensional delta function expressing momentum conservation:

$$(2\pi)^4 \delta_D^4(p_1 + p_2 - p_3 - p_4) = 2\pi \delta_D(E_1 + E_2 - E_3 - E_4) (2\pi)^3 \delta_D^3(\mathbf{p}_1 + \mathbf{p}_2 - \mathbf{p}_3 - \mathbf{p}_4). \quad (31.44)$$

Exercise 31.5 Detailed balance.

1. Show that the rates balance in thermodynamic equilibrium,

$$f_1 f_2 (1 \mp f_3)(1 \mp f_4) = f_3 f_4 (1 \mp f_1)(1 \mp f_2) . \quad (31.45)$$

2. Conclude that, if each particle type i has a thermodynamic distribution with its own temperature T_i and chemical potential μ_i , then

$$\begin{aligned} & - f_1 f_2 (1 \mp f_3)(1 \mp f_4) + f_3 f_4 (1 \mp f_1)(1 \mp f_2) \\ & = f_1 f_2 (1 \mp f_3)(1 \mp f_4) \left[-1 + \exp \left(\frac{E_1 - \mu_1}{T_1} + \frac{E_2 - \mu_2}{T_2} + \frac{-E_3 + \mu_3}{T_3} + \frac{-E_4 + \mu_4}{T_4} \right) \right] . \end{aligned} \quad (31.46)$$

Solution.

1. Equation (31.45) is true if and only if

$$\frac{f_1}{1 \mp f_1} \frac{f_2}{1 \mp f_2} = \frac{f_3}{1 \mp f_3} \frac{f_4}{1 \mp f_4} . \quad (31.47)$$

But

$$\frac{f}{1 \mp f} = e^{(-E+\mu)/T} , \quad (31.48)$$

so (31.47) is true if and only if

$$\frac{-E_1 + \mu_1}{T} + \frac{-E_2 + \mu_2}{T} = \frac{-E_3 + \mu_3}{T} + \frac{-E_4 + \mu_4}{T} , \quad (31.49)$$

which is true in thermodynamic equilibrium because

$$E_1 + E_2 = E_3 + E_4 , \quad \mu_1 + \mu_2 = \mu_3 + \mu_4 . \quad (31.50)$$

31.7 Non-equilibrium recombination

At times well before recombination, the ionization state of the baryonic gas was well described by thermodynamic equilibrium. However, as recombination approached, the recombination rate could not keep up with the adiabatic decrease in density and temperature. Consequently recombination was delayed slightly compared to what would be expected in thermodynamic equilibrium. To model the CMB precisely, it is necessary to worry about the details of non-equilibrium recombination.

Although the ionization state was out of equilibrium, elastic collisions between electrons, ions, and neutrals kept the velocity distributions of electrons and baryons in mutual thermodynamic equilibrium at a common kinetic temperature $T_e = T_b$.

Recombination to and photoionization out of bound state i of hydrogen destroys and creates a free electron. The electron collision integral $C_i[f_e]$ corresponding to this process is given by, from equation (31.40) with

stimulated processes from protons, electrons, and hydrogen atoms neglected because of their small occupation numbers,

$$C_i[f_e] \frac{g_p d^3 p_p}{m_p (2\pi)^3} = \int \langle |\mathcal{M}|^2 \rangle_i [-f_p f_e (1 + f_\gamma) + f_i f_\gamma] (2\pi)^4 \delta_D^4(p_p + p_e - p_i - p_\gamma) \frac{g_p d^3 p_p}{2m_p (2\pi)^3} \frac{g_e d^3 p_e}{2m_e (2\pi)^3} \frac{d^3 p_i}{2m_p (2\pi)^3} \frac{d^3 p_\gamma}{2p_\gamma (2\pi)^3} . \quad (31.51)$$

The $-f_p f_e$ term in the integrand corresponds to direct recombination, the $-f_p f_e f_\gamma$ term to stimulated recombination, and the $f_i f_\gamma$ term to photoionization. Because the proton and hydrogen atom are so massive, they remain essentially at rest during a recombination or photoionization, so the mean squared amplitude $\langle |\mathcal{M}|^2 \rangle_i$ for photoionization out of and recombination into bound state i is essentially independent of the proton and hydrogen momenta. Integrating the collision integral (31.51) over the proton and hydrogen momenta yields

$$C_i[f_e] = \frac{1}{8m_p^2} \int \langle |\mathcal{M}|^2 \rangle_i 2\pi \delta_D(E_p + E_e - E_i - E_\gamma) [-n_p f_e (1 + f_\gamma) + n_i (g_p/g_i) f_\gamma] \frac{d^3 p_\gamma}{2p_\gamma (2\pi)^3} , \quad (31.52)$$

one of the integrations over momenta being swallowed by the momentum-conserving Dirac delta function $(2\pi)^3 \delta_D^3(\mathbf{p}_p + \mathbf{p}_e - \mathbf{p}_i - \mathbf{p}_\gamma)$. Again because the proton and hydrogen atom are so massive, the photon is emitted and absorbed isotropically. Integrating over directions $\hat{\mathbf{p}}_\gamma$ of the photon momentum yields 4π . Integrating over the photon energy p_γ swallows the energy-conserving delta function, yielding

$$C_i[f_e] = \frac{p_\gamma}{16\pi m_p^2} \langle |\mathcal{M}|^2 \rangle_i [-n_p f_e (1 + f_\gamma) + n_i (g_p/g_i) f_\gamma] . \quad (31.53)$$

If the hydrogenic state i is in energy level n , then energy conservation requires that the energy $E_\gamma \equiv p_\gamma$ of the photon be the sum of the electron kinetic energy and the binding energy (ionization energy) of the level,

$$\frac{p_e^2}{2m_e} + E_n = p_\gamma . \quad (31.54)$$

In the situation of cosmological recombination under consideration, the photons, whose numbers overwhelm those of electrons, have a thermal (Planckian) momentum distribution at temperature T_γ . Elastic collisions between electrons keep their distribution close to thermal (Maxwellian). Since electron energies redshift faster than photon energies, $p_e^2/(2m) \propto a^{-2}$ versus $p_\gamma \propto a^{-1}$, the electron temperature is slightly below that of photons. However, electron-photon collisions keep the electron temperature closely equal to the photon temperature, $T_e = T_\gamma$, up to and through recombination. After recombination, electron-photon collisions become rare enough that the electron kinetic temperature drops below the photon temperature (Scott and Moss, 2009). For completeness, the treatment in this section allows different electron and photon temperatures, although the two temperatures will be set equal in subsequent sections.

Substituting the Boltzmann distribution (31.15) at temperature T_e for the electron occupation number f_e , and the Planckian distribution (10.130) at temperature T_γ for the photon occupation number f_γ , brings the

electron collision integral (31.53) to

$$C_i[f_e] = \frac{p_\gamma}{16\pi m_p^2} \langle |\mathcal{M}|^2 \rangle_i \left[\frac{-n_p n_e (1/g_e) (m_e T_e / 2\pi)^{-3/2} e^{-p_e^2 / (2m_e T_e)} + n_i (g_p / g_i) e^{-p_\gamma / T_\gamma}}{1 - e^{-p_\gamma / T_\gamma}} \right]. \quad (31.55)$$

Finally, integrating the collision integral (31.55) over electron momenta gives

$$\int C_i[f_e] \frac{g_e d^3 p_e}{m_p (2\pi)^3} = -n_p n_e [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + n_i \beta_i(T_\gamma), \quad (31.56)$$

where $\alpha_i(T_e)$ and $\alpha_i^{\text{stim}}(T_e, T_\gamma)$ are thermally averaged direct and stimulated recombination rate coefficients to state i , and $\beta_i(T_\gamma)$ is the photoionization rate coefficient out of bound state i . The direct recombination rate $\alpha_i(T_e)$ depends only on the electron temperature T_e , while the photoionization rate $\beta_i(T_\gamma)$ depends only on the photon temperature T_γ . The stimulated recombination rate $\alpha_i^{\text{stim}}(T_e, T_\gamma)$ depends on both temperatures. In cosmological recombination, stimulated recombination is a small correction of order $e^{-E_n/T}$, which can be neglected. If stimulated recombination is neglected, then detailed balance imposes

$$\beta_i(T) = \alpha_i(T) \left(\frac{n_p n_e}{n_i} \right)_{\text{TE}} = \alpha_i(T) \frac{g_p g_e}{g_i} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{3/2} e^{-E_n/T}. \quad (31.57)$$

The Boltzmann equation for electrons, equation (31.37), is a sum over recombinations to and photoionizations out of bound states i ,

$$\frac{1}{a^3} \frac{dn_e a^3}{dt} = -n_p n_e \sum_i [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + \sum_i n_i \beta_i(T_\gamma). \quad (31.58)$$

Let X_i denote the ratio of the number density of species i to the nuclear proton density n_+ ,

$$X_i \equiv \frac{n_i}{n_+}. \quad (31.59)$$

The ratio is defined so that the electron fraction is unity, $X_e = 1$, when the plasma is fully ionized. Since $n_+ a^3$ is constant as the Universe expands, the Boltzmann equation (31.58) can be written as an equation for the evolution of the electron fraction,

$$\frac{dX_e}{dt} = -X_p X_e n_+ \sum_i [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + \sum_i X_i \beta_i(T_\gamma). \quad (31.60)$$

Equation (31.60) gives the rate of change of the electron fraction for a pure hydrogen gas. If other elements are included, notably helium, additional processes of recombination to and ionization out of bound states of those elements should be adjoined.

31.8 Recombination: Peebles approximation

Recombination is dominated by hydrogen, the dominant chemical element. The second most abundant element is helium, which is largely neutral by the time of recombination. Peebles (1968) argued that the overall

hydrogen density and the predominance of hydrogen atoms in the ground state, equation (31.22), would have the consequence that the gas would be optically thick to Lyman transitions, that is, to transitions to the ground state, but optically thin in transitions to excited states. Consequently any continuum or line Lyman photon emitted as a result of a recombination or transition to the ground state would be quickly absorbed. On the other hand radiative transitions between excited levels $n \geq 2$ would proceed rapidly without hindrance, leading to a thermal distribution among the excited levels. Since the dominant excited level would be $n = 2$, equation (31.22), Peebles (1968) argued that recombination could be approximated by a 3-level system consisting of protons and of $n = 2$ and $n = 1$ levels of hydrogen.

Since transitions from the continuum to $n = 1$ were ineffective, the rate of change of the proton fraction $X_p \equiv n_p/n_+$ was dominated by recombinations to and photoionizations out of the $n = 2$ level,

$$\frac{dX_p}{dt} = -X_p X_e n_+ \alpha_2 + X_2 \beta_2 . \quad (31.61)$$

Equation (31.61) ignores stimulated recombination, which is a $e^{-E_2/T} \ll 1$ correction to the rate.

Peebles (1968) argued that successful recombination to the $n = 1$ ground state would be dominated by slow leakage out of the $n = 2$ level, which occurred by two processes. The first process is 2-photon decay out of the $2s$ state, which occurs at a rate $A_{2s} = 8.22458 \text{ s}^{-1}$. The second process is that, although most decays out of the $2p$ state produced a Lyman α photon that was immediately absorbed by a nearby hydrogen atom, the expansion of the Universe redshifted the emitted photon, and a small fraction P_S of the emitted Lyman α photons succeeded in redshifting out of the line without being reabsorbed. The fraction P_S of emitted photons that escape in an expanding medium can be approximated using the Sobolev formalism, §31.10. Thus the rate of change of the fraction $X_1 \equiv n_1/n_+$ of hydrogen atoms in the ground $n = 1$ level is

$$\frac{dX_1}{dt} = X_2 A_{21} - X_1 B_{12} , \quad (31.62)$$

where the effective spontaneous decay rate A_{21} from the $n = 2$ levels to the ground $n = 1$ level is

$$A_{21} = \frac{g_{2s}}{g_2} A_{2s-1s} + \frac{g_{2p}}{g_2} A_{2p-1s} P_S , \quad (31.63)$$

with P_S the Sobolev escape probability given by equation (31.101). Equation (31.63) assumes that $2s$ and $2p$ are populated in the ratio $(g_{2s}/g_2) : (g_{2p}/g_2) = \frac{1}{4} : \frac{3}{4}$ of their statistical weights. The value of the spontaneous decay coefficient A_{2p-1s} itself is not actually needed since it cancels in the Sobolev approximation, equation (31.102),

$$\frac{g_{2p}}{g_2} A_{2p-1s} P_S = \frac{1}{X_1 n_+} \frac{g_1}{g_2} \frac{8\pi H}{\lambda_{2p-1s}^3} , \quad (31.64)$$

where H is the Hubble parameter. The statistical weight factor is $g_1/g_2 = \frac{1}{4}$.

Detailed balance requires that dX_p/dt , equation (31.61), must vanish in thermodynamic equilibrium (TE), so the ratio of photoionization to recombination rate coefficients must be

$$\frac{\beta_2}{\alpha_2} = \left(\frac{n_p n_e}{n_2} \right)_{\text{TE}} = \frac{g_p g_e}{g_2} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{3/2} e^{-E_2/T} , \quad (31.65)$$

the statistical weight factor being $g_p g_e / g_2 = \frac{1}{4}$. Hence equation (31.61) may be written

$$\frac{dX_p}{dt} = X_p X_e n_+ \alpha_2 \left(-1 + \frac{1}{b_{p2}} \right), \quad (31.66)$$

where the departure coefficient b_{p2} is the value of $n_p n_e / n_2$ relative to its value in thermodynamic equilibrium,

$$b_{p2} \equiv \frac{n_p n_e}{n_2} \bigg/ \left(\frac{n_p n_e}{n_2} \right)_{\text{TE}} = \frac{X_p X_e n_+}{X_2} \frac{g_2}{g_p g_e} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{-3/2} e^{E_2/T}. \quad (31.67)$$

Similarly, detailed balance requires that dX_1/dt , equation (31.62), must vanish in thermodynamic equilibrium, so the ratio of radiative excitation to decay rate coefficients must be

$$\frac{B_{12}}{A_{21}} = \left(\frac{X_2}{X_1} \right)_{\text{TE}} = \frac{g_2}{g_1} e^{-E_{12}/T}, \quad (31.68)$$

with $E_{12} \equiv E_1 - E_2$ and $g_2/g_1 = 4$. Thus equation (31.62) may be written

$$\frac{dX_1}{dt} = X_1 B_{12} (b_{21} - 1), \quad (31.69)$$

where the departure coefficient b_{21} is the value of n_2/n_1 relative to its value in thermodynamic equilibrium,

$$b_{21} \equiv \frac{n_2}{n_1} \bigg/ \left(\frac{n_2}{n_1} \right)_{\text{TE}} = \frac{X_2 g_1}{X_1 g_2} e^{E_{12}/T}. \quad (31.70)$$

Since the population of the $n = 2$ level was so much smaller than the populations either of protons or of the ground $n = 1$ level of hydrogen, Peebles (1968) argued that the rate of change of X_2 must be negligible relative to the rates of change of X_p and X_1 ,

$$\frac{dX_2}{dt} = -\frac{dX_p}{dt} - \frac{dX_1}{dt} = X_p X_e n_+ \alpha_2 - X_2 \beta_2 - X_2 A_{21} + X_1 B_{12} \approx 0. \quad (31.71)$$

The approximation (31.71) of vanishing dX_2/dt allows X_2 to be eliminated in favour of X_1 ,

$$\frac{X_2}{X_1} = \frac{(X_p X_e / X_1) n_+ \alpha_2 + B_{12}}{\beta_2 + A_{21}}. \quad (31.72)$$

Given the detailed balance relations (31.65) between β_2 and α_2 , and (31.68) between B_{12} and A_{21} , the relation (31.72) may also be written as an expression for the departure coefficient b_{21} ,

$$b_{21} = \frac{b_{p1} \beta_2 + A_{21}}{\beta_2 + A_{21}}, \quad (31.73)$$

in terms of the departure coefficient b_{p1} , the value of $n_p n_e / n_1$ relative to its value in thermodynamic equilibrium,

$$b_{p1} \equiv \frac{n_p n_e}{n_1} \bigg/ \left(\frac{n_p n_e}{n_1} \right)_{\text{TE}} = \frac{X_p X_e n_+}{X_1} \frac{g_1}{g_p g_e} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{-3/2} e^{E_1/T}. \quad (31.74)$$

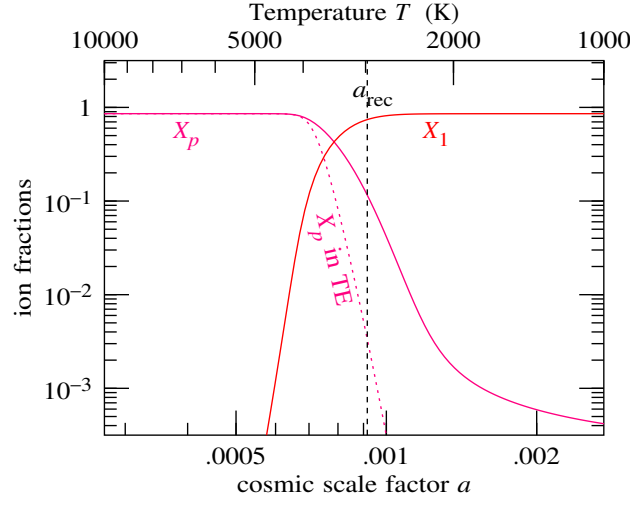


Figure 31.2 Non-equilibrium hydrogen ion fractions as a function of cosmic scale factor a scaled to $a_0 = 1$. The total hydrogen fraction, $X_H = f_H = 0.86$, is the fraction of nuclear protons that are hydrogen nuclei, equation (31.83).

The statistical weight factor is $g_1/(g_p g_e) = 1$. Equation (31.73) allows the departure coefficients b_{21} and $b_{p2} \equiv b_{p1}/b_{21}$ in the recombination equations (31.66) and (31.69) to be eliminated in favour of b_{p1} , yielding

$$\frac{d \ln X_p}{dt} = X_e n_+ \alpha_2 \frac{A_{21}}{\beta_2 + A_{21}} \left(\frac{1}{b_{p1}} - 1 \right), \quad (31.75a)$$

$$\frac{d \ln X_1}{dt} = B_{12} \frac{\beta_2}{\beta_2 + A_{21}} (b_{p1} - 1). \quad (31.75b)$$

Equations (31.75a) and (31.75b) combine to give $d(X_p + X_1)/dt = 0$ in accordance with the condition (31.71), but each of equations (31.75) is written in a form that remains finite as respectively $X_p \rightarrow 0$ and $X_1 \rightarrow 0$.

Equations (31.75) combine to give the rate of change of the logarithmic departure coefficient $\ln b_{p1}$,

$$\frac{d \ln b_{p1}}{dt} = \frac{d \ln X_p}{dt} + \frac{d \ln X_e}{dt} - \frac{d \ln X_1}{dt} + \frac{d}{dt} \ln \left(n_+ T^{-3/2} e^{E_1/T} \right). \quad (31.76)$$

Given that helium is largely neutral by the time of recombination, charge conservation in the pure hydrogen gas implies that

$$X_e = X_p, \quad (31.77)$$

so the time derivative of $\ln X_e$ in equation (31.76) is the same as that for $\ln X_p$. The time derivatives of the temperature and density follow from $T \propto a^{-1}$ and $n_+ \propto a^{-3}$, equations (31.1) and (31.4). The differential equation (31.76) is stiff. Near thermodynamic equilibrium, the logarithmic departure coefficient $\ln b_{p1}$ is near

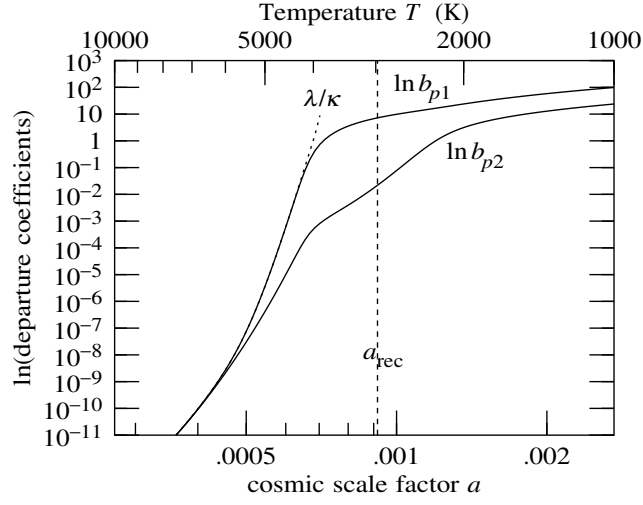


Figure 31.3 Logarithmic ionized-to-bound level departure coefficients $\ln b_{p1}$ and $\ln b_{p2}$, equations (31.74) and (31.67). Logarithmic departure coefficients are zero in thermodynamic equilibrium (the departure coefficients themselves are unity). The increasingly positive values of $\ln b_{p1}$ and $\ln b_{p2}$ mean that protons become over-abundant compared to thermodynamic equilibrium, that is, recombination is not keeping pace with the cosmological decrease in temperature. The $n = 1$ ground level is further from thermodynamic equilibrium with protons than the $n = 2$ and other excited levels. When first coming out of thermodynamic equilibrium, the logarithmic departure coefficient $\ln b_{p1}$ is approximated by its steady state value λ/κ , equation (31.81), indicated by the dotted line.

zero, and then the factors involving b_{p1} on the right hand sides of equations (31.75) become small,

$$\frac{1}{b_{p1}} - 1 = e^{-\ln b_{p1}} - 1 \approx -\ln b_{p1}, \quad b_{p1} - 1 = e^{\ln b_{p1}} - 1 \approx \ln b_{p1}. \quad (31.78)$$

The $d \ln X_i / dt$ derivatives in equation (31.76) are then proportional to $\ln b_{p1}$ with a negative coefficient $-\kappa$,

$$\frac{d \ln X_p}{dt} + \frac{d \ln X_e}{dt} - \frac{d \ln X_1}{dt} \approx -\kappa \ln b_{p1}. \quad (31.79)$$

Thus the differential equation (31.76) takes the form

$$\frac{d \ln b_{p1}}{dt} \approx -\kappa \ln b_{p1} + \lambda, \quad (31.80)$$

where the forcing term λ is the remaining, last, term on the right hand side of the differential equation (31.76). The κ term tends to drive $\ln b_{p1}$ exponentially to zero, that is, into thermodynamic equilibrium, while the forcing term λ drives $\ln b_{p1}$ away from zero. The differential equation (31.80) is stiff when κ is much larger than the absolute value of λ .

A solution to the stiffness problem is to evaluate the thermodynamic equilibrium value of $\kappa/|\lambda|$, and if it

exceeds some threshold (say 10^2), then set $\ln b_{p1}$ to the steady state solution of equation (31.80), which is

$$\ln b_{p1} \approx \frac{\lambda}{\kappa} . \quad (31.81)$$

A differential equation solver that can cope with stiff equations will in effect impose the solution (31.81). Given the logarithmic departure coefficient $\ln b_{p1}$, the neutral X_1 and ionized X_p hydrogen fractions follow, in a form that remains numerically well-behaved even when X_1 or X_p is tiny, as

$$X_1 = \frac{4f_H^2 e^q}{(1 + \sqrt{1 + 4f_H e^q})^2} , \quad X_p = \frac{2f_H}{1 + \sqrt{1 + 4f_H e^q}} , \quad (31.82)$$

where f_H ,

$$f_H \equiv n_H/n_+ = 1 - f_n/f_+ , \quad (31.83)$$

is the fraction of nuclear protons that are hydrogen nuclei, and q is

$$q \equiv \ln \left(\frac{X_1}{X_p X_e} \right) = \ln \left[n_+ \frac{g_1}{g_p g_e} \left(\frac{m_e T}{2\pi \hbar^2} \right)^{-3/2} \right] + \frac{\chi_H}{T} - \ln b_{p1} . \quad (31.84)$$

The statistical weight factor is $g_1/(g_p g_e) = 1$.

Only when $\kappa/|\lambda|$ falls below the threshold is it necessary to start solving the differential equation numerically. It is better to solve directly for the proton fraction X_p rather than the departure coefficient b_{p1} , since as recombination freezes out, X_p changes slowly, whereas b_{p1} continues to evolve, and solving for X_p from b_{p1} becomes numerically unstable. The differential equation governing X_p is, from equation (31.75a),

$$\boxed{\frac{dX_p}{dt} = \left(X_1 \frac{g_2}{g_1} \beta_2 e^{(E_2 - E_1)/T} - X_e X_p n_+ \alpha_2 \right) \frac{A_{21}}{\beta_2 + A_{21}}} . \quad (31.85)$$

The statistical weight factor is $g_2/g_1 = 4$. Figure 31.2 shows the resulting non-equilibrium H ion fractions, and Figure 31.3 shows the logarithmic departure coefficients $\ln b_{p1}$ and $\ln b_{p2}$. Exercise 31.6 asks you to write code to solve equation (31.85).

Exercise 31.6 Recombination. Write code that implements the recombination of hydrogen.

1. Well before recombination, the ionization state is near ionization equilibrium. As suggested in the text, calculate the coefficients κ and λ that go into equation (31.80) in thermodynamic equilibrium. If $\kappa/|\lambda|$ exceeds some threshold, then set the logarithmic departure coefficient to $\ln b_{p1} = \lambda/\kappa$, equation (31.81). Thence deduce the ionization fractions X_1 and X_p , equation (31.82).
2. Once $\kappa/|\lambda|$ falls below the threshold, solve the evolution equation (31.85) for the proton fraction X_p numerically. Approximate the recombination rate α_2 to the $n = 2$ level of hydrogen by formula (31.92a).

Solution. See Figures 31.2 and 31.3.

31.9 Recombination: Seager et al. approximation

Seager, Sasselov, and Scott (1999) provide an improved approximation for recombination based on the Peebles (1968) approximation, but with the inclusion of helium, and with the $n = 2$ recombination coefficients adjusted to fit the results of a detailed calculation of recombination by Seager, Sasselov, and Scott (2000) that includes explicit treatment of up to 300 levels of H, 200 levels of He, and 100 levels of He^+ , plus one each of e , p , H^- , and He^{++} , plus the ground levels of molecular hydrogen species H_2 and H_2^+ .

Seager et al.'s approximation has been refined by Wong, Moss, and Scott (2008) to include the semi-forbidden decay $\text{He } 2p^3P_1 \rightarrow 1s^1S_0$ from the triplet $2p$ state of helium, and the scattering of $\text{He } 2p \rightarrow 1s$ photons by neutral hydrogen. Chluba and Thomas (2011) have developed an even more comprehensive approach to recombination. The various refinements affect the electron fraction X_e at the percent level. The present section follows the simpler work of Seager, Sasselov, and Scott (1999).

Seager, Sasselov, and Scott (1999) adjoin to the hydrogenic recombination equation (31.85) an equivalent equation for helium, protons p being replaced by singly-ionized helium He^+ in its ground state. The effective spontaneous decay $A_{\text{He}21}$ from the singlet $n = 2$ levels to the ground $n = 1$ level of neutral He is, analogous to the hydrogenic equation (31.63),

$$A_{\text{He}21} = \frac{g_{\text{He}2s}}{g_{\text{He}2}} A_{\text{He}2s-1s} + \frac{g_{\text{He}2p}}{g_{\text{He}2}} A_{\text{He}2p-1s} P_S e^{(E_{\text{He}2s} - E_{\text{He}2p})/T} . \quad (31.86)$$

The extra factor of $e^{(E_{\text{He}2s} - E_{\text{He}2p})/T}$ takes into account that the $2p$ state lies slightly but appreciably above the $2s$ state in energy, so its population in thermodynamic equilibrium is reduced by a corresponding Boltzmann factor. The statistical weight factors are $g_{\text{He}2s}/g_{\text{He}2} = \frac{1}{4}$ and $g_{\text{He}2p}/g_{\text{He}2} = \frac{3}{4}$. As in the hydrogenic case, equation (31.64), the value of $A_{\text{He}2p-1s}$ cancels against the Sobolev probability P_S , equation (31.102),

$$\frac{g_{\text{He}2p}}{g_{\text{He}2}} A_{\text{He}2p-1s} P_S = \frac{1}{X_{\text{He}1} n_+} \frac{g_{\text{He}1}}{g_{\text{He}2}} \frac{8\pi H}{\lambda_{\text{He}2p-1s}^3} , \quad (31.87)$$

the statistical weight factor being $g_{\text{He}1}/g_{\text{He}2} = \frac{1}{4}$.

In thermodynamic equilibrium, He^{++} combines to He^+ at a redshift of $z \sim 6,000$, a factor of 6 higher than recombination, Figure 31.1. By the time recombination approaches, little He^{++} remains. He^{++} is well-approximated throughout as being in thermodynamic equilibrium with He^+ .

Charge conservation implies that the electron fraction density X_e is

$$X_e = X_p + X_{\text{He}^+} + 2X_{\text{He}^{++}} . \quad (31.88)$$

The relevant atomic physics is as follows. The wavelengths of the $2 \rightarrow 1$ transitions of hydrogen and helium are

$$\lambda_{\text{H}2p-1s} = 121.5682 \text{ nm} , \quad \lambda_{\text{He}2p-1s} = 58.4334 \text{ nm} , \quad \lambda_{\text{He}2s-1s} = 60.1404 \text{ nm} . \quad (31.89)$$

Ionization energies of hydrogen and helium, commonly quoted in units of cm^{-1} , are

$$\chi_{\text{H}} = 10,967,877.17 \text{ cm}^{-1} , \quad \chi_{\text{He}} = 19,831,066.9 \text{ cm}^{-1} , \quad \chi_{\text{He}^+} = 43,890,887.89 \text{ cm}^{-1} . \quad (31.90)$$

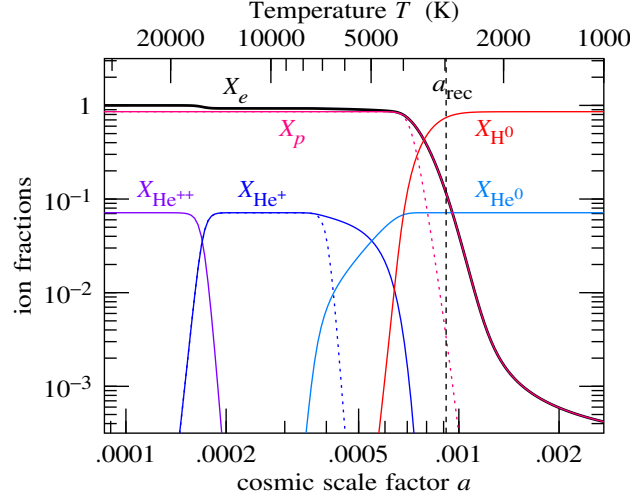


Figure 31.4 Non-equilibrium hydrogen and helium ion fractions as a function of cosmic scale factor a scaled to $a_0 = 1$. Dotted lines show X_p and X_{He^+} in thermodynamic equilibrium. The total hydrogen and helium fractions are $X_{\text{H}} = 1 - f_n/f_+ = 0.86$ and $X_{4\text{He}} = \frac{1}{2}f_n/f_+ = 0.07$.

The spontaneous 2-photon $2s \rightarrow 1s$ transition rates of hydrogen and neutral helium are

$$A_{\text{H}2s-1s} = 8.22458 \text{ s}^{-1}, \quad A_{\text{He}2s-1s} = 51.3 \text{ s}^{-1}. \quad (31.91)$$

The effective recombination rates to $n = 2$ levels of hydrogen and neutral helium are

$$\alpha_{\text{H}2}(T) = 1.14 \times 10^{-19} \frac{4.309 (T/10^4 \text{ K})^{-0.6166}}{1 + 0.6703 (T/10^4 \text{ K})^{0.5300}} \text{ m}^3 \text{ s}^{-1}, \quad (31.92a)$$

$$\alpha_{\text{He}2}(T) = 10^{-16.744} \left[\sqrt{T/3 \text{ K}} \left(1 + \sqrt{T/3 \text{ K}} \right)^{1-p} \left(1 + \sqrt{T/10^{5.114} \text{ K}} \right)^{1+p} \right]^{-1} \text{ m}^3 \text{ s}^{-1}, \quad (31.92b)$$

with $p = 0.711$. The factor of 1.14 in the hydrogenic recombination rate (31.92a) is a fudge factor introduced by Seager, Sasselov, and Scott (1999) that adjusts the Hummer's (1994) calculated rate coefficient to achieve agreement with the multi-level numerical computation of Seager, Sasselov, and Scott (2000). The helium recombination rate (31.92b) is from Hummer and Storey (1998). The statistical weight factor that goes into the ratio $\beta_{\text{He}2}/\alpha_{\text{He}2}$ of photoionization to recombination rates for He, analogous to the hydrogenic ratio (31.65), is $g_{\text{He}^+}g_e/g_{\text{He}2} = 2 \times 2/4 = 1$.

Figure 31.4 shows the recombination of hydrogen and helium in the Seager, Sasselov, and Scott (1999) approximation. The Figure shows that the recombination of singly-ionized helium is, like the recombination of protons, delayed compared to thermodynamic equilibrium. Even so, helium is almost entirely neutral by the time of recombination, so in practice helium has little effect on recombination.

31.10 Sobolev escape probability

The Sobolev escape probability formalism applies to a uniformly expanding medium such as an FLRW universe. Suppose that a photon is emitted in a transition $2 \rightarrow 1$ between two atomic levels 2 and 1. The line is narrow, but not infinitely narrow. As a result of natural and Doppler broadening (the specifics are unimportant here), the line is emitted with some line profile ϕ_λ , which can be taken to be normalized to

$$\int_{\lambda=0}^{\infty} \phi_\lambda d \ln \lambda = 1 . \quad (31.93)$$

The emitted photon travels through the medium, and has some probability of being absorbed by other atoms in level 1 before the photon is redshifted out of the line. Since the line is narrow, the photon is either absorbed nearby, or else it escapes the line completely. In the approximation that the properties of the medium change little over the small distance between emission and absorption, detailed balance implies that the line profile for absorption is the same as that for emission. The cross-section σ_λ for absorption at wavelength λ is

$$\sigma_\lambda = \sigma \phi_\lambda , \quad (31.94)$$

where $\sigma \equiv \int_0^\infty \sigma_\lambda d \ln \lambda$ is the cross-section integrated over the line profile. By detailed balance, the integrated cross-section is related to the Einstein coefficient A_{21} for spontaneous emission by

$$\sigma = \frac{1}{8\pi c} \frac{g_2}{g_1} \lambda_{21}^3 A_{21} . \quad (31.95)$$

The optical depth $d\tau_\lambda$, the differential probability for the photon to be absorbed, as the photon passes through a distance $dl = c dt$ is

$$d\tau_\lambda = n_1 \sigma_\lambda dl = n_1 c \sigma \phi_\lambda dt . \quad (31.96)$$

The medium is expanding with Hubble parameter H , and the photon wavelength λ redshifts by $d \ln \lambda = H dt$ in time dt . Therefore the optical depth to absorption as the photon redshifts through an interval $d \ln \lambda$ of wavelength is

$$d\tau_\lambda = \tau_S \phi_\lambda d \ln \lambda , \quad (31.97)$$

where τ_S is the Sobolev optical depth

$$\tau_S \equiv \frac{n_1 c \sigma}{H} = n_1 \frac{g_2}{g_1} \frac{\lambda_{21}^3 A_{21}}{8\pi H} . \quad (31.98)$$

The optical depth τ_λ for the photon to redshift from an emitted wavelength λ to infinite wavelength is

$$\tau_\lambda \equiv \tau_S \int_\lambda^\infty \phi_{\lambda'} d \ln \lambda' . \quad (31.99)$$

The probability for a photon emitted at wavelength λ to escape from the line without being reabsorbed is the exponential $e^{-\tau_\lambda}$ of the optical depth. The escape probability averaged over the emitted line profile defines

the Sobolev escape probability P_S ,

$$\begin{aligned} P_S &\equiv \int_0^\infty e^{-\tau_\lambda} \phi_\lambda d \ln \lambda = \frac{1}{\tau_S} \int_0^{\tau_S} e^{-\tau_\lambda} d\tau_\lambda \\ &= \frac{1 - e^{-\tau_S}}{\tau_S} , \end{aligned} \quad (31.100)$$

which is evidently independent of the shape of the line profile (just so long as the line is narrow). The Sobolev escape probability P_S varies from 0 as $\tau_S \rightarrow \infty$ to 1 as $\tau_S \rightarrow 0$.

For large Sobolev optical depth τ_S , the Sobolev escape probability approximates the reciprocal of the Sobolev optical depth (31.98),

$$P_S = \frac{1}{\tau_S} \quad (\tau_S \gg 1) . \quad (31.101)$$

The rate per unit time and volume at which photons are emitted and escape is then

$$n_2 A_{21} P_S = \frac{n_2 g_1}{n_1 g_2} \frac{8\pi H}{\lambda_{21}^3} . \quad (31.102)$$

Cosmological perturbations: the hydrodynamic approximation

The simple model in Chapter 30 of the evolution of cosmological perturbations misses some processes that affect in observationally distinctive ways the power spectra of fluctuations both of the CMB and of the distribution of matter.

The most important missing element is baryons, which were neglected in Chapter 30 on the grounds that baryons are gravitationally sub-dominant, having a density $\Omega_b/\Omega_c \approx 1/5$ of the non-baryonic dark matter density. Photons and baryons are coupled by electron-photon scattering, which causes the photons and baryons to behave effectively as a single photon-baryon fluid prior to recombination. Baryons add mass density but no pressure to the photon-baryon fluid, reducing the sound speed of the photon-baryon fluid below its relativistic limit of $\sqrt{1/3}$, §32.4. The reduction in sound speed becomes greater as the ratio of matter to radiation density increases after matter-radiation equality. The baryon mass loading enhances compression (odd) peaks and weakens rarefaction (even) peaks in the power spectrum of the CMB, §32.10. The change in sound speed modifies the relation between the sound horizon and physical distance, resulting in observationally distinctive shifts in the locations of peaks as a function of harmonic number in the power spectrum of the CMB, Figure 34.7. After recombination, baryons decouple from the photons and behave like matter. Oscillations in the photon-baryon fluid at recombination produce an imprint, called **baryon acoustic oscillations**, in the matter power spectrum, Figure 32.4, analogous to the acoustic oscillations in the CMB power spectrum.

A second important effect missing from the simple model of Chapter 30 is dissipation that results from the finite mean free path of electron-photon scattering, which causes photons and baryons not to be perfectly coupled, §32.7. Dissipation damps oscillations of the baryon-photon fluid at smaller scales, reducing power in higher order peaks in the CMB.

A third modification is to treat neutrinos separately from photons, §32.11. Like photons, neutrinos are relativistic, but unlike photons, neutrinos stream freely.

A varying sound speed, dissipation, and freely-streaming neutrinos, can all be modelled in a hydrodynamic approximation that treats the photon-baryon fluid, and the neutrinos, as imperfect fluids. An imperfect fluid is characterized by the first three moments of its momentum distribution, the monopole, dipole, and quadrupole, or equivalently the density, bulk velocity, and pressure, but unlike a perfect fluid the pressure is allowed to be anisotropic. Equations governing the anisotropy can be derived by appealing to a Boltzmann

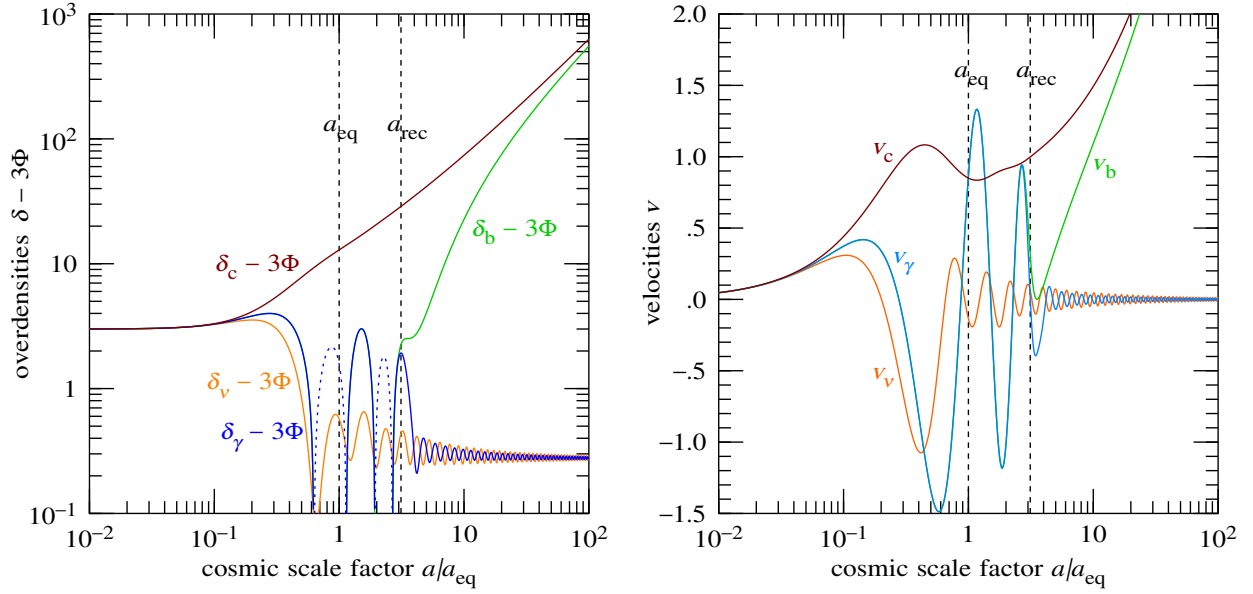


Figure 32.1 (Left) Overdensities $\delta - 3\Phi$, and (right) bulk velocities v in the hydrodynamic approximation as a function of cosmic scale factor a/a_{eq} , at wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$, for non-baryonic dark matter (c), baryons (b), photons (γ), and neutrinos (ν). The cosmological model is the standard model adopted in this book, a flat Λ CDM model with concordance parameters $\Omega_\Lambda = 0.69$ and $\Omega_m = 0.31$, and adiabatic initial conditions, §32.3. The overdensities and velocities of relativistic species are related to their monopole and dipole moments by $\delta_\gamma - 3\Phi = 3(\Theta_0 - \Phi)$, $\delta_\nu - 3\Phi = 3(\mathcal{N}_0 - \Phi)$, $v_\gamma = 3\Theta_1$, $v_\nu = 3\mathcal{N}_1$. The results may be compared to those in the simple approximation, Figure 30.2, and from a Boltzmann computation, Figure 33.1.

treatment, Chapter 33. Given the anisotropy, the evolution of the density and bulk velocity of an imperfect fluid is governed by the equations of conservation of its energy and momentum.

The approximate anisotropic pressure in the hydrodynamic approximation is not sufficiently accurate to provide a reliable source for the difference $\Psi - \Phi$ in scalar gravitational potentials. Thus in the hydrodynamic approximation, as in the simple approximation, the two scalar potentials are set equal, $\Psi = \Phi$.

Figure 32.1 shows the overdensity and bulk velocity of the 4 species, non-baryonic dark matter, baryons, photons, and neutrinos, calculated in the hydrodynamic treatment of this Chapter, as a function of cosmic scale factor, in a flat Λ CDM cosmological model at an illustrative wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$. Figure 32.2 shows photon and neutrino multipoles up to the quadrupole $\ell = 2$, the largest multipole computed in the hydrodynamic approximation. The hydrodynamic approach yields a fair approximation to more accurate calculations that follow higher order multipole moments of the photon and neutrino distributions using the Boltzmann equation, Chapter 33.

This Chapter starts, §32.2, with a summary of the equations in the hydrodynamic approximation. The

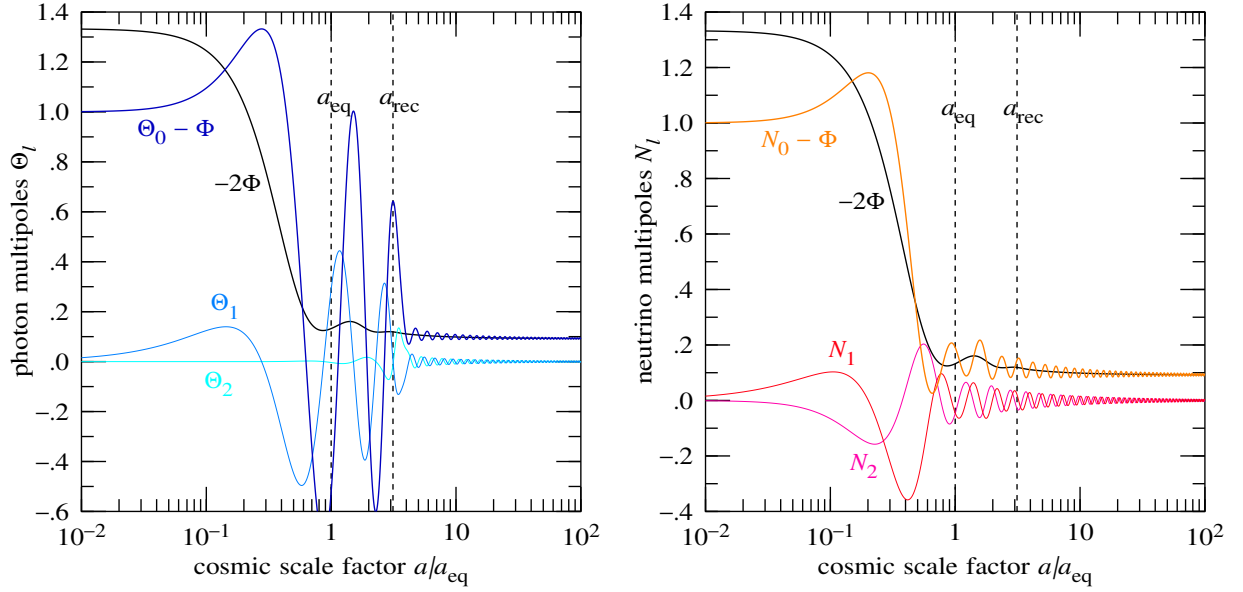


Figure 32.2 (Left) Photon and (right) neutrino multipoles in the hydrodynamic approximation as a function of cosmic scale factor a/a_{eq} , at wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$. The cosmological model is the same as in Figures 32.1–33.1, §32.3. The multipoles may be compared to those from a Boltzmann computation, Figure 33.2.

remainder of the Chapter is concerned with finding approximations to the hydrodynamic system of equations (32.6)–(32.13), so as to gain a physical understanding of their solutions.

Section 32.4 presents the tight-coupling approximation, which effectively treats photons and baryons as a single fluid with a common bulk velocity. The tight-coupling approximation, valid well before recombination, treats the photon-baryon fluid as a perfect fluid, as in the simple approximation of Chapter 30, but the mass density contributed by baryons reduces the sound speed of the fluid.

Sections 32.6–32.10 examine the consequences of allowing quadrupole anisotropy in the photon distribution (shear viscosity), and a small velocity difference between photons and baryons (heat conduction), both of which lead to dissipation.

Section 32.11 considers neutrinos, which stream freely. After recombination, photons also stream freely, as do baryons.

32.1 Electron-photon (Thomson) scattering

For some time before and after recombination, photons and baryons were coupled principally by nonrelativistic electron-photon (Thomson) scattering. The inverse comoving mean free path l_{T}^{-1} to Thomson scattering

is

$$l_T^{-1} \equiv \bar{n}_e \sigma_T a , \quad (32.1)$$

where σ_T is the Thomson cross-section. The Thomson cross-section is proportional to the square of the classical electron radius r_e ,

$$\sigma_T = \frac{8\pi}{3} r_e^2 , \quad r_e = \frac{e^2}{m_e c^2} . \quad (32.2)$$

The inverse comoving mean free path l_T^{-1} is evaluated in Exercise 32.1. In calculating fluctuations in the CMB, Chapter 34, it is convenient to introduce the (dimensionless) Thomson scattering optical depth τ , which starts from zero, $\tau_0 = 0$, at the present time, and increases going backwards in time η to higher redshift,

$$\tau \equiv \int_{\eta}^{\eta_0} \bar{n}_e \sigma_T a d\eta . \quad (32.3)$$

The conformal time derivative of the Thomson optical depth τ equals minus the inverse comoving mean free path,

$$\dot{\tau} \equiv \frac{d\tau}{d\eta} \equiv -\bar{n}_e \sigma_T a . \quad (32.4)$$

Exercise 32.1 Thomson scattering rate. Let f_+ be the proton fraction (31.5), and X_e be the ionization fraction (31.3). Show that the (dimensionless) ratio of the inverse comoving electron-photon (Thomson) mean free path $l_T^{-1} = \bar{n}_e \sigma_T a$ to the inverse comoving Hubble distance $a_{\text{eq}} H_{\text{eq}}/c$ at matter-radiation equality is

$$\begin{aligned} \frac{c \bar{n}_e \sigma_T a}{a_{\text{eq}} H_{\text{eq}}} &= \frac{3c \sigma_T f_+ X_e H_{\text{eq}}}{16\pi G m_b} \frac{\Omega_b}{\Omega_m} \left(\frac{a}{a_{\text{eq}}} \right)^{-2} \\ &= 0.033 h f_+ X_e \frac{H_{\text{eq}}}{H_0} \frac{\Omega_b}{\Omega_m} \left(\frac{a}{a_{\text{eq}}} \right)^{-2} = 500 X_e \left(\frac{a}{a_{\text{eq}}} \right)^{-2} , \end{aligned} \quad (32.5)$$

the Hubble parameter H_{eq} at matter-radiation equality being related to the present-day Hubble parameter H_0 by equation (30.41).

32.2 Summary of equations in the hydrodynamic approximation

The hydrodynamic approximation is derived by suitably truncating the full set of Boltzmann equations, §33.1, at the quadrupole moment. The equations governing the evolution of scalar fluctuations in non-baryonic cold dark matter, baryons, photons, and neutrinos at comoving wavenumber k in the hydrodynamic approximation are as follows (compare to the equations in the simple approximation, §30.7, and in a full Boltzmann treatment, §33.1). The equations for non-baryonic cold dark matter (c) follow from conservation

of energy-momentum, and are the same as those (30.52) in the simple approximation (recall that overdot signifies the derivative $d/d\eta$ with respect to conformal time),

$$\dot{\delta}_c - k v_c - 3\dot{\Phi} = 0 , \quad (32.6a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k \Psi = 0 . \quad (32.6b)$$

Equations for baryons (b) are similar to those (32.6) for the non-baryonic dark matter, except that photon-electron scattering causes a transfer of momentum between photons and baryons when their bulk velocities are not equal,

$$\dot{\delta}_b - k v_b - 3\dot{\Phi} = 0 , \quad (32.7a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k \Psi = -\frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) , \quad (32.7b)$$

where R is $\frac{3}{4}$ the baryon-to-photon density ratio, equation (32.46). The equations of conservation of energy and momentum of photons (γ) are

$$\dot{\Theta}_0 - k \Theta_1 - \dot{\Phi} = 0 , \quad (32.8a)$$

$$\dot{\Theta}_1 + \frac{k}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3} \Psi = \frac{1}{3} |\dot{\tau}| (v_b - 3\Theta_1) . \quad (32.8b)$$

The photon quadrupole moment Θ_2 can be approximated by an expression that interpolates between the tight-coupling limit $|\dot{\tau}| \gg k_s$, equation (33.83), and the free-streaming limit $|\dot{\tau}| \ll k_s$, equation (33.84), where k_s is an interpolation constant, which numerical comparison to full Boltzmann computations indicates is adequately approximated by twice the inverse Hubble distance at recombination, $k_s \approx 2a_{\text{rec}}H_{\text{rec}}$ (or $k_s \approx a_{\text{eq}}H_{\text{eq}}$, for standard Λ CDM cosmological parameters),

$$\Theta_2 = \frac{1}{1 + (|\dot{\tau}|/k_s)^2} \left(\frac{|\dot{\tau}|^2}{k_s^2} \Theta_2^{\text{tight}} + \Theta_2^{\text{free}} \right) , \quad (32.9a)$$

$$\Theta_2^{\text{tight}} = -\frac{8k}{15|\dot{\tau}|} \Theta_1 , \quad \Theta_2^{\text{free}} = -(\Theta_0 + \Psi) - \frac{3}{k\eta} \Theta_1 . \quad (32.9b)$$

As commented after equation (32.67), the factor $\frac{8}{15}$ in equation (32.9b) includes the effect of polarization; without polarization, the factor is $\frac{4}{9}$. Energy-momentum conservation of neutrinos (ν) implies

$$\dot{\mathcal{N}}_0 - k \mathcal{N}_1 - \dot{\Phi} = 0 , \quad (32.10a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3} (\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3} \Psi = 0 . \quad (32.10b)$$

The neutrino quadrupole $\mathcal{N}_2$ may be approximated by, equation (34.50),

$$\mathcal{N}_2 = -(\mathcal{N}_0 + \Psi) - \frac{3}{k\eta} \mathcal{N}_1 . \quad (32.11)$$

The Einstein energy equation is

$$-k^2\Phi - 3\frac{\dot{a}}{a}F = 4\pi Ga^2(\bar{\rho}_c\delta_c + \bar{\rho}_b\delta_b + 4\bar{\rho}_\gamma\Theta_0 + 4\bar{\rho}_\nu\mathcal{N}_0), \quad (32.12)$$

where F is defined by equation (30.55). The non-vanishing photon and neutrino quadrupoles Θ_2 and $\mathcal{N}_2$ are a source for the difference $\Psi - \Phi$ in scalar gravitational potentials, equation (29.56d). However, the hydrodynamic approximations (32.9) and (32.11) are not sufficiently accurate to serve as a reliable source for $\Psi - \Phi$. Therefore in the hydrodynamic approximation the two potentials are set equal, as in the simple approximation (30.57),

$$\Psi = \Phi. \quad (32.13)$$

Exercise 32.2 Program the equations in the hydrodynamic approximation. Upgrade the code you wrote in Exercise 30.11 to implement the hydrodynamic approximation, equations (32.6)–(32.13). Explore the evolution of the gravitational potential Φ , and of the 4 species of mass-energy, non-baryonic dark matter, baryons, photons, and neutrinos. You will upgrade this code to a Boltzmann code in Exercise 33.1.

Solution. See Figures 32.1, 32.2, and 32.3.

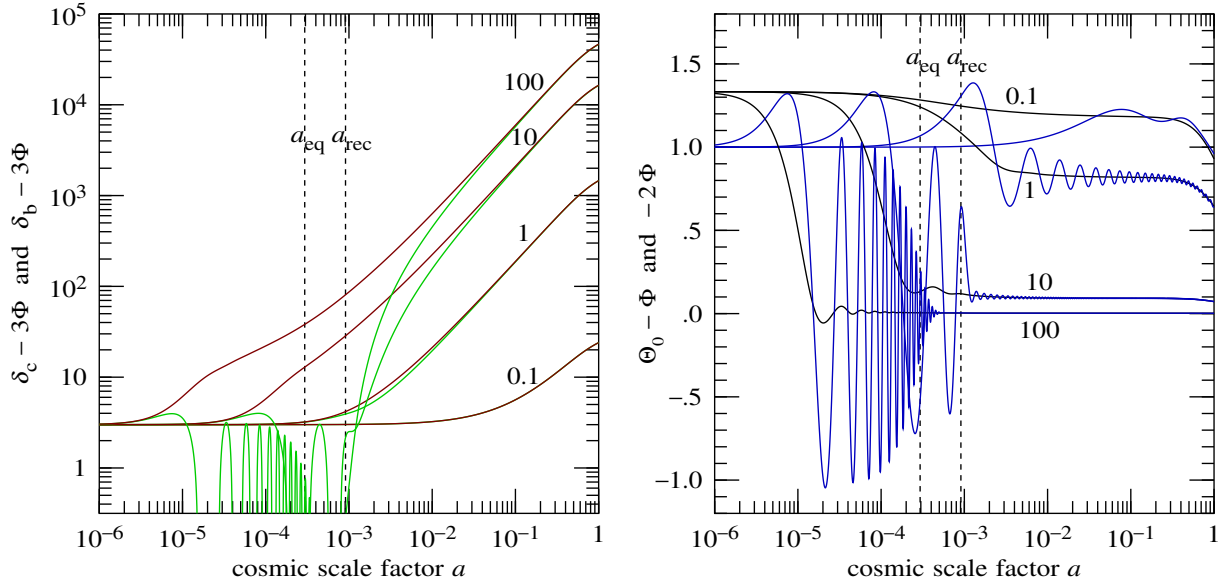


Figure 32.3 (Left) Overdensities $\delta_c - 3\Phi$ and $\delta_b - 3\Phi$ of non-baryonic dark matter (brown) and baryonic matter (green), and (right) radiation monopole $\Theta_0 - 3\Phi$ (blue), and minus twice the scalar potential, -2Ψ (black), as a function of cosmic scale factor a in the hydrodynamic approximation. Curves are labelled with the comoving wavenumber $k/(a_{\text{eq}} H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. The results may be compared to those in the simple approximation, Figure 30.1, and using a Boltzmann computation, Figure 33.3.

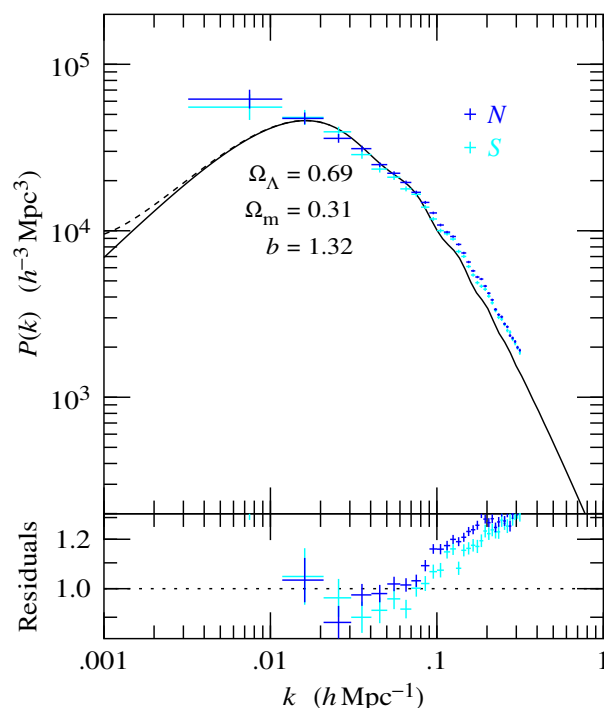


Figure 32.4 Model matter power spectrum computed in the hydrodynamic approximation, compared to observations from the North (N) and South (S) Galactic Caps of the Sloan Digital Sky Survey IV (Gil-Marín et al., 2020). The predicted power spectrum has been multiplied, arbitrarily, by a bias factor of $b^2 = 1.32^2$. The model power spectrum may be compared to those computed in the simple approximation, Figure 30.15, and from a Boltzmann computation, Figure 33.5.

Exercise 32.3 Power spectrum of matter fluctuations: hydrodynamic approximation. Upgrade the code you wrote in Exercise 30.16 to compute the power spectrum of matter fluctuations in the hydrodynamic approximation. Comment on how the power spectrum differs from that in the simple approximation.

Solution. See Figure 32.4. The cosmological model is the standard flat Λ CDM model described in §32.3. The model power spectrum differs from that in the simple approximation, Figure 30.15, firstly in that power is slightly reduced at smaller scales (larger wavenumbers k), and secondly in that the power spectrum shows wiggles, commonly called baryon acoustic oscillations, or BAO. Both effects arise from the finite contribution of baryons to the matter power spectrum.

The possibility of scale-dependent bias between galaxies and matter, coupled with the effects of nonlinear growth of power, complicates the relation between the observed galaxy power spectrum and the linear matter power spectrum. BAO persist in the presence of scale-dependent bias, providing a cosmic ruler that links the comoving scale of distance in galaxy clustering to that in the CMB. Gil-Marín et al. (2020) were interested

primarily in the scale of the BAO. In Figure 10.5 they allowed for possible scale-dependent bias and incipient non-linearity by applying a more or less arbitrary polynomial correction to the model power spectrum.

Exercise 32.4 Effect of massive neutrinos on the matter power spectrum.

1. Incorporate 1 or more species of massive neutrino into the Friedmann equations describing the evolution of the background FLRW geometry.
2. Compute the effect of 1 or more species of massive neutrino on the matter power spectrum. For simplicity, assume an abrupt transition of the neutrino evolution equations from relativistic to nonrelativistic.

Solution

1. Neutrinos decouple while relativistic, at around $e\bar{e}$ -annihilation, and inherit a relativistic thermodynamic distribution from that time. Since decoupling, neutrinos free-streamed, with particle momenta p and temperature T redshifting as $p \propto T \propto a^{-1}$. The energy density $\rho(m, T)$ of a single species of neutrino of mass m at temperature T is (units $c = 1$)

$$\rho(m, T) = \int_0^\infty \sqrt{p^2 + m^2} \frac{1}{e^{p/T} + 1} \frac{4\pi p^2 dp}{(2\pi\hbar)^3} = \frac{7\pi^2 T^4}{240 \hbar^3} R(m/T) , \quad (32.14)$$

where $R(\mu)$ is the integral

$$R(\mu) \equiv \frac{120}{7\pi^4} \int_0^\infty \frac{\sqrt{x^2 + \mu^2}}{e^x + 1} x^2 dx \rightarrow \begin{cases} 1 & \mu \rightarrow 0 , \\ \alpha\mu & \mu \rightarrow \infty , \end{cases} \quad (32.15)$$

with

$$\alpha = \frac{180 \zeta(3)}{7\pi^4} = 0.3173 . \quad (32.16)$$

The neutrino pressure $p(m, T)$ (not to be confused with neutrino particle momentum p) is

$$p(m, T) = \frac{1}{3} \int_0^\infty \frac{p^2}{\sqrt{p^2 + m^2}} \frac{1}{e^{p/T} + 1} \frac{4\pi p^2 dp}{(2\pi\hbar)^3} , \quad (32.17)$$

which can be expressed in terms of the neutrino density $\rho(m, T)$ as

$$p(m, T) = \frac{1}{3} \left[\rho(m, T) - \frac{\partial \rho(m, T)}{\partial \ln m} \right] . \quad (32.18)$$

An approximation good to 1% for the density $\rho(m, T)$, and which yields an approximation good to 4% for the pressure $p(m, T)$ given in terms of $\rho(m, T)$ by formula (32.18), is

$$R(\mu) \approx \sqrt{\frac{1 + \beta\mu^2 + \gamma\alpha^2\mu^4}{1 + \gamma\mu^2}} , \quad (32.19)$$

where the constants α (equation (32.16)), β, γ are chosen such that both density ρ and pressure p have the correct asymptotic behaviour at both $\mu \rightarrow 0$ and $\mu \rightarrow \infty$,

$$\beta = \frac{10}{7\pi^2} + \gamma = 0.2902 , \quad \gamma = \frac{10 [-7\pi^2 + 3240 \zeta(3)^2]}{49\pi^8 - 486000 \zeta(3)\zeta(5)} = 0.1454 . \quad (32.20)$$

A simple approximation that reproduces the correct asymptotic behaviour of the density $\rho(m, T)$ at large and small temperature is to adopt an abrupt change from relativistic to nonrelativistic at $T = \alpha m$,

$$\rho(m, T) \approx \frac{7\pi^2 T^3}{240 \hbar^3} \begin{cases} T & T \geq \alpha m , \\ \alpha m & T \leq \alpha m . \end{cases} \quad (32.21)$$

2. The approximation (32.21) for the neutrino density suggests adopting an abrupt transition of the neutrino evolution equations from relativistic, equations (32.10) and (32.11), to nonrelativistic at $T = \alpha m$, with α from equation (32.16). The nonrelativistic equations are as for non-baryonic cold dark matter, equations (32.6). Conservation of energy and momentum at the transition requires that the neutrino overdensity δ_ν and bulk velocity v_ν are

$$\delta_\nu = 3\mathcal{N}_0 , \quad (32.22a)$$

$$v_\nu = 3\mathcal{N}_1 . \quad (32.22b)$$

32.3 Standard cosmological parameters

Unless otherwise stated, all computations of cosmological perturbations carried out in this book are for a standard flat Λ CDM cosmological model with parameters consistent with those reported by the Planck Collaboration (2018). This section gives the standard parameters adopted in this book.

The CMB power spectrum constrains the physical density Ωh^2 of dark matter and baryonic components more precisely than the density Ω relative to the critical density. The physical matter densities $\Omega_c h^2$ of non-baryonic cold dark matter and $\Omega_b h^2$ of baryonic matter today are taken to be, in the standard model,

$$\Omega_c h^2 = 0.12 , \quad \Omega_b h^2 = 0.022 . \quad (32.23)$$

The conversion factor between Ωh^2 and mass density ρ today is

$$\rho = \frac{3\Omega H^2}{8\pi G c^2} = 6.44932 \times 10^{-26} \Omega h^2 \text{ kg m}^{-3} . \quad (32.24)$$

The matter density Ω_m in non-baryonic cold dark matter and baryonic components today is taken to be, in the standard model,

$$\Omega_m = 0.31 . \quad (32.25)$$

The individual non-baryonic cold dark matter and baryonic densities are then

$$\Omega_c = 0.262 , \quad \Omega_b = 0.048 . \quad (32.26)$$

Together, equations (32.23) and (32.25) yield a Hubble parameter H_0 today of

$$H_0 \equiv 100 h \text{ km s}^{-1} \text{ Mpc}^{-1} = 67.7 \text{ km s}^{-1} \text{ Mpc}^{-1} . \quad (32.27)$$

The CMB temperature T_0 today is (Fixsen, 2009)

$$T_0 = 2.7255 \text{ K} , \quad (32.28)$$

implying a physical photon density of, Exercise 10.3,

$$\Omega_\gamma h^2 = 2.4728 \times 10^{-5} . \quad (32.29)$$

The standard model adopted here assumes $N_{\text{eff}} = 3$ species of massless neutrino that decouple just before electron-positron annihilation, implying that the neutrino temperature after $e\bar{e}$ -annihilation is $T_\nu/T_\gamma = (4/11)^{1/3}$, Exercise 10.21. The energy-weighted effective number of relativistic particle species at recombination is then, equation (10.153b),

$$g_\rho = 2 \left[1 + \left(\frac{4}{11} \right)^{4/3} \frac{7}{8} N_{\text{eff}} \right] = 3.36 . \quad (32.30)$$

In reality, neutrinos are not quite decoupled by $e\bar{e}$ -annihilation. In a more accurate treatment, the neutrino temperature after $e\bar{e}$ -annihilation is slightly larger than $T_\nu/T_\gamma = (4/11)^{1/3}$, and the effective number g_ρ of relativistic species at recombination is correspondingly slightly larger. It is conventional to quote the increase in g_ρ as if it were an increase in the effective number of neutrino types in equation (32.30), $N_{\text{eff}} = 3.04$ (Mangano et al., 2002). In the approximation of $N_{\text{eff}} = 3$ massless neutrinos adopted here, the physical density of neutrinos today is

$$\Omega_\nu h^2 = 1.68 \times 10^{-5} . \quad (32.31)$$

The ratio of the physical matter density from equations (32.23) to the physical radiation density implied by equations (32.29) and (32.31) implies a redshift of matter-radiation equality of

$$1 + z_{\text{eq}} = 3415 . \quad (32.32)$$

If neutrinos have masses as indicated by neutrino oscillation data, §10.25, then at least 2 of the 3 neutrino species are nonrelativistic today, even though they were relativistic at recombination. If the third species is taken to be massless, then the neutrino masses are

$$m_{\nu_1} = 0 \text{ eV} , \quad m_{\nu_2} = 0.01 \text{ eV} , \quad m_{\nu_3} = 0.05 \text{ eV} . \quad (32.33)$$

The corresponding physical neutrino mass density today is, in place of equation (32.31),

$$\Omega_\nu h^2 = 1.3 \times 10^{-3} . \quad (32.34)$$

The assumption of spatial flatness implies vanishing spatial curvature,

$$\Omega_k = 0 . \quad (32.35)$$

In the standard Λ CDM model, the remaining density is taken to be vacuum energy, equivalent to a cosmological constant, with density

$$\Omega_\Lambda = 1 - \Omega_k - \Omega_m - \Omega_r = 0.69 . \quad (32.36)$$

Recombination is affected by the helium mass fraction $Y_{\text{He}} \equiv \rho_{\text{He}}/(\rho_{\text{H}} + \rho_{\text{He}})$, taken to be (Cyburt et al., 2016)

$$Y_{\text{He}} = 0.25 . \quad (32.37)$$

Given the helium fraction (32.37) and the Peebles approximation to recombination, §31.8, along with the standard parameters adopted here, the redshift of recombination, where the Thomson optical depth is unity, is

$$1 + z_{\text{rec}} = 1092 . \quad (32.38)$$

In integrating the simple or hydrodynamic or Boltzmann equations, §30.7 or §32.2 or §33.1, I find it convenient to work in units where the scale factor and Hubble parameter are one at matter-radiation equality, $a_{\text{eq}} = H_{\text{eq}} = 1$. With the standard parameters adopted here, the scale factor and Hubble parameter today are related to those at matter-radiation equality by

$$\frac{a_0}{a_{\text{eq}}} = 3415 , \quad \frac{H_0}{H_{\text{eq}}} = 6.363 \times 10^{-6} . \quad (32.39)$$

The scale factor and Hubble parameter at recombination, where the Thomson optical depth τ is one, are related to those at matter-radiation equality by

$$\frac{a_{\text{rec}}}{a_{\text{eq}}} = 3.13 , \quad \frac{H_{\text{rec}}}{H_{\text{eq}}} = 0.147 . \quad (32.40)$$

The Hubble distance today relative to those at matter-radiation equality and at recombination are

$$\frac{c}{a_0 H_0} = 46.0 \frac{c}{a_{\text{eq}} H_{\text{eq}}} = 21.1 \frac{c}{a_{\text{rec}} H_{\text{rec}}} . \quad (32.41)$$

In cosmology, distances are commonly reported in units of $h^{-1}\text{Mpc}$, or, if the Hubble parameter h today is considered to be known, in Mpc . The Hubble distance today is

$$\frac{c}{a_0 H_0} = 2997.92458 h^{-1}\text{Mpc} = 4.43 \text{ Gpc} . \quad (32.42)$$

The horizon distance today is

$$\eta_0 = 147 \frac{c}{a_{\text{eq}} H_{\text{eq}}} = 3.20 \frac{c}{a_0 H_0} = 9600 h^{-1}\text{Mpc} = 14.2 \text{ Gpc} . \quad (32.43)$$

The age of the Universe today is, equation (10.15),

$$t_0 = 0.955 H_0^{-1} = 13.8 \text{ Gyr} . \quad (32.44)$$

32.4 The photon-baryon fluid in the tight-coupling approximation

Prior to recombination, nonrelativistic electron-photon (Thomson) scattering kept photons tightly coupled to electrons, and Coulomb scattering kept electrons tightly coupled to baryons (nuclei, mostly protons and

helium ions). Thus photons and baryons/electrons behaved as a single tightly-coupled fluid. The baryonic fluid contributed negligible pressure to the combined baryon-photon fluid, but it contributed a finite energy density that became increasingly important as recombination approached. The density loading decreased the sound speed of the photon-baryon fluid, equation (32.50), to the point that at recombination the sound speed was about 80% of the negligible-baryon sound speed of $\sqrt{1/3}$.

Electron-photon scattering transfers momentum between the baryonic fluid and photons, but it does not transfer energy, since the baryons and electrons, being nonrelativistic, have negligible pressure, so their energy is just that of their rest mass. Consequently the energy conservation equation (30.13a) holds separately for each of the photon and baryon fluids. However, the exchange of momentum means that the momentum equation (30.13b) does not hold separately for each fluid. Rather, electron-photon scattering couples the fluids so that their bulk velocities are the same to a good approximation,

$$v_b = v_\gamma , \quad (32.45)$$

the photon bulk velocity being related to the photon dipole by $v_\gamma = 3\Theta_1$, equation (30.15). The approximation (32.45) is called the **tight-coupling** approximation. The right panel of Figure 32.1 illustrates that the equality (32.45) of baryon and photon bulk velocities holds up to recombination, but then breaks down as the scattering mean free path becomes large, and baryons and photons are released from each other's grasp.

Define R to be $\frac{3}{4}$ the baryon-to-photon density ratio,

$$R \equiv \frac{3\bar{\rho}_b}{4\bar{\rho}_\gamma} = R_a \frac{a}{a_{\text{eq}}} , \quad R_a = \frac{3g_\rho\Omega_b}{8\Omega_m} \approx 0.2 , \quad (32.46)$$

with $g_\rho = 3.36$ being the energy-weighted effective number of relativistic particle species at around the time of recombination, equation (10.153b). The energy flux of the combined photon-baryon fluid is, from equation (30.9b),

$$f_\gamma + f_b = (\bar{\rho}_\gamma + \bar{p}_\gamma)v_\gamma + \bar{\rho}_b v_b = \frac{4}{3}\bar{\rho}_\gamma v(1 + R) , \quad (32.47)$$

where v is the common bulk velocity of the photon-baryon fluid. The equation of momentum conservation of the combined baryon-photon fluid is then a sum of the photon-velocity equation (30.13b) with $w = 1/3$, and R times the baryon-velocity equation (30.13b) with $w = 0$. The resulting momentum conservation equation is

$$(1 + R)\dot{v} + R\frac{\dot{a}}{a}v + \frac{k}{3}\delta_\gamma = -k(1 + R)\Psi . \quad (32.48)$$

Combining the photon energy conservation equation (30.13a) with the momentum conservation equation (32.48), and substituting $\delta_\gamma = 3\Theta_0$, equation (30.15), yields

$$\left[\frac{d^2}{d\eta^2} + \frac{R}{1 + R} \frac{\dot{a}}{a} \frac{d}{d\eta} + \frac{k^2}{3(1 + R)} \right] (\Theta_0 - \Phi) = -\frac{k^2}{3(1 + R)} [(1 + R)\Psi + \Phi] , \quad (32.49)$$

which coincides with equation (30.14) for $w = 1/[3(1 + R)]$, and which goes over to the earlier radiation equation (30.47) in the limit $R \rightarrow 0$ of negligible baryons. The term proportional to the first derivative $d/d\eta$

on the left hand side of equation (32.49) is an adiabatic damping term. In the absence of this term, and in the absence of a driving potential, equation (32.49) would reduce to a wave equation with sound speed

$$c_s = \sqrt{\frac{1}{3(1+R)}} . \quad (32.50)$$

The coefficient of the adiabatic damping term in equation (32.49) is, given that $R \propto a$, equation (32.46),

$$\frac{R}{1+R} \frac{\dot{a}}{a} = -2 \frac{\dot{c}_s}{c_s} . \quad (32.51)$$

The **sound horizon distance** η_s is defined to be the distance travelled by a sound wave since the initial time $\eta = 0$,

$$\eta_s \equiv \int_0^\eta c_s d\eta . \quad (32.52)$$

Recast in terms of the sound horizon distance η_s , the differential equation (32.49) is

$$\left(\frac{d^2}{d\eta_s^2} - \frac{c'_s}{c_s} \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = -k^2 [(1+R)\Psi + \Phi] , \quad (32.53)$$

where prime ' denotes derivatives with respect to the sound horizon distance, $c'_s = dc_s/d\eta_s$ (to be distinguished from the derivative with respect to conformal time η).

32.5 WKB approximation

Equation (32.53) is an equation for a forced, damped harmonic oscillator. The forcing terms are those on the right hand side of the equation, while the damping term is the first derivative term on the left hand side. There is a general method, called the WKB approximation (Wentzel, 1926; Kramers, 1926; Brillouin, 1926), to obtain the homogeneous solutions for a damped harmonic oscillator when the damping rate is small compared to the frequency.

Denote the coefficient of the damping term by 2κ . The homogeneous version of equation (32.53) is then

$$\left(\frac{d^2}{d\eta_s^2} + 2\kappa \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = 0 . \quad (32.54)$$

In case being considered, the damping rate κ is the adiabatic rate

$$\kappa = -\frac{1}{2} \frac{c'_s}{c_s} , \quad (32.55)$$

but the WKB method works for more general κ , provided that κ is small compared to the wavenumber of the sound wave, $\kappa \ll k$. The homogeneous wave equation (32.54) can be solved approximately by introducing a frequency ω defined by

$$\Theta_0 - \Phi \propto e^{\int \omega d\eta_s} . \quad (32.56)$$

The homogeneous wave equation (32.54) is then equivalent to

$$\omega' + \omega^2 + 2\kappa\omega + k^2 = 0 . \quad (32.57)$$

To the extent that the damping parameter is much smaller than the frequency, $\kappa \ll k \sim \omega$, the frequency ω is approximately constant, so that ω' can be neglected in equation (32.57). With ω' neglected, the solution of equation (32.57) is

$$\omega = -\kappa \pm i\sqrt{k^2 - \kappa^2} \approx -\kappa \pm ik , \quad (32.58)$$

where the last approximation holds because $\kappa \ll k$. Equation (32.58) is called the **WKB approximation**. Thus the homogeneous solutions of the wave equation (32.54) are approximately

$$\Theta_0 - \Phi \propto e^{-\int \kappa d\eta_s \pm ik\eta_s} . \quad (32.59)$$

32.5.1 Radiation in the tight-coupling approximation

In the tight-coupling approximation, the damping rate κ in the differential equation (32.54) is the adiabatic damping rate (32.55). The integral of the adiabatic damping term is $\int \kappa_a d\eta_s = -\frac{1}{2} \ln c_s$, whose exponential is

$$e^{-\int \kappa_a d\eta_s} = \sqrt{c_s} . \quad (32.60)$$

In the WKB approximation, the homogeneous solutions to the wave equation (32.54) are

$$\Theta_0 - \Phi \propto \sqrt{c_s} e^{\pm ik\eta_s} . \quad (32.61)$$

This shows that, as the sound speed decreased thanks to the increasing baryon-to-photon density in the expanding Universe, the amplitude of a sound wave decreased as the square root of the sound speed.

32.6 Including quadrupole pressure in the momentum conservation equation

The tight-coupling approximation treats the photon-baryon fluid as a perfect fluid, that is, the pressure is taken to be isotropic in the fluid frame. A better approximation is to allow the photons a small quadrupole anisotropy, which allows diffusive dissipation, §32.7.

The scalar part of the momentum conservation equation (29.52b) in general depends not only on the isotropic pressure p , but also on a traceless quadrupole pressure. Let the dimensionless scalar quadrupole q be defined by its relation to the trace-free quadrupole component of the energy-momentum tensor,

$$T_{\text{quad}}^{ab} = (\bar{\rho} + \bar{p})q \left(\frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) , \quad (\bar{\rho} + \bar{p})q \equiv \left(\hat{k}_a \hat{k}_b - \frac{1}{3} \delta_{ab} \right) T^{ab} . \quad (32.62)$$

For relativistic species such as photons, the dimensionless quadrupole q is related to the quadrupole moment Θ_2 by, equation (33.53d),

$$q = -2\Theta_2 . \quad (32.63)$$

In the presence of a quadrupole pressure, the momentum conservation equation (29.52b) includes a term

$$D_m T_{\text{quad}}^{ma} = \frac{1}{a}(\bar{\rho} + \bar{p})\nabla_a q . \quad (32.64)$$

The net effect is to modify all momentum conservation equations by replacing $\Psi \rightarrow \Psi + q$. The scalar bulk velocity equation (30.13b) is thus modified to

$$\dot{v} + (1 - 3w)\frac{\dot{a}}{a}v + wk\delta = -k(\Psi + q) . \quad (32.65)$$

32.7 Photon diffusion (Silk damping)

The tight coupling between photons and baryons is not perfect, because the mean free path for electron-photon scattering is finite, not zero. The imperfect coupling causes sound waves to damp at scales comparable to and below the mean free path. The damping is greater at smaller scales, leading to a systematic reduction in CMB power at smaller scales by an approximately Gaussian factor, equation (32.84). The damping reduces power, but it does not smooth out the acoustic oscillation structure of the CMB power spectrum, which remains intact.

For photon multipoles $\ell \geq 2$, the electron-photon scattering term on the right hand side of the photon Boltzmann hierarchy (33.81) acts as a damping term that tends to drive the multipoles exponentially into equilibrium (the solution to the homogeneous equation $\dot{\Theta}_\ell + |\dot{\tau}|\Theta_\ell = 0$ is a decaying exponential). As seen in §32.4, in the tight-coupling approximation the monopole and dipole oscillate with a natural frequency of $\omega = c_s k$, where c_s is the sound speed. These oscillations provide a source that propagates upward to higher harmonic numbers ℓ . For scales much larger than a mean free path, $k/|\dot{\tau}| \ll 1$, the time derivative is small compared to the scattering term, $|\dot{\Theta}| \sim c_s k|\Theta| \ll |\dot{\tau}|\Theta$, reflecting the near-equilibrium response of the higher harmonics. For multipoles $\ell \geq 2$, the dominant term on the left hand side of the Boltzmann hierarchy (33.81) is the lowest order multipole, which acts as a driver. Solution of the Boltzmann equations (33.81) then requires that

$$\Theta_{\ell+1} \sim \frac{k}{|\dot{\tau}|}\Theta_\ell \quad \text{for } \ell \geq 2 . \quad (32.66)$$

The relation (32.66) implies that higher order photon multipoles are successively smaller than lower orders, $|\Theta_{\ell+1}| \ll |\Theta_\ell|$, for scales much larger than a mean free path, $k/|\dot{\tau}| \ll 1$. This accords with the physical expectation that electron-photon scattering tends to drive the photon distribution to near isotropy.

To lowest order, dissipation can be taken into account by including the photon quadrupole Θ_2 in the Boltzmann hierarchy (33.81) of photon multipole equations, but still neglecting the higher multipoles, $\Theta_\ell = 0$ for $\ell \geq 3$. According to the estimate (32.66), this approximation is valid for scales much larger than a mean free path, $k/|\dot{\tau}| \ll 1$. The approximation of truncating at the quadrupole is equivalent to a diffusion approximation. In the diffusion approximation, the photon quadrupole equation (33.81c) reduces to

$$\Theta_2 = -\frac{4k}{9|\dot{\tau}|}\Theta_1 . \quad (32.67)$$

Substituted into the photon momentum equation (32.8b), the photon quadrupole Θ_2 (32.67) acts as a source of friction on the photon dipole Θ_1 . In hydrodynamics of near-equilibrium fluids, such a quadrupole moment is called **shear viscosity**. When polarization is included, which modifies the factor on the right hand side of the photon quadrupole equation (33.81c), the factor $\frac{4}{9}$ in equation (32.67) is increased by a factor of $\frac{6}{5}$ to $\frac{8}{15}$, equation (35.72), as already adopted in equations (32.9).

The diffusive damping resulting from a small photon quadrupole Θ_2 conserves the energy and momentum of the photon fluid (by itself, irrespective of baryons), so that covariant momentum conservation $D_m T^{mn} = 0$ continues to hold true within the photon fluid. The contribution of a quadrupole pressure to the momentum conservation equation was discussed in §32.6.

32.8 Viscous baryon drag damping

A second source of damping of sound waves, distinct from the photon diffusion of §32.7, arises from the viscous drag on photons that results from a small difference $v_b - 3\Theta_1$ between the baryon and photon bulk velocities. In contrast to photon diffusion, viscous baryon drag transfers momentum between photons and baryons. In hydrodynamics of near-equilibrium fluids, this effect is called **heat conduction**.

An expression for the bulk velocity difference $v_b - 3\Theta_1$ follows from either of the momentum conservation equations (32.8b) or (32.7b) for photons or baryons,

$$v_b - 3\Theta_1 = \frac{3}{|\dot{\tau}|} \left(\dot{\Theta}_1 + \frac{k}{3}\Theta_0 + \frac{k}{3}\Psi \right) = -\frac{R}{|\dot{\tau}|} \left(\dot{v}_b + \frac{\dot{a}}{a}v_b + k\Psi \right) \approx -\frac{3R}{|\dot{\tau}|} \left(\dot{\Theta}_1 + \frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}\Psi \right) . \quad (32.68)$$

The bulk velocity difference $v_b - 3\Theta_1$ is small because the scattering factor $|\dot{\tau}|$ is large. The final approximation of equations (32.68) follows from replacing v_b with $3\Theta_1$ to lowest order, which is valid because the expression is already of linear order. Taking a linear combination of the second and fourth expressions in equations (32.68) so as to eliminate $\dot{\Theta}_1$ gives

$$v_b - 3\Theta_1 \approx \frac{3R}{(1+R)|\dot{\tau}|} \left(\frac{k}{3}\Theta_0 - \frac{\dot{a}}{a}\Theta_1 \right) . \quad (32.69)$$

On the right hand side of equation (32.69), the wavenumber k is large compared to $\dot{a}/a$ at the subhorizon scales where dissipation is important, so the bulk velocity difference reduces to

$$v_b - 3\Theta_1 \approx \frac{Rk}{(1+R)|\dot{\tau}|} \Theta_0 . \quad (32.70)$$

It is tempting to insert the approximation (32.70) directly into the right hand sides of the photon and baryon momentum conservation equations (32.8b) and (32.7b), but the result is not of the desired precision, since the right hand sides of the momentum equations are multiplied by the large factor $|\dot{\tau}|$, amplifying imprecision in the approximation (32.70). A precise approach is to start with the equation of conservation of total momentum of the photon-baryon fluid, which is a sum of the momentum conservation equations (32.8b)

and (32.7b) for photons and baryons,

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi + \frac{R}{3}\left(\dot{v}_b + \frac{\dot{a}}{a}v_b + k\Psi\right) = 0. \quad (32.71)$$

Rewriting the baryon velocity as the photon velocity plus a small difference, $v_b = 3\Theta_1 + (v_b - 3\Theta_1)$, brings the momentum conservation equation (32.71) to

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi + R\left(\dot{\Theta}_1 + \frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}\Psi\right) + \frac{R}{3}\left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)(v_b - 3\Theta_1) = 0. \quad (32.72)$$

The term in equation (32.72) involving the velocity difference $v_b - 3\Theta_1$ is proportional to

$$\left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)(v_b - 3\Theta_1) = \left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)\frac{Rk}{(1+R)|\dot{\tau}|}\Theta_0 \approx \frac{Rk}{(1+R)|\dot{\tau}|}\dot{\Theta}_0 \approx \frac{Rk^2}{(1+R)|\dot{\tau}|}\Theta_1, \quad (32.73)$$

the second step of which invokes the approximation (32.70), and the last two steps of which retain only the dominant term at the subhorizon scales $k\eta \gg 1$ where dissipation is important.

Substituting the approximation (32.73), and the diffusive approximation (32.67) for the radiation quadrupole Θ_2 , brings the photon-baryon momentum conservation equation (32.72) to

$$(1+R)\dot{\Theta}_1 + R\frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}[\Theta_0 + (1+R)\Psi] + \frac{k^2}{3|\dot{\tau}|}\left(\frac{8}{9} + \frac{R^2}{1+R}\right)\Theta_1 = 0. \quad (32.74)$$

The final terms proportional to the comoving Thomson mean free path $1/|\dot{\tau}|$ on the left hand side of equation (32.74) are the dissipative terms. The $8/9$ term is from photon diffusion, while the $R^2/(1+R)$ term is from baryon drag.

32.9 Photon-baryon wave equation with dissipation

Eliminating the dipole Θ_1 in equation (32.74) in favour of the monopole Θ_0 using the photon monopole equation (32.8) yields a second order differential equation for $\Theta_0 - \Phi$,

$$\left\{\frac{d^2}{d\eta^2} + \left[\frac{R}{(1+R)}\frac{\dot{a}}{a} + \frac{k^2}{3(1+R)|\dot{\tau}|}\left(\frac{8}{9} + \frac{R^2}{1+R}\right)\right]\frac{d}{d\eta} + \frac{k^2}{3(1+R)}\right\}(\Theta_0 - \Phi) = -\frac{k^2}{3(1+R)}[(1+R)\Psi + \Phi]. \quad (32.75)$$

Recast in terms of the sound horizon distance η_s defined by equation (32.52), the wave equation (32.75) becomes

$$\left\{\frac{d^2}{d\eta_s^2} + \left[-\frac{c'_s}{c_s} + \frac{k^2 c_s}{|\dot{\tau}|}\left(\frac{8}{9} + \frac{R^2}{1+R}\right)\right]\frac{d}{d\eta_s} + k^2\right\}(\Theta_0 - \Phi) = -k^2[(1+R)\Psi + \Phi], \quad (32.76)$$

where prime $'$ denotes derivative with respect to sound horizon distance, $c'_s = dc_s/d\eta_s$. Equations (32.75) and (32.76) differ from the earlier dissipation-free equations (32.49) and (32.53) by the inclusion of dissipation

terms proportional to the Thomson scattering mean free path $l_T = 1/|\dot{\tau}|$. WKB solution of equations such as (32.76) was discussed in §32.5.

In Exercise 35.7 it is found that polarization increases the photon diffusion contribution in equation (32.76) by a factor of $\frac{6}{5}$ from $\frac{8}{9}$ to $\frac{16}{15}$,

$$\frac{8}{9} \rightarrow \frac{16}{15} . \quad (32.77)$$

The terms proportional to the linear derivative $d/d\eta_s$ in equation (32.76) are damping terms, which may be collected into an overall damping coefficient κ ,

$$\left(\frac{d^2}{d\eta_s^2} + 2\kappa \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = -k^2 [(1+R)\Psi + \Phi] . \quad (32.78)$$

The damping coefficient κ is a sum of adiabatic κ_a and dissipative κ_d parts,

$$\kappa \equiv \kappa_a + \kappa_d , \quad \kappa_a = -\frac{1}{2} \frac{d \ln c_s}{d\eta_s} , \quad \kappa_d = \frac{k^2 c_s}{2|\dot{\tau}|} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) . \quad (32.79)$$

In the dissipative damping coefficient κ_d , the $16/15$ term arises from photon diffusion, while the $R^2/(1+R)$ term arises from baryon drag. At recombination, where $R \approx 0.6$, dissipation by photon diffusion and baryon drag are in the ratio $(16/15)/[R^2/(1+R)] \approx 5$. Thus photon diffusion dominates the dissipation, but baryon drag contributes non-negligibly.

In the WKB approximation, §32.5, the homogeneous solutions of equation (32.78) are

$$\Theta_0 - \Phi \propto \sqrt{c_s} e^{-\int \kappa_d d\eta_s} e^{\pm i k \eta_s} . \quad (32.80)$$

The dissipative factor $e^{-\int \kappa_d d\eta_s}$ involves an integral of the dissipative damping coefficient over the sound horizon distance, which may be written

$$\int \kappa_d d\eta_s = \frac{k^2}{k_d^2} , \quad (32.81)$$

where k_d^{-1} is the damping scale defined by, from equation (32.79) along with the definition (32.52) of η_s and the relation (32.50) between c_s and R ,

$$\frac{1}{k_d^2} \equiv \int \frac{c_s}{2|\dot{\tau}|} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) d\eta_s = \int \frac{1}{6|\dot{\tau}|(1+R)} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) d\eta . \quad (32.82)$$

The damping scale k_d^{-1} is roughly the geometric mean of the scattering mean free path l_T and the horizon distance η , as might be expected for a random walk by increments l_T over a time η ,

$$k_d^{-1} \sim \sqrt{l_T \eta} . \quad (32.83)$$

The resulting dissipative damping factor is

$$e^{-\int \kappa_d d\eta_s} = e^{-k^2/k_d^2} . \quad (32.84)$$

Thus the effect of dissipation is to damp temperature fluctuations exponentially at scales smaller than the diffusion scale k_d . The diffusion scale k_d is evaluated in Exercise 32.6.

32.10 Baryon loading

The driving potential on the right hand side of the wave equation (32.78) causes $\Theta_0 - \Phi$ to oscillate not around zero, but rather around the offset $-\left[(1+R)\Psi + \Phi\right]$. At scales well inside the sound horizon, $k\eta_s \gg 1$, this driving potential also varies slowly compared to the wave frequency. To the extent that the driving potential is slowly varying, the complete solution of the inhomogeneous wave equation (32.78) well inside the horizon is

$$\Theta_0 + (1+R)\Psi \propto \sqrt{c_s} e^{-k^2/k_d^2} e^{\pm i k \eta_s} . \quad (32.85)$$

As will be seen in Chapter 34, equation (34.17), the monopole contribution to CMB fluctuations is not the photon monopole Θ_0 by itself, but rather $\Theta_0 + \Psi$, which is the monopole redshifted by the potential Ψ . This redshifted monopole is

$$\Theta_0 + \Psi = -R\Psi + A\sqrt{c_s} e^{-k^2/k_d^2} e^{\pm i k \eta_s} , \quad (32.86)$$

with some constant amplitude A . Thus the redshifted monopole $\Theta_0 + \Psi$ oscillates about the offset $-R\Psi$. Physically, the gravity of baryons enhances sound wave compressions while weakening rarefactions. The offset of the redshifted temperature monopole translates into an amplification of compression (odd) peaks in the CMB, and a weakening of rarefaction (even) peaks in the CMB, as is observed in the CMB.

Exercise 32.5 Behaviour of radiation in the presence of damping. Confirm that, for $\kappa \ll k$, the homogeneous solutions of equation (32.78) are approximately $\Theta_0 - \Phi \propto e^{-\int \kappa d\eta_s \pm i k \eta_s}$. Hence find the retarded Green's function, and write down the general solution to equation (32.78). Convince yourself that $\Theta_0 - \Phi$ is a decaying wave that oscillates around $-\left[(1+R)\Psi + \Phi\right]$.

Solution. The general solution of equation (32.78) is, with $y \equiv k\eta_s$ and $\beta \equiv \int \kappa d\eta_s$,

$$\Theta_0(y) - \Phi(y) = e^{-\beta} (A_0 \cos y + A_1 \sin y) - \int_0^y \{[1+R(y')] \Psi(y') + \Phi(y')\} e^{-(\beta-\beta')} \sin(y-y') dy' , \quad (32.87)$$

where A_0 and A_1 are constants.

Exercise 32.6 Diffusion scale. Show that the dimensionless ratio of the damping scale k_d defined by (32.82) to the comoving Hubble distance $c/(a_{\text{eq}} H_{\text{eq}})$ at matter-radiation equality is given by

$$\frac{a_{\text{eq}}^2 H_{\text{eq}}^2}{c^2 k_d^2} = \frac{8\sqrt{2}\pi G m_b}{9c\sigma_T f_+ H_{\text{eq}}} \frac{\Omega_m}{\Omega_b} \int_0^{a/a_{\text{eq}}} \frac{(a/a_{\text{eq}})^2}{X_e \sqrt{1+(a/a_{\text{eq}})(1+R)}} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) d(a/a_{\text{eq}}) . \quad (32.88)$$

If hydrogen is taken to be fully ionized and helium neutral, which is a reasonable approximation in the run-up to recombination, then $X_e = f_H$. For constant X_e , the integral on the right hand side of equation (32.88) can be done analytically. With a normalized to $a_{\text{eq}} = 1$,

$$f(a) \equiv \int_0^a \frac{a^2}{\sqrt{1+a}} \frac{\frac{16}{15}(1+R) + R^2}{(1+R)^2} da \approx \int_0^a \frac{a^2 da}{\sqrt{1+a}} . \quad (32.89)$$

The last approximation is correct to order unity for any a . Conclude that, neglecting the effect of recombination on the electron fraction X_e ,

$$\frac{a_{\text{eq}}^2 H_{\text{eq}}^2}{c^2 k_d^2} = \frac{6.83 h^{-1}}{f_+ f_H} \frac{H_0}{H_{\text{eq}}} \frac{\Omega_m}{\Omega_b} f(a/a_{\text{eq}}) = 6 \times 10^{-4} f(a/a_{\text{eq}}) \approx 0.0035 , \quad (32.90)$$

the final value being the approximate value at recombination.

32.11 Neutrinos

Before electron-positron annihilation at temperature $T \approx 1 \text{ MeV}$, weak interactions were fast enough that scattering between neutrinos, antineutrinos, electrons, and positrons kept neutrinos and antineutrinos in thermodynamic equilibrium with baryons. After $e\bar{e}$ annihilation, neutrinos and antineutrinos decoupled, rather like photons decoupled at recombination. After decoupling, neutrinos streamed freely. In Exercise 32.7 you will show that, in an approximation developed in §34.6.2, the effective sound speed in neutrinos was about the speed of light, in contrast to photons where collisional isotropization leads to a sound speed about $1/\sqrt{3}$ the speed of light.

Exercise 32.7 Generic behaviour of neutrinos. Insert the approximate value (34.50) of the neutrino quadrupole $\mathcal{N}_2$ into the neutrino energy and momentum conservation equations (32.10) to obtain the differential equation

$$\left(\frac{d^2}{d\eta^2} + \frac{2}{\eta} \frac{d}{d\eta} + k^2 \right) (\mathcal{N}_0 - \Phi) = -k^2 (\Psi + \Phi) . \quad (32.91)$$

What kind of equation is this? What are its solutions? Find the Green's function solution driven by a prescribed potential $\Psi + \Phi$, subject to the initial condition that $\mathcal{N}_0 - \Phi = \zeta_\nu$. Convince yourself that $\mathcal{N}_0 - \Phi$ is a decaying wave that oscillates around $-(\Psi + \Phi)$. Exercise 35.8 generalizes this exercise to the case of vector and tensor fluctuations.

Solution. The Green's function solution of equation (32.91) is with $y \equiv k\eta$,

$$\mathcal{N}_0 - \Phi = \zeta_\nu \frac{\sin y}{y} - \int_0^y [\Psi(y') + \Phi(y')] \sin(y - y') \frac{y'}{y} dy' . \quad (32.92)$$

Cosmological perturbations: Boltzmann treatment

Chapters 30 and 32 treated cosmological perturbations in the approximations that matter and radiation behaved as respectively perfect and imperfect fluids. The fluid approximation truncates the momentum dis-

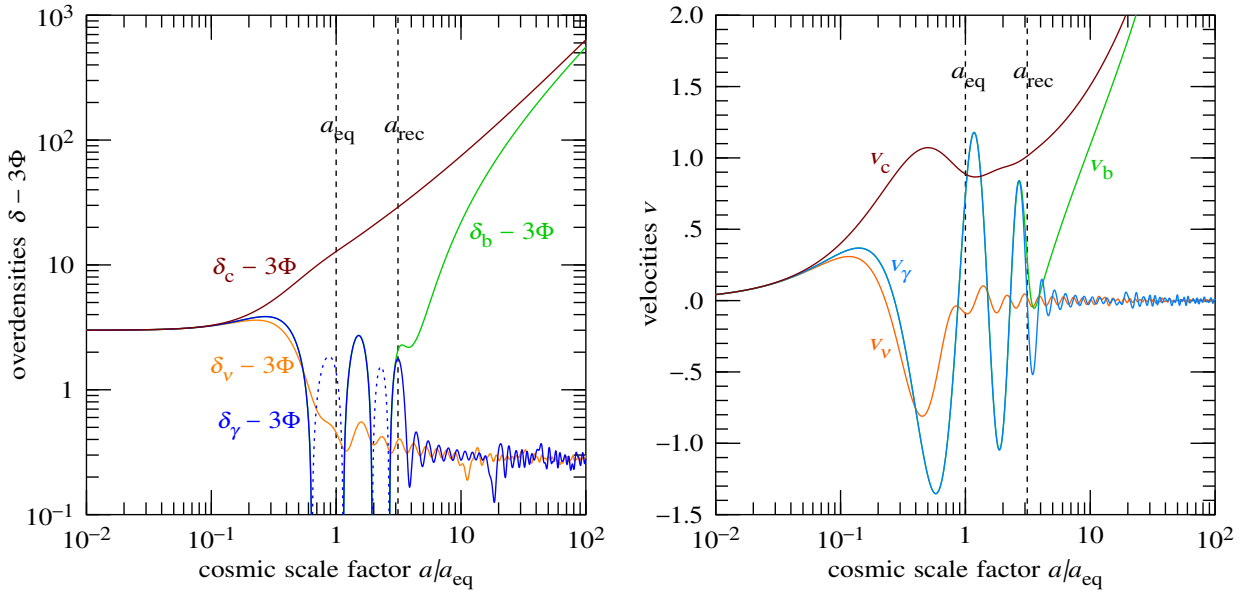


Figure 33.1 (Left) Overdensities $\delta - 3\Phi$, and (right) bulk velocities v in a Boltzmann treatment as a function of cosmic scale factor a/a_{eq} , at wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$, for non-baryonic dark matter (c), baryons (b), photons (γ), and neutrinos (ν). The cosmological model is the standard model adopted in this book, a flat Λ CDM model with concordance parameters $\Omega_\Lambda = 0.69$ and $\Omega_m = 0.31$ and adiabatic initial conditions, §32.3. The overdensities and velocities of relativistic species are related to their monopole and dipole moments by $\delta_\gamma - 3\Phi = 3(\Theta_0 - \Phi)$, $\delta_\nu - 3\Phi = 3(\mathcal{N}_0 - \Phi)$, $v_\gamma = 3\Theta_1$, $v_\nu = 3\mathcal{N}_1$. The computation shown here includes photon and neutrino multipoles up to $\ell_{\text{max}} = 32$. Compare these results to the simple and hydrodynamic computations, Figures 30.2 and 32.1.

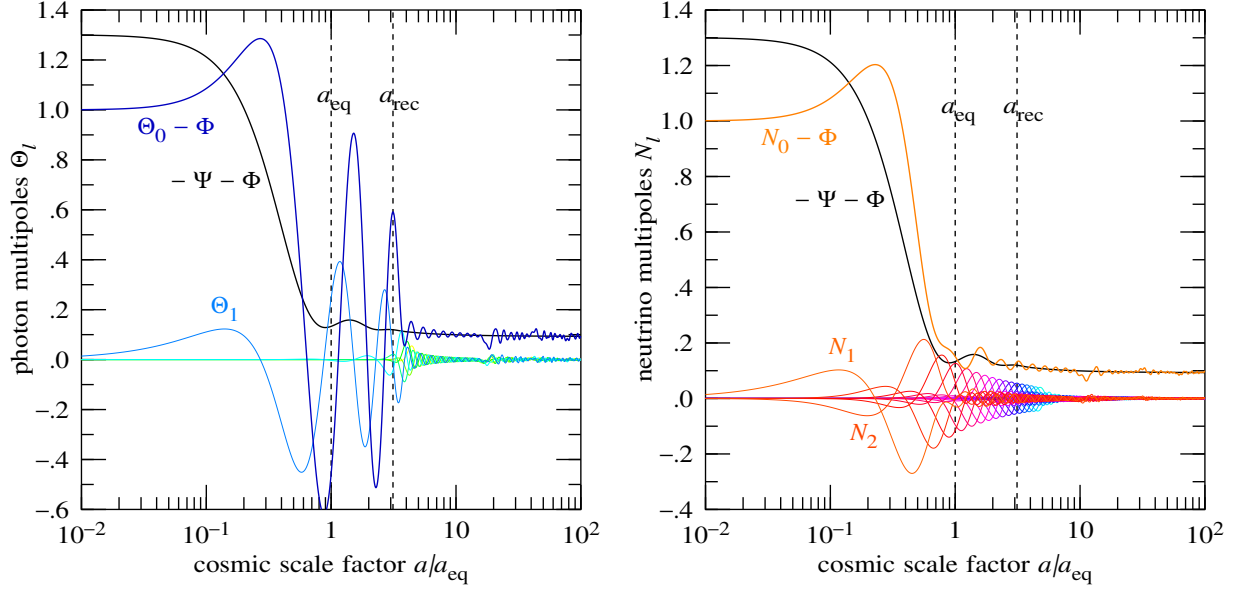


Figure 33.2 (Left) Photon multipoles up to $\ell = 6$, and (right) neutrino multipoles up to $\ell = 32$, as a function of cosmic scale factor a/a_{eq} , at wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$. The cosmological model is the same as in Figure 33.1, §32.3. The thick (black) line shows $-\Psi - \Phi$, about which the photon and neutrino monopoles $\Theta_0 - \Phi$ and $N_0 - \Phi$ mostly oscillate (except near recombination, where the photon monopole $\Theta_0 - \Phi$ oscillates about $-(1+R)\Psi - \Phi$). The computation includes photon and neutrino multipoles up to $\ell_{\text{max}} = 32$. The unphysical jitter in the modes for $a/a_{\text{eq}} \gtrsim 10$ is a symptom of the computation ceasing to be reliable once multipoles higher than those computed become significant. The multipoles may be compared to those in the hydrodynamic approximation, Figure 32.2.

tribution at the quadrupole momentum moment. However, higher order multipole moments of the photon distribution become important near recombination, and a fully satisfactory treatment of the CMB requires following these moments. The evolution of the complete set of multipole moments is governed by the collisional Boltzmann equation.

A Boltzmann treatment is needed in any case to determine how the Boltzmann equations should best be truncated to give the hydrodynamic treatment of Chapter 32. The purpose of the present Chapter is to give an account of the Boltzmann equation as it applies to cosmological perturbation theory.

Figure 33.1 shows the overdensity and bulk velocity of the 4 species, non-baryonic dark matter, baryons, photons, and neutrinos, calculated in the Boltzmann treatment of this Chapter, as a function of cosmic scale factor, at an illustrative wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 10$. Figure 33.2 shows photon multipoles up to $\ell = 6$ and neutrino multipoles up to $\ell = 32$ in a Boltzmann computation that include multipoles up to $\ell_{\text{max}} = 32$ for both photons and neutrinos.

33.1 Summary of equations in the Boltzmann treatment

The Boltzmann treatment uses the Boltzmann hierarchy of equations to follow the evolution of multipole moments of relativistic species, photons and neutrinos, up to some maximum harmonic $\ell_{\max}-1$. The hierarchy is truncated by invoking a suitable approximation for the $\ell_{\max}$ th harmonic. The Boltzmann treatment yields the hydrodynamic approximation, §32.2, when $\ell_{\max} = 2$.

In the Boltzmann treatment, the equations for non-baryonic cold dark matter (c) and baryons (b) are the same as those in the hydrodynamic approximation, equations (32.6) and (32.7),

$$\dot{\delta}_c - k v_c - 3\dot{\Phi} = 0 , \quad (33.1a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k \Psi = 0 , \quad (33.1b)$$

and

$$\dot{\delta}_b - k v_b - 3\dot{\Phi} = 0 , \quad (33.2a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k \Psi = -\frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) . \quad (33.2b)$$

The equations for photons (γ) are given by the Boltzmann hierarchy (33.81),

$$\dot{\Theta}_0 - k \Theta_1 - \dot{\Phi} = 0 , \quad (33.3a)$$

$$\dot{\Theta}_1 + \frac{k}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3} \Psi = \frac{1}{3} |\dot{\tau}| (v_b - 3\Theta_1) , \quad (33.3b)$$

$$\dot{\Theta}_2 + \frac{k}{5} (2\Theta_1 - 3\Theta_3) = -\frac{3}{4} |\dot{\tau}| \Theta_2 , \quad (33.3c)$$

$$\dot{\Theta}_\ell + \frac{k}{2\ell+1} [\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] = -|\dot{\tau}| \Theta_\ell \quad (\ell \geq 3) . \quad (33.3d)$$

As commented after equations (33.81), the factor $\frac{3}{4}$ in equation (33.3c) includes the effect of polarization; without polarization, the factor is $\frac{9}{10}$. The $\ell_{\max}$ th harmonic $\Theta_{\ell_{\max}}$ may be approximated by an expression that interpolates between the tight-coupling limit $|\dot{\tau}| \gg k_s$, equation (33.83), and the free-streaming limit $|\dot{\tau}| \ll k_s$, equation (33.84),

$$\Theta_{\ell_{\max}} = \frac{1}{1 + (|\dot{\tau}|/k_s)^2} \left(\frac{|\dot{\tau}|^2}{k_s^2} \Theta_{\ell_{\max}}^{\text{tight}} + \Theta_{\ell_{\max}}^{\text{free}} \right) , \quad (33.4a)$$

$$\Theta_{\ell_{\max}}^{\text{tight}} = -\frac{(1 + \frac{1}{3}\delta_{\ell_{\max}2}) \ell_{\max} k}{(2\ell_{\max} + 1)|\dot{\tau}|} \Theta_{\ell_{\max}-1} , \quad \Theta_{\ell_{\max}}^{\text{free}} = -(\Theta_{\ell_{\max}-2} + \delta_{\ell_{\max}2} \Psi) - \frac{2\ell_{\max}-1}{k\eta} \Theta_{\ell_{\max}-1} . \quad (33.4b)$$

Equations (33.4) reduces to the hydrodynamic approximation (32.9) when $\ell_{\max} = 2$. As in the hydrodynamic case, numerical experiment indicates that the interpolation constant k_s is adequately approximated by $k_s \approx 2a_{\text{rec}}H_{\text{rec}}$ (or $k_s \approx a_{\text{eq}}H_{\text{eq}}$, for standard Λ CDM cosmological parameters). The equations for neutrinos (ν) are given by a Boltzmann hierarchy (33.91) which looks like that for photons, but without the scattering

terms,

$$\dot{\mathcal{N}}_0 - k\mathcal{N}_1 - \dot{\Phi} = 0 , \quad (33.5a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3}(\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3}\Psi = 0 , \quad (33.5b)$$

$$\dot{\mathcal{N}}_\ell + \frac{k}{2\ell+1}[\ell\mathcal{N}_{\ell-1} - (\ell+1)\mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) . \quad (33.5c)$$

The $\ell_{\max}$ th harmonic $\mathcal{N}_{\ell_{\max}}$ may be approximated by, equation (33.92),

$$\mathcal{N}_{\ell_{\max}} = -(\mathcal{N}_{\ell_{\max}-2} + \delta_{\ell_{\max}2}\Psi) - \frac{2\ell_{\max}-1}{k\eta}\mathcal{N}_{\ell_{\max}-1} . \quad (33.6)$$

The Einstein energy and quadrupole pressure equations are

$$-k^2\Phi - 3\frac{\dot{a}}{a}F = 4\pi Ga^2(\bar{\rho}_c\delta_c + \bar{\rho}_b\delta_b + 4\bar{\rho}_\gamma\Theta_0 + 4\bar{\rho}_\nu\mathcal{N}_0) , \quad (33.7a)$$

$$k^2(\Psi - \Phi) = -32\pi Ga^2(\bar{\rho}_\gamma\Theta_2 + \bar{\rho}_\nu\mathcal{N}_2) , \quad (33.7b)$$

where F is defined by equation (30.55).

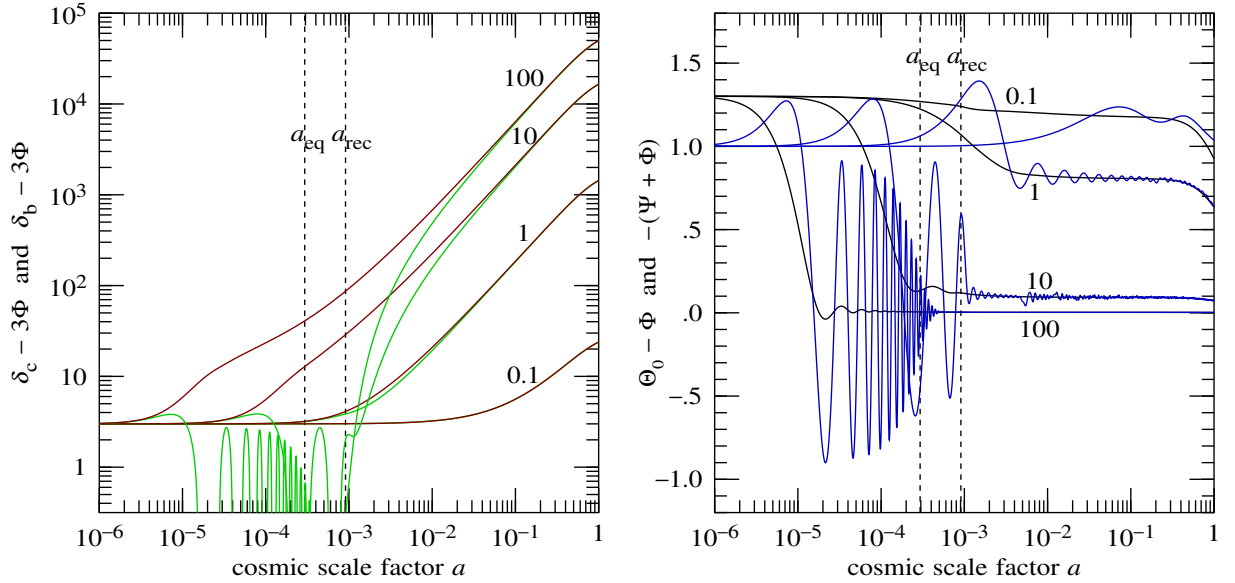


Figure 33.3 (Left) Overdensities $\delta_c - 3\Phi$ and $\delta_b - 3\Phi$ of non-baryonic dark matter (brown) and baryonic matter (green), and (right) radiation monopole $\Theta_0 - 3\Phi$ (blue), and minus the sum of the scalar potentials, $-(\Psi + \Phi)$ (black), as a function of cosmic scale factor a . Curves are labelled with the comoving wavenumber $k/(a_{\text{eq}} H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. The cosmological model is as in §32.3. Compare these results to the simple and hydrodynamic computations, Figures 30.1 and 32.3.

Exercise 33.1 Program the Boltzmann equations. Upgrade the code you wrote in Exercise 32.2 to implement the system of Boltzmann equations (32.6)–(32.13). Initial conditions for neutrinos, and for the two scalar potentials Ψ and Φ , are derived in Exercise 33.5. Explore the evolution of the 2 scalar potentials and of the 4 species of mass-energy, non-baryonic dark matter, baryons, photons, and neutrinos.

Solution. See Figures 33.1, 33.2, 33.3 and 33.4. The computations here included photon and neutrino multipoles up to $\ell_{\max} = 32$.

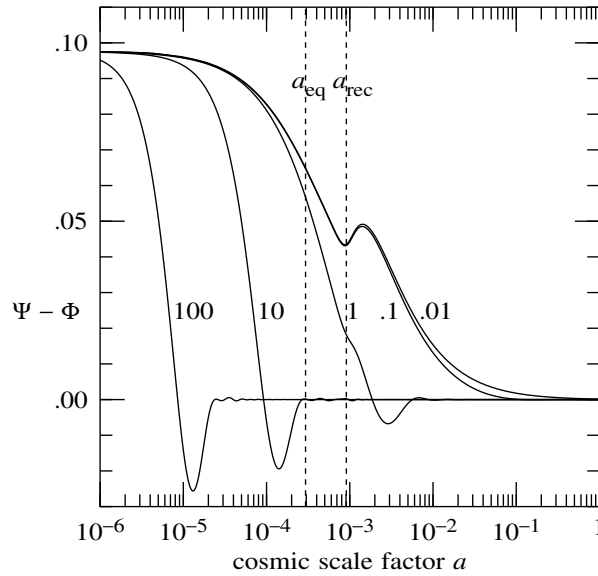


Figure 33.4 Difference $\Psi - \Phi$ in scalar potentials as a function of cosmic scale factor a . The cosmological model is the same as in Figure 33.1, §32.3. Curves are labelled with the wavenumber $k/(a_{\text{eq}} H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. The difference $\Psi - \Phi$ is sourced principally by neutrino anisotropy before recombination, and by photon and neutrino anisotropy after recombination. The computation includes photon and neutrino multipoles up to $\ell_{\max} = 32$.

Exercise 33.2 Power spectrum of matter fluctuations: Boltzmann treatment. Upgrade the code you wrote in Exercise 32.3 to compute the power spectrum of matter fluctuations in a Boltzmann computation.

Solution. See Figure 33.5.

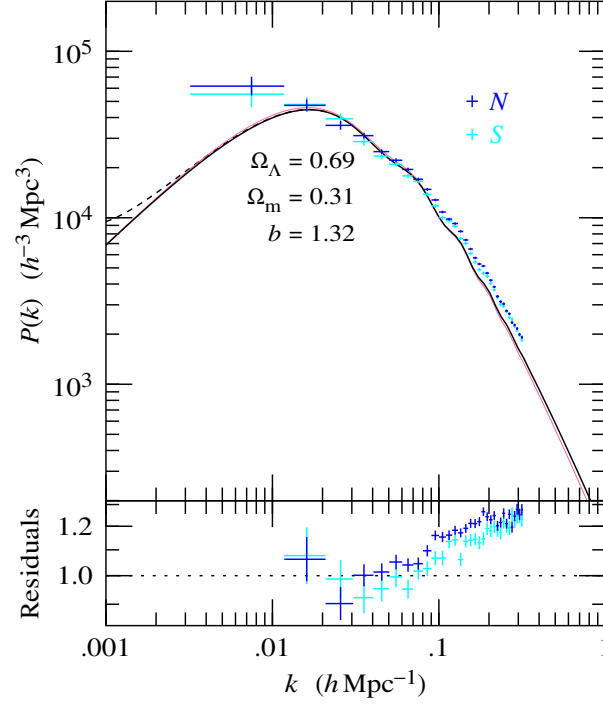


Figure 33.5 Model matter power spectrum computed from a Boltzmann computation, compared to observations from the North (N) and South (S) Galactic Caps of the Sloan Digital Sky Survey IV (Gil-Marín et al., 2020). The cosmological model is the same as in Figure 33.1, §32.3. The model power spectrum may be compared to those computed in the simple and hydrodynamic approximations, Figure 30.15 and Figure 32.4. The thin (pink) line is the model power spectrum in the hydrodynamic approximation.

33.2 Boltzmann equation in a perturbed FLRW geometry

The Boltzmann equation was introduced in §31.5. The left hand side of the Boltzmann equation (31.32) is, for either massless or massive particles,

$$\frac{df}{d\lambda} = p^m \partial_m f + \frac{dp^a}{d\lambda} \frac{\partial f}{\partial p^a} = E \partial_0 f + p^a \partial_a f + \frac{d\hat{\mathbf{p}}}{d\lambda} \cdot \frac{\partial f}{\partial \hat{\mathbf{p}}} + \frac{dp}{d\lambda} \frac{\partial f}{\partial p}. \quad (33.8)$$

Here λ is an affine parameter along the worldline of a particle, and $p^m \equiv \{E, \mathbf{p}\}$ is the tetrad-frame momentum of the particle. Both $d\hat{\mathbf{p}}/d\lambda$ and $\partial f/\partial \hat{\mathbf{p}}$ vanish in the unperturbed background, so $d\hat{\mathbf{p}}/d\lambda \cdot \partial f/\partial \hat{\mathbf{p}}$ is of second order, and can be neglected to linear order, so that

$$\frac{df}{d\lambda} = E \partial_0 f + p^a \partial_a f + \frac{dp}{d\lambda} \frac{\partial f}{\partial p}. \quad (33.9)$$

The expression (33.9) for the left hand side $df/d\lambda$ of the Boltzmann equation involves $dp/d\lambda$, which in free-fall is determined by the usual geodesic equation

$$\frac{dp^k}{d\lambda} + \Gamma_{mn}^k p^m p^n = 0 . \quad (33.10)$$

Since $E^2 - p^2 = m^2$, it follows that the equation of motion for the magnitude p of the tetrad-frame momentum is related to the equation of motion for the tetrad-frame energy E by

$$p \frac{dp}{d\lambda} = E \frac{dE}{d\lambda} . \quad (33.11)$$

The equation of motion for the tetrad-frame energy $E \equiv p^0$ is

$$\frac{dE}{d\lambda} = -\Gamma_{mn}^0 p^m p^n = \Gamma_{0a0} p^a E + \Gamma_{0ab} p^a p^b . \quad (33.12)$$

From this it follows that

$$\frac{d \ln p}{d\lambda} = \frac{E}{p^2} \frac{dE}{d\lambda} = E \left(\frac{E \hat{p}^a}{p} \Gamma_{0a0} + \hat{p}^a \hat{p}^b \Gamma_{0ab} \right) = E \left(-\frac{\dot{a}}{a^2} + \frac{E \hat{p}^a}{p} \Gamma_{0a0} + \hat{p}^a \hat{p}^b \hat{\Gamma}_{0ab} \right) , \quad (33.13)$$

where in the last expression the tetrad connection Γ_{0ab} , equation (29.24b), has been separated into its unperturbed and perturbed parts $-(\dot{a}/a^2)\delta_{ab}$ and $\hat{\Gamma}_{0ab}$.

In practice, the integration variable used to evolve equations is the conformal time η , not the affine parameter λ . The relation between conformal time η and affine parameter λ is

$$\frac{d\eta}{d\lambda} = p^\eta = e_m{}^\eta p^m = (\delta_m^n + \varphi_m{}^n) \hat{e}_n{}^\eta p^m = \frac{1}{a} [E(1 - \varphi_{00}) - p^a \varphi_{a0}] , \quad (33.14)$$

whose reciprocal is to linear order

$$\frac{d\lambda}{d\eta} = \frac{a}{E} \left(1 + \varphi_{00} + \frac{p^a}{E} \varphi_{a0} \right) . \quad (33.15)$$

With conformal time η as the integration variable, the equation of motion (33.13) for the magnitude p of the tetrad-frame momentum becomes, to linear order,

$$\frac{d \ln p}{d\eta} = -\frac{\dot{a}}{a} \left(1 + \varphi_{00} + \frac{p^a}{E} \varphi_{a0} \right) + \frac{E \hat{p}^a}{p} \Gamma_{0a0} + \hat{p}^a \hat{p}^b \hat{\Gamma}_{0ab} . \quad (33.16)$$

With the collision term restored, the Boltzmann equation (31.32) expressed with respect to conformal time η is

$$\boxed{\frac{df}{d\eta} = \frac{\partial f}{\partial \eta} + v^a \nabla_a f + \frac{d \ln p}{d\eta} \frac{\partial f}{\partial \ln p} = \frac{d\lambda}{d\eta} C[f]} , \quad (33.17)$$

where $v^a \equiv p^a/E$ is the tetrad-frame particle velocity, and $d\lambda/d\eta$ and $d \ln p/d\eta$ are given by equations (33.15)

and (33.16). Expressions for $d\lambda/d\eta$ and $d\ln p/d\eta$ in terms of the vierbein perturbations in a general gauge are left as Exercise 33.3. In conformal Newtonian gauge, the factor $d\lambda/d\eta$, equation (33.15), is

$$\frac{d\lambda}{d\eta} = \frac{a}{E}(1 + \Psi) . \quad (33.18)$$

In conformal Newtonian gauge, and including only scalar fluctuations, the factor $d\ln p/d\eta$, equation (33.16), is

$$\frac{d\ln p}{d\eta} = -\frac{\dot{a}}{a} + \dot{\Phi} - \frac{E\hat{p}^a}{p}\nabla_a\Psi . \quad (33.19)$$

To unperturbed order, the Boltzmann equation (33.17) is

$$\frac{d\overset{\circ}{f}}{d\eta} = \frac{\partial\overset{\circ}{f}}{\partial\eta} - \frac{\dot{a}}{a}\frac{\partial\overset{\circ}{f}}{\partial\ln p} = \frac{a}{E}C[\overset{\circ}{f}] , \quad (33.20)$$

where $C[\overset{\circ}{f}]$ is the unperturbed collision term, the factor a/E coming from $d\lambda/d\eta = a/E$ to unperturbed order, equation (33.15). The second term in the middle expression of equation (33.20) reflects the fact that the tetrad-frame momentum p redshifts as $p \propto 1/a$ as the Universe expands, a statement that is true for both massive and massless particles, equation (10.69).

Subtracting off the unperturbed part (33.20) of the Boltzmann equation (33.17) gives the perturbation of the Boltzmann equation

$$\boxed{\frac{d\overset{1}{f}}{d\eta} = \frac{\partial\overset{1}{f}}{\partial\eta} + v^a\nabla_a\overset{1}{f} - \frac{\dot{a}}{a}\frac{\partial\overset{1}{f}}{\partial\ln p} + G\frac{\partial\overset{\circ}{f}}{\partial\ln p} = \frac{a}{E}C[\overset{1}{f}] + \frac{d\overset{1}{\lambda}}{d\eta}C[\overset{\circ}{f}]} , \quad (33.21)$$

where $d\ln p/d\eta$ multiplying $\partial\overset{1}{f}/\partial\ln p$ has been replaced by $-\dot{a}/a$ to linear order, equation (33.16), and G (not to be confused with the Einstein tensor) defined by

$$G \equiv \frac{d\ln(ap)}{d\eta} \quad (33.22)$$

expresses the peculiar gravitational redshifting of particles. In conformal Newtonian gauge, and including only scalar fluctuations, the gravitational redshift term G is

$$G = \dot{\Phi} - \frac{E\hat{p}^a}{p}\nabla_a\Psi . \quad (33.23)$$

In conformal Newtonian gauge, the perturbed part of $d\lambda/d\eta$ which appears on the right hand side of the Boltzmann equation (33.21) is

$$\frac{d\overset{1}{\lambda}}{d\eta} = \frac{a}{E}\Psi . \quad (33.24)$$

Exercise 33.3 Boltzmann equation factors in a general gauge. Show that in a general gauge, and including not just scalar but also vector and tensor fluctuations, equation (33.15) is

$$\frac{d\lambda}{d\eta} = \frac{a}{E} \left[1 + \psi + \frac{p^a}{E} (\nabla_a \tilde{w} + \tilde{w}_a) \right] , \quad (33.25)$$

while equation (33.16) yields the gravitational redshift term

$$\begin{aligned} G \equiv \frac{d \ln(ap)}{d\eta} = \dot{\phi} + \frac{E \hat{p}^a}{p} \left[-\nabla_a \psi + \left(\frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \frac{m^2}{E^2} \right) (\nabla_a \tilde{w} + \tilde{w}_a) \right] \\ + \hat{p}^a \hat{p}^b \left[-\nabla_a \nabla_b (w - \dot{h}) - \frac{1}{2} (\nabla_a W_b + \nabla_b W_a) + \nabla_b \tilde{w}_a + \dot{h}_{ab} \right] . \end{aligned} \quad (33.26)$$

33.3 Non-baryonic cold dark matter

Non-baryonic cold dark matter is by assumption nonrelativistic and collisionless. The unperturbed mean density is $\bar{\rho}_c$, which evolves with cosmic scale factor a as

$$\bar{\rho}_c \propto a^{-3} . \quad (33.27)$$

Since dark matter particles are nonrelativistic, the energy of a dark matter particle is its rest-mass energy, $E_c = m_c$, and its momentum is the nonrelativistic momentum $p_c^a = m_c v_c^a$.

The energy-momentum tensor T_c^{mn} of the dark matter is obtained from integrals over the dark matter phase-space distribution f_c , equation (10.122). The energy and momentum moments of the distribution define the dark matter overdensity δ_c and bulk velocity v_c , while the pressure is of order v_c^2 , and can be neglected to linear order (note the different fonts for particle velocity v and bulk velocity v),

$$T_c^{00} \equiv \int f_c m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \equiv \bar{\rho}_c (1 + \delta_c) , \quad (33.28a)$$

$$T_c^{0a} \equiv \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \equiv \bar{\rho}_c v_c^a , \quad (33.28b)$$

$$T_c^{ab} \equiv \int f_c m_c v_c^a v_c^b \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = 0 . \quad (33.28c)$$

Non-baryonic cold dark matter is collisionless, so the collision term in the Boltzmann equation is zero, $C[f_c] = 0$, and the dark matter satisfies the collisionless Boltzmann equation

$$\frac{df_c}{d\eta} = 0 . \quad (33.29)$$

The energy and momentum moments of the Boltzmann equation (33.17) yield equations for the overdensity

δ_c and bulk velocity v_c , which in the conformal Newtonian gauge are

$$\begin{aligned} 0 &= \int \frac{df_c}{d\eta} m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = \frac{\partial}{\partial\eta} \int f_c m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} + \nabla_a \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} - \int \left(\frac{\dot{a}}{a} - \dot{\Phi} \right) \frac{\partial f}{\partial \ln p} m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &= \frac{\partial \bar{\rho}_c (1 + \delta_c)}{\partial\eta} + \nabla_a (\bar{\rho}_c v_c^a) + 3 \left(\frac{\dot{a}}{a} - \dot{\Phi} \right) \bar{\rho}_c , \end{aligned} \quad (33.30a)$$

$$\begin{aligned} 0 &= \int \frac{df_c}{d\eta} m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = \frac{\partial}{\partial\eta} \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} + \nabla_b \int f_c m_c v_c^a v_c^b \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &\quad - \int \left(\frac{\dot{a}}{a} - \dot{\Phi} + \frac{E \hat{p}^b}{p} \nabla_b \Psi \right) \frac{\partial f}{\partial \ln p} m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &= \frac{\partial \bar{\rho}_c v_c^a}{\partial\eta} + 4 \left(\frac{\dot{a}}{a} - \dot{\Phi} \right) \bar{\rho}_c v_c^a + \bar{\rho}_c \nabla_a \Psi . \end{aligned} \quad (33.30b)$$

The $\dot{\Phi} \bar{\rho}_c v_c^a$ term on the last line of equation (33.30b) can be dropped, since the potential Φ and the bulk velocity v_c^a are both of first order, so their product is of second order. Subtracting the unperturbed part from equations (33.30a) and (33.30b) gives equations for the dark matter overdensity δ_c and bulk velocity $\mathbf{v}_c$,

$$\dot{\delta}_c + \nabla \cdot \mathbf{v}_c - 3\dot{\Phi} = 0 , \quad (33.31a)$$

$$\dot{\mathbf{v}}_c + \frac{\dot{a}}{a} \mathbf{v}_c + \nabla \Psi = 0 . \quad (33.31b)$$

Transformed into Fourier space, and decomposed into scalar v_c and vector $\mathbf{v}_{c,\perp}$ parts, the velocity 3-vector $\mathbf{v}_c$ is

$$\mathbf{v}_c = -i\hat{\mathbf{k}} v_c + \mathbf{v}_{c,\perp} . \quad (33.32)$$

For the scalar modes under consideration, only the scalar part of the dark matter equations (33.31) is relevant:

$$\dot{\delta}_c - k v_c - 3\dot{\Phi} = 0 , \quad (33.33a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k \Psi = 0 . \quad (33.33b)$$

Equations (33.33) reproduce the equations (30.52) derived previously from conservation of energy and momentum.

Exercise 33.4 Moments of the non-baryonic cold dark matter Boltzmann equation. Confirm equations (33.30).

33.4 Boltzmann equation for the temperature fluctuation

The full hierarchy of Boltzmann equations is needed only to describe relativistic species (photons and neutrinos); nonrelativistic species (non-baryonic dark matter and baryons) are described adequately by the equations of conservation of energy and momentum. Cosmological fluctuations in relativistic species are commonly characterized in terms of a temperature fluctuation Θ ,

$$\Theta(\eta, \mathbf{x}, \mathbf{p}) \equiv \frac{\delta T(\eta, \mathbf{x}, \mathbf{p})}{T(\eta)}. \quad (33.34)$$

In most of this book Θ refers to photons, but in this section the temperature fluctuation Θ refers to any species, bosonic or fermionic, massless or massive (neutrinos have small masses, §10.25). At early times, collisions drove the occupation number f into thermodynamic equilibrium at each comoving position $\mathbf{x}$, so that initially the temperature fluctuation was a function $\Theta(\eta, \mathbf{x})$ only of time and position, not of particle momentum $\mathbf{p}$. This explains why the preferred fluctuation variable is the temperature fluctuation Θ , and not the perturbation δf of the occupation number; the latter depends on momentum p even in thermodynamic equilibrium. As collisions peter out, around recombination in the case of photons, and around $e\bar{e}$ -annihilation in the case of neutrinos, free-streaming allows the temperature fluctuation Θ to become anisotropic.

For a relativistic species, the unperturbed occupation number in thermodynamic equilibrium is

$$\overset{\circ}{f} = \frac{1}{e^{p/T} \mp 1}, \quad (33.35)$$

where the sign is $-$ for bosons (photons) and $+$ for fermions (neutrinos). The unperturbed Boltzmann equation (33.20) can be recast as an equation for the background temperature $T(\eta)$,

$$\frac{d \ln(aT)}{d\eta} = \frac{a}{E} C[\overset{\circ}{f}] / \frac{\partial \overset{\circ}{f}}{\partial \ln T}, \quad (33.36)$$

where it follows from equation (33.35) that (the partial derivative with respect to temperature $\partial/\partial \ln T$ is at constant momentum p)

$$\frac{\partial \overset{\circ}{f}}{\partial \ln T} = \overset{\circ}{f} (1 \pm \overset{\circ}{f}) \frac{p}{T}, \quad (33.37)$$

with $+$ and $-$ for bosons and fermions respectively. Equation (33.36) shows that if the collision term $C[\overset{\circ}{f}]$ vanishes, then the background temperature redshifts as $T \propto a^{-1}$. In practice, the collision term $C[\overset{\circ}{f}]$ was negligible for both photons and neutrinos since the end of $e\bar{e}$ -annihilation. Photons continued to exchange energy with electrons and baryons, but the effect on the photons was negligible because they overwhelmingly outnumbered electrons and baryons, equation (10.104). Although the heating term $d \ln(aT)/d\eta$ is negligible in the situation at hand, it is retained temporarily for completeness in the next paragraph.

The definition (33.34) of the temperature fluctuation Θ is to be interpreted as meaning that the perturbation to the occupation number is

$$\delta f = \frac{\partial \overset{\circ}{f}}{\partial \ln T} \delta \ln T = \frac{\partial \overset{\circ}{f}}{\partial \ln T} \Theta. \quad (33.38)$$

Two of the terms on the left hand side of the perturbed Boltzmann equation (33.21) rearrange to

$$\begin{aligned} \frac{\partial \overset{1}{f}}{\partial \eta} + \frac{d \ln p}{d \eta} \frac{\partial \overset{1}{f}}{\partial \ln p} &= \frac{\partial \overset{0}{f}}{\partial \ln T} \dot{\Theta} + \left[\left(\frac{d \ln T}{d \eta} \frac{\partial}{\partial \ln T} - \frac{\dot{a}}{a} \frac{\partial}{\partial \ln p} \right) \frac{\partial \overset{0}{f}}{\partial \ln T} \right] \Theta \\ &= \frac{\partial \overset{0}{f}}{\partial \ln T} \left[\dot{\Theta} + \frac{d \ln(aT)}{d \eta} \frac{\partial \ln(\partial \overset{0}{f} / \partial \ln T)}{\partial \ln T} \Theta \right]. \end{aligned} \quad (33.39)$$

The collision terms on the right hand side of the perturbed Boltzmann equation (33.21) are

$$\frac{a}{E} C[\overset{1}{f}] + \frac{d \lambda}{d \eta} C[\overset{0}{f}] = \frac{\partial \overset{0}{f}}{\partial \ln T} \left[C[\Theta] + \frac{d \ln(aT)}{d \eta} \frac{E}{a} \frac{d \lambda}{d \eta} \right], \quad (33.40)$$

where $C[\overset{0}{f}]$ has been eliminated in favour of $d \ln(aT)/d \eta$ using equation (33.36), and $C[\Theta]$ is the scaled collision term defined by

$$C[\Theta] \equiv \frac{a}{E} C[\overset{1}{f}] / \frac{\partial \overset{0}{f}}{\partial \ln T}. \quad (33.41)$$

The perturbed Boltzmann equation (33.21) thus becomes

$$\frac{d \Theta}{d \eta} = \frac{\partial \Theta}{\partial \eta} + v^a \nabla_a \Theta - G + \frac{d \ln(aT)}{d \eta} \frac{\partial \ln(\partial \overset{0}{f} / \partial \ln T)}{\partial \ln T} \Theta = C[\Theta] + \frac{d \ln(aT)}{d \eta} \frac{E}{a} \frac{d \lambda}{d \eta}, \quad (33.42)$$

where the gravitational redshift term G gets a minus sign from $\partial \overset{0}{f} / \partial \ln p = -\partial \overset{0}{f} / \partial \ln T$.

In practice the heating terms proportional to $d \ln(aT)/d \eta$, though important during for example electron-positron annihilation, are negligible for both photons and neutrinos during the time before and through recombination when anisotropies in the CMB are developing. The Boltzmann equation (33.42) then reduces to

$$\frac{d \Theta}{d \eta} = \frac{\partial \Theta}{\partial \eta} + v^a \nabla_a \Theta - G = C[\Theta]. \quad (33.43)$$

As long as the particles are relativistic, the particle velocity is one, $v = 1$; but equation (33.43) allows for a general non-unit velocity v to accommodate neutrinos, which have small masses.

Fourier transforming the Boltzmann equation (33.43) over spatial position $\mathbf{x}$ yields the Boltzmann equation for the Fourier components $\Theta(\eta, \mathbf{k}, \mathbf{p})$ of the temperature fluctuation,

$$\boxed{\frac{d \Theta}{d \eta} = \frac{\partial \Theta}{\partial \eta} - i v k \mu \Theta - G = C[\Theta]}, \quad (33.44)$$

where μ is the cosine of the angle between the wavevector $\mathbf{k}$ and the photon momentum $\mathbf{p}$,

$$\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}. \quad (33.45)$$

In Fourier space, the gravitational redshift term G , in conformal Newtonian gauge and including only scalar fluctuations, is, equation (33.23),

$$G = \dot{\Phi} + \frac{i k \mu}{v} \Psi. \quad (33.46)$$

33.5 Spherical harmonics of the temperature fluctuation

It is natural to expand the (photon or neutrino) temperature fluctuation Θ in spherical harmonics. The various components of the energy-momentum tensor T^{mn} are determined by the monopole, dipole, and quadrupole harmonics of the particle distribution. Scalar fluctuations are those that are rotationally symmetric about the wavevector direction $\hat{\mathbf{k}}$, which correspond to spherical harmonics with zero azimuthal quantum number, $m = 0$. Expanded in spherical harmonics $Y_{\ell m}(\hat{\mathbf{p}})$, and with only scalar terms retained, the temperature fluctuation Θ can be written

$$\begin{aligned}\Theta(\eta, \mathbf{k}, \mathbf{p}) &= \sum_{\ell=0}^{\infty} (-i)^{\ell} \sqrt{4\pi(2\ell+1)} \Theta_{\ell}(\eta, \mathbf{k}, p) Y_{\ell 0}(\hat{\mathbf{p}}) \\ &= \sum_{\ell=0}^{\infty} (-i)^{\ell} (2\ell+1) \Theta_{\ell}(\eta, \mathbf{k}, p) P_{\ell}(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}}) ,\end{aligned}\quad (33.47)$$

where P_{ℓ} are Legendre polynomials, §33.14. The choice of normalization of the scalar harmonics Θ_{ℓ} is *not* the same as the traditional normalization $\Theta = \sum_{\ell=0}^{\infty} \Theta_{\ell 0} Y_{\ell 0}$, but is conventional in studies of the CMB. The factor of $(-i)^{\ell}$ makes Θ_{ℓ} real, and the normalization factor removes square root factors in the Boltzmann hierarchy. The harmonics in the traditional and CMB conventions are related by $\Theta_{\ell 0} = (-i)^{\ell} \sqrt{4\pi(2\ell+1)} \Theta_{\ell}$. The scalar harmonics Θ_{ℓ} are angular integrals of the temperature fluctuation Θ over momentum directions $\hat{\mathbf{p}}$,

$$\Theta_{\ell}(\eta, \mathbf{k}, p) = i^{\ell} \int \Theta(\eta, \mathbf{k}, \mathbf{p}) P_{\ell}(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}}) \frac{d\mathbf{p}}{4\pi} . \quad (33.48)$$

Expanded into the scalar harmonics $\Theta_{\ell}(\eta, \mathbf{k}, p)$, equation (33.47), the left hand side of the Boltzmann equation (33.44) in conformal Newtonian gauge is

$$\frac{d\Theta_0}{d\eta} = \dot{\Theta}_0 - vk\Theta_1 - \dot{\Phi} , \quad (33.49a)$$

$$\frac{d\Theta_1}{d\eta} = \dot{\Theta}_1 + \frac{vk}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3v} \Psi , \quad (33.49b)$$

$$\frac{d\Theta_{\ell}}{d\eta} = \dot{\Theta}_{\ell} + \frac{vk}{2\ell+1} [\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] \quad (\ell \geq 2) . \quad (33.49c)$$

33.6 The Boltzmann equation for massless particles

The Boltzmann equation (33.44) and its harmonic expansion (33.49) are valid for massive as well as massless particles, to allow for neutrino masses. The case of massive neutrinos will be resumed in §33.13; but for the next several sections, particles (photons and neutrinos) will be taken to be massless.

Photons are massless, and neutrinos (probably) have small enough masses that they can be treated as massless through recombination. The velocities of massless particles are always one, $v = 1$. For massless

particles, the Boltzmann equation for the temperature fluctuation Θ is equation (33.43) with $v = 1$. The left hand side of the Boltzmann equation expands in scalar harmonics Θ_ℓ as equations (33.49) again with $v = 1$.

As remarked at the beginning of §33.4, at early times photons and neutrinos have distributions in thermodynamic equilibrium, as a result of which their initial temperature fluctuations Θ are independent of particle momentum $\mathbf{p}$. For massless particles ($v = 1$) the left hand side of the Boltzmann equation (33.43) is independent of the magnitude p of the particle momentum. As will be seen in §33.8, equation (33.68), Thomson scattering leaves the magnitude p_γ of the photon momentum essentially unchanged. Consequently the temperature fluctuations $\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}})$ of photons, and of neutrinos as long as they are relativistic, depend on the direction $\hat{\mathbf{p}}$ but not magnitude p of the particle momentum,

$$\Theta(\eta, \mathbf{k}, \mathbf{p}) = \Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}) \quad \text{for photons and relativistic neutrinos.} \quad (33.50)$$

33.7 Energy-momentum tensor for massless particles

Perturbations $\overset{1}{T}{}^{kl}$ to the energy-momentum tensor of particles involve integrals (10.122) over the perturbed occupation number $\overset{1}{f}$. For massless particles, these integrals take the form, where $F(\hat{\mathbf{p}})$ is some arbitrary function of the momentum direction $\hat{\mathbf{p}}$,

$$\int \overset{1}{f} p^2 F(\hat{\mathbf{p}}) \frac{g d^3 p}{p(2\pi\hbar)^3} = \int \frac{\partial \overset{0}{f}}{\partial \ln T} p^2 \frac{g 4\pi p^2 dp}{p(2\pi\hbar)^3} \int \Theta F(\hat{\mathbf{p}}) \frac{d\mathbf{o}_{\mathbf{p}}}{4\pi} = 4\bar{\rho} \int \Theta F(\hat{\mathbf{p}}) \frac{d\mathbf{o}_{\mathbf{p}}}{4\pi}, \quad (33.51)$$

in which the last expression is true because

$$\int \frac{\partial \overset{0}{f}}{\partial \ln T} p^2 \frac{g 4\pi p^2 dp}{p(2\pi\hbar)^3} = 4 \int \overset{0}{f} p \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = 4\bar{\rho}, \quad (33.52)$$

which follows from $\partial \overset{0}{f} / \partial \ln T = -\partial \overset{0}{f} / \partial \ln p$ and an integration by parts. The perturbation of the energy density, energy flux, monopole pressure, and quadrupole pressure of massless particles are then, with integrals over Θ converted to harmonics Θ_ℓ using equations (33.48),

$$\overset{1}{T}{}^{00} = 4\bar{\rho} \Theta_0, \quad (33.53a)$$

$$\hat{k}_a \overset{1}{T}{}^{0a} = -i 4\bar{\rho} \Theta_1, \quad (33.53b)$$

$$\frac{1}{3} \delta_{ab} \overset{1}{T}{}^{ab} = \frac{4}{3} \bar{\rho} \Theta_0, \quad (33.53c)$$

$$\left(\frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) \overset{1}{T}{}^{ab} = -4\bar{\rho} \Theta_2. \quad (33.53d)$$

33.8 Nonrelativistic electron-photon (Thomson) scattering

The dominant process that couples photons and baryons is electron-photon scattering

$$e + \gamma \leftrightarrow e' + \gamma'. \quad (33.54)$$

The Lorentz-invariant mean amplitude squared for unpolarized nonrelativistic electron-photon (Thomson) scattering from initial photon momentum $\mathbf{p}_\gamma$ to final momentum $\mathbf{p}_{\gamma'}$ of the same magnitude, $p_{\gamma'} = p_\gamma$, is, in units $c = \hbar = 1$, equation (47.41),

$$\langle |\mathcal{M}|^2 \rangle = (8\pi\alpha)^2 \frac{1 + (\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})^2}{2}, \quad (33.55)$$

where $\alpha \equiv e^2/(\hbar c)$ is the fine-structure constant. The unpolarized mean amplitude squared (33.55) is the polarized amplitude squared (35.54) averaged over polarization states of the incoming photon (that's what the adjective mean refers to here), and summed over polarization states of the scattered photon. The differential cross-section $d\sigma_T/d\Omega'$ for unpolarized Thomson scattering into an interval $d\Omega'$ of solid angle about scattered photon direction $\hat{\mathbf{p}}'$ is related to the squared amplitude $\langle |\mathcal{M}|^2 \rangle$ by, equation (47.42),

$$\frac{d\sigma_T}{d\Omega'} = \frac{\langle |\mathcal{M}|^2 \rangle}{(8\pi m_e)^2} = \frac{\alpha^2}{m_e^2} \frac{1 + (\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})^2}{2}. \quad (33.56)$$

The coefficient of the differential cross-section is, with units c and $\hbar$ restored,

$$\left(\frac{\alpha \hbar}{m_e c} \right)^2 = r_e^2 = \left(\frac{e^2}{m_e c^2} \right)^2, \quad (33.57)$$

where $r_e \equiv e^2/m_e c^2$ is the classical electron radius. The total Thomson cross-section σ_T is

$$\sigma_T \equiv \int \frac{d\sigma_T}{d\Omega'} d\Omega' = \frac{8\pi}{3} r_e^2. \quad (33.58)$$

33.9 The photon collision term for electron-photon scattering

Electron-photon scattering keeps electrons and photons close to mutual thermodynamic equilibrium, and their unperturbed distributions can be taken to be in thermodynamic equilibrium. The unperturbed photon collision term for electron-photon scattering therefore vanishes, because of detailed balance, Exercise 31.5,

$$C[f_\gamma^0] = 0. \quad (33.59)$$

Thanks to detailed balance, the combination of rates in the collision integral (31.40) almost cancels, so can be treated as being of linear order in perturbation theory. This allows other factors in the collision integral to be approximated by their unperturbed values.

The photon collision term for electron-photon scattering follows from the general expression (31.40). To unperturbed order, the energies of the electrons, which are nonrelativistic, may be set equal to their rest masses, $E_e = m_e$. Since photons are massless, their energies are just equal to their momenta, $E_\gamma = p_\gamma$. The electron occupation number is small, $f_e \ll 1$, so the Fermi blocking factors for electrons may be neglected, $1 - f_e = 1$. These considerations bring the photon collision term for electron-photon scattering to, from the

general expression (31.40),

$$C[\dot{f}_\gamma] = \frac{1}{16} \int \langle |\mathcal{M}|^2 \rangle [-f_e f_\gamma (1 + f_{\gamma'}) + f_{e'} f_{\gamma'} (1 + f_\gamma)] (2\pi)^4 \delta_D^4(p_e + p_\gamma - p_{e'} - p_{\gamma'}) \frac{2 d^3 p_e}{m_e (2\pi)^3} \frac{d^3 p_{e'}}{m_e (2\pi)^3} \frac{d^3 p_{\gamma'}}{p_{\gamma'} (2\pi)^3} . \quad (33.60)$$

The various integrations over momenta are most conveniently carried out as follows. The energy-conserving integral is best done over the energy of the scattered photon γ' , which is scattered into an interval $do_{\gamma'}$ of solid angle:

$$\int 2\pi \delta_D(E_e + E_\gamma - E_{e'} - E_{\gamma'}) \frac{d^3 p_{\gamma'}}{E_{\gamma'} (2\pi)^3} = p_{\gamma'} \frac{do_{\gamma'}}{(2\pi)^2} \approx p_\gamma \frac{do_{\gamma'}}{(2\pi)^2} . \quad (33.61)$$

The approximation in the last step of equation (33.61), replacing the energy $p_{\gamma'}$ of the scattered photon by the energy p_γ of the incoming photon, is valid because, thanks to the smallness of the combination of rates in the collision integral (33.60), it suffices to treat the photon energy to unperturbed order. As seen below, equation (33.66), the energy difference $p_\gamma - p_{\gamma'}$ between the incoming and scattered photons is of linear order in electron velocities.

The momentum-conserving integral is best done over the momentum of the scattered electron, which is e' for outgoing scatterings $e + \gamma \rightarrow e' + \gamma'$, and e for incoming scatterings $e + \gamma \leftarrow e' + \gamma'$. In the former case ($e + \gamma \rightarrow e' + \gamma'$),

$$\int (2\pi)^3 \delta_D^3(\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{e'} - \mathbf{p}_{\gamma'}) \frac{d^3 p_{e'}}{m_e (2\pi)^3} = \frac{1}{m_e} , \quad (33.62)$$

and the result is the same, $1/m_e$, in the latter case ($e + \gamma \leftarrow e' + \gamma'$). The energy- and momentum-conserving integrals having been done, the electron e' in the latter case may be relabelled e . So relabelled, the combination of rate factors in the collision integral (33.60) becomes

$$-f_e f_\gamma (1 + f_{\gamma'}) + f_e f_{\gamma'} (1 + f_\gamma) = f_e (-f_\gamma + f_{\gamma'}) . \quad (33.63)$$

Notice that the stimulated ($f_e f_\gamma f_{\gamma'}$) terms cancel. The energy- and momentum-conserving integrations (33.61) and (33.62) bring the photon collision term (33.60) to

$$C[\dot{f}_\gamma] = \frac{p_\gamma}{16\pi m_e^2} \int \langle |\mathcal{M}|^2 \rangle f_e (-f_\gamma + f_{\gamma'}) \frac{2 d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi} . \quad (33.64)$$

The collision integral (33.64) involves the difference $-f_\gamma + f_{\gamma'}$ between the occupancy of the initial and final photon states. To linear order, the difference is

$$-f_\gamma + f_{\gamma'} = -\overset{0}{f}(\mathbf{p}_\gamma) + \overset{0}{f}(\mathbf{p}_{\gamma'}) - \overset{1}{f}(\mathbf{p}_\gamma) + \overset{1}{f}(\mathbf{p}_{\gamma'}) = \frac{\partial \overset{0}{f}_\gamma}{\partial \ln T} \left[\frac{p_\gamma - p_{\gamma'}}{p_\gamma} - \Theta(\mathbf{p}_\gamma) + \Theta(\mathbf{p}_{\gamma'}) \right] . \quad (33.65)$$

The first term $(p_\gamma - p_{\gamma'})/p_\gamma$ arises because the incoming and scattered photon energies differ slightly. The

difference in photon energies is given by energy conservation:

$$\begin{aligned}
p_\gamma - p_{\gamma'} &= E_{e'} - E_e \\
&= \left(m_e + \frac{p_e'^2}{2m_e} \right) - \left(m_e + \frac{p_e^2}{2m_e} \right) \\
&= \frac{(\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{\gamma'})^2 - p_e^2}{2m_e} \\
&= \frac{(\mathbf{p}_\gamma - \mathbf{p}_{\gamma'}) \cdot (2\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{\gamma'})}{2m_e} \\
&\approx (\mathbf{p}_\gamma - \mathbf{p}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} ,
\end{aligned} \tag{33.66}$$

the last line of which follows because the photon momentum is small compared to the electron momentum,

$$p_\gamma \sim T \sim \frac{p_e^2}{m_e} \ll p_e . \tag{33.67}$$

Because the photon energy difference is of first order, and the temperature fluctuation is already of first order, it suffices to regard the temperature fluctuation Θ as being a function only of the direction $\hat{\mathbf{p}}_\gamma$ of the photon momentum, not of its energy:

$$\Theta(\mathbf{p}_\gamma) \approx \Theta(\hat{\mathbf{p}}_\gamma) . \tag{33.68}$$

The linear approximations (33.66) and (33.68) bring the difference (33.65) between the initial and final photon occupancies to

$$-f_\gamma + f_{\gamma'} = \frac{\partial f_\gamma^0}{\partial \ln T} \left[(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right] . \tag{33.69}$$

Inserting this difference in occupancies into the collision integral (33.64) yields

$$C[f_\gamma] = \frac{p_\gamma}{16\pi m_e^2} \frac{\partial f_\gamma^0}{\partial \ln T} \int \langle |\mathcal{M}|^2 \rangle f_e \left[(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right] \frac{2 d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi} , \tag{33.70}$$

or, switching to $C[\Theta]$ defined by equation (33.41),

$$C[\Theta] = \frac{a}{16\pi m_e^2} \int \langle |\mathcal{M}|^2 \rangle f_e \left[(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right] \frac{2 d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi} . \tag{33.71}$$

The amplitude squared $\langle |\mathcal{M}|^2 \rangle$, equation (33.55), is independent of electron momenta, so the integration over electron momentum in the collision integral (33.71) is straightforward. The unperturbed electron density $\bar{n}_e$ and the electron bulk velocity $\mathbf{v}_e$ are defined by

$$\bar{n}_e \equiv \int f_e^0 \frac{2 d^3 p_e}{(2\pi)^3} , \quad \bar{n}_e \mathbf{v}_e \equiv \int f_e^0 \frac{\mathbf{p}_e}{m_e} \frac{2 d^3 p_e}{(2\pi)^3} . \tag{33.72}$$

Coulomb scattering keeps electrons and ions tightly coupled, so the electron bulk velocity $\mathbf{v}_e$ equals the baryon bulk velocity $\mathbf{v}_b$,

$$\mathbf{v}_e = \mathbf{v}_b . \tag{33.73}$$

Integration over the electron momentum brings the collision integral (33.71) to

$$C[\Theta] = \frac{\bar{n}_e a}{16\pi m_e^2} \int \langle |\mathcal{M}|^2 \rangle [(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'})] \frac{do_{\gamma'}}{4\pi} . \quad (33.74)$$

Finally, the collision integral (33.74) must be integrated over the direction $\hat{\mathbf{p}}_{\gamma'}$ of the scattered photon. The integration is facilitated if the angular dependence of the amplitude squared $\langle |\mathcal{M}|^2 \rangle$ given by equation (33.55) is expanded in Legendre polynomials, $\frac{1}{2}(1 + \mu^2) = \frac{2}{3} [1 + \frac{1}{2}P_2(\mu)]$. Inserting the amplitude squared $\langle |\mathcal{M}|^2 \rangle$ into the collision integral (33.74) brings it to

$$C[\Theta] = |\dot{\tau}| \int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] [(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'})] \frac{do_{\gamma'}}{4\pi} , \quad (33.75)$$

where $\dot{\tau} \equiv -\bar{n}_e \sigma_T a$ is the scattering rate (32.4). Equation (33.75) (unlike equation (33.74)) remains dimensionally correct even when units of c and $\hbar$ are restored (both sides have units $1/\eta$). The $\hat{\mathbf{p}}_{\gamma'} \cdot \mathbf{v}_b$ term in the integrand of (33.75) is odd, and vanishes on angular integration:

$$\int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] \hat{\mathbf{p}}_{\gamma'} \frac{do_{\gamma'}}{4\pi} = 0 . \quad (33.76)$$

Similarly, the angular integral over the quadrupole of quantities independent of $\hat{\mathbf{p}}_{\gamma'}$ vanishes:

$$\int P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'}) [\hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma)] \frac{do_{\gamma'}}{4\pi} = 0 . \quad (33.77)$$

The collision integral (33.75) thus reduces to

$$C[\Theta(\mathbf{x}, \hat{\mathbf{p}}_\gamma)] = |\dot{\tau}| \left\{ \hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b(\mathbf{x}) - \Theta(\mathbf{x}, \hat{\mathbf{p}}_\gamma) + \int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] \Theta(\mathbf{x}, \hat{\mathbf{p}}_{\gamma'}) \frac{do_{\gamma'}}{4\pi} \right\} , \quad (33.78)$$

where the dependence of various quantities on comoving position $\mathbf{x}$ has been made explicit. Now transform to Fourier space (in effect, replace comoving position $\mathbf{x}$ by comoving wavevector $\mathbf{k}$). Replace the baryon bulk velocity by its scalar part, $\mathbf{v}_b = i\hat{\mathbf{k}}v_b$. To perform the remaining angular integral over the photon direction $\hat{\mathbf{p}}_{\gamma'}$, expand the Legendre polynomial $P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})$ in the integrand in spherical harmonics using the addition theorem (33.103), expand the temperature fluctuation $\Theta(\mathbf{k}, \hat{\mathbf{p}}_{\gamma'})$ in scalar multipole moments according to equation (33.47), and invoke orthogonality of the spherical harmonics. With $\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma$, these manipulations bring the photon collision integral (33.78) at last to

$$C[\Theta(\mathbf{k}, \hat{\mathbf{p}}_\gamma)] = |\dot{\tau}| [-i\mu v_b(\mathbf{k}) - \Theta(\mathbf{k}, \mu) + \Theta_0(\mathbf{k}) - \frac{1}{2}\Theta_2(\mathbf{k})P_2(\mu)] . \quad (33.79)$$

33.10 Boltzmann equation for photons

Inserting the collision term (33.79) into equation (33.44) with unit velocity, $v = 1$, yields the Boltzmann equation for the photon temperature fluctuation $\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}_\gamma)$, for scalar fluctuations in conformal Newtonian

gauge,

$$\boxed{\frac{d\Theta}{d\eta} = \frac{\partial\Theta}{\partial\eta} - ik\mu\Theta - \dot{\Phi} - ik\mu\Psi = |\dot{\tau}| \left[-i\mu v_b - \Theta + \Theta_0 - \frac{1}{2}\Theta_2 P_2(\mu) \right]} . \quad (33.80)$$

Expanded into the scalar harmonics $\Theta_\ell(\eta, \mathbf{k})$, the photon Boltzmann equation (33.80) yields the hierarchy of photon multipole equations

$$\dot{\Theta}_0 - k\Theta_1 - \dot{\Phi} = 0 , \quad (33.81a)$$

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi = \frac{1}{3}|\dot{\tau}|(v_b - 3\Theta_1) , \quad (33.81b)$$

$$\dot{\Theta}_2 + \frac{k}{5}(2\Theta_1 - 3\Theta_3) = -\frac{9}{10}|\dot{\tau}|\Theta_2 , \quad (33.81c)$$

$$\dot{\Theta}_\ell + \frac{k}{2\ell+1}[\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] = -|\dot{\tau}|\Theta_\ell \quad (\ell \geq 3) . \quad (33.81d)$$

When polarization is included, the factor $\frac{9}{10}$ on the right hand side of equation (33.81c) is decreased by a factor $\frac{5}{6}$ to $\frac{3}{4}$, Exercise 35.7,

$$\frac{9}{10} \rightarrow \frac{3}{4} . \quad (33.82)$$

The Boltzmann hierarchy (33.81) shows that all the photon multipoles except the photon monopole Θ_0 are affected by electron-photon scattering, but only the photon dipole Θ_1 depends directly on one of the baryon variables, the baryon bulk velocity v_b . The dependence on the baryon velocity v_b reflects the fact that, to linear order, there is a transfer of momentum between photons and baryons, but no transfer of number or of energy.

33.10.1 Truncating the photon Boltzmann hierarchy

Photons are tightly coupled to baryons by scattering well before recombination, and stream freely well after recombination. The two regimes are sufficiently different to require different truncations of the Boltzmann hierarchy.

As argued in §32.7, prior to recombination, when $|\dot{\tau}|$ is large, scattering keeps successive multipoles smaller by factors of $k/|\dot{\tau}|$, equation (32.66). Keeping only the dominant $\Theta_{\ell-1}$ term on the left hand side of the Boltzmann hierarchy (33.81) for $\ell \geq 2$ implies

$$\Theta_\ell \approx -\frac{(1 + \frac{1}{9}\delta_{\ell 2})\ell k}{(2\ell+1)|\dot{\tau}|}\Theta_{\ell-1} \quad \text{for } \ell \geq 2 . \quad (33.83)$$

When polarization is included, the factor $1 + \frac{1}{9}\delta_{\ell 2} = \frac{10}{9}$ for $\ell = 2$ on the right hand side of equation (33.83) is changed to $1 + \frac{1}{3}\delta_{\ell 2} = \frac{4}{3}$ for $\ell = 2$, Exercise 35.7.

After recombination, photons stream freely, allowing the photon distribution to develop higher order

multipoles comparable to lower orders. A better approximation in the free-streaming regime is the same as that for neutrinos, equation (33.92),

$$\Theta_\ell \approx -(\Theta_{\ell-2} + \delta_{\ell 2} \Psi) - \frac{2\ell-1}{k\eta} \Theta_{\ell-1} \quad \text{for } \ell \geq 2. \quad (33.84)$$

The truncation of the photon Boltzmann hierarchy adopted in equations (33.4) is an interpolation between the scattering and free-streaming regimes (33.83) and (33.84).

33.11 Baryons

The equations governing baryonic matter are similar to those governing non-baryonic cold dark matter, §33.3, except that baryons are collisional. Coulomb scattering between electrons and ions keep baryons tightly coupled to each other. Electron-photon scattering then couples baryons to photons.

Since the unperturbed distribution of baryons is in thermodynamic equilibrium, the unperturbed collision term vanishes for each species of baryonic matter, as it did for photons, equation (33.59),

$$C[f_b] = 0. \quad (33.85)$$

For the perturbed baryon distribution, only the first and second moments of the phase-space distribution are important, since these govern the baryon overdensity δ_b and bulk velocity $\mathbf{v}_b$. The relevant collision term is the electron collision term associated with electron-photon scattering. Since electron-photon scattering neither creates nor destroys electrons, the zeroth moment of the electron collision term vanishes,

$$\int C[f_e] \frac{2d^3p_e}{m_e(2\pi)^3} = 0. \quad (33.86)$$

The first moment of the electron collision term is most easily determined from the fact that electron-photon collisions must conserve the total momentum of electron and photons:

$$\int C[f_e] m_e \mathbf{v}_e \frac{2d^3p_e}{m_e(2\pi)^3} + \int C[f_\gamma] \mathbf{p}_\gamma \frac{2d^3p_\gamma}{p_\gamma(2\pi)^3} = 0. \quad (33.87)$$

Substituting the expression (33.78) for the photon collision integral into equation (33.87), separating out factors depending on the magnitude p_γ and direction $\hat{\mathbf{p}}_\gamma$ of the photon momentum, and taking into consideration that the integral terms in equation (33.78), when multiplied by $\hat{\mathbf{p}}_\gamma$, are odd in $\hat{\mathbf{p}}_\gamma$, and therefore vanish on integration over directions $\hat{\mathbf{p}}_\gamma$, gives

$$\int C[f_e] m_e \mathbf{v}_e \frac{2d^3p_e}{m_e(2\pi)^3} = \bar{n}_e \sigma_T a \int \frac{\partial f_\gamma}{\partial \ln T} p_\gamma^2 \frac{24\pi p_\gamma^2 dp_\gamma}{p_\gamma(2\pi)^3} \int [-\hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b + \Theta(\hat{\mathbf{p}}_\gamma)] \hat{\mathbf{p}}_\gamma \frac{d\Omega_\gamma}{4\pi}. \quad (33.88)$$

The integral over the magnitude p_γ of the photon momentum in equation (33.88) yields $4\bar{\rho}$, in accordance with equation (33.52). Transformed into Fourier space, and with only scalar terms retained, the collision

integral (33.88) becomes, with $\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma$,

$$\begin{aligned} \hat{\mathbf{k}} \cdot \int C[f_e] m_e \mathbf{v}_e \frac{2 d^3 p_e}{m_e (2\pi)^3} &= 4 \bar{\rho}_\gamma \bar{n}_e \sigma_T a \int [i\mu v_b + \Theta] \mu \frac{do_\gamma}{4\pi} \\ &= \frac{4}{3} i \bar{\rho}_\gamma \bar{n}_e \sigma_T a (v_b - 3\Theta_1) . \end{aligned} \quad (33.89)$$

The result is that the equations governing the baryon overdensity δ_b and scalar bulk velocity v_b look like those (33.33) governing non-baryonic cold dark matter, except that the velocity equation has an additional source (33.89) arising from momentum transfer with photons through electron-photon scattering:

$$\dot{\delta}_b - k v_b - 3\dot{\Phi} = 0 , \quad (33.90a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k \Psi = - \frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) , \quad (33.90b)$$

where $R \equiv \frac{3}{4} \bar{\rho}_b / \bar{\rho}_\gamma$ is $\frac{3}{4}$ the baryon-to-photon density ratio, equation (32.46).

33.12 Boltzmann equation for relativistic neutrinos

Neutrinos oscillations indicate that at least two of the three neutrino types have mass, §10.25; but the masses are (probably) small enough that all three neutrinos types were relativistic until some time after recombination, equation (10.112). As long as neutrinos are relativistic, the hierarchy of Boltzmann equations is the same as that for photons, equations (33.81), but without the scattering terms,

$$\dot{\mathcal{N}}_0 - k \mathcal{N}_1 - \dot{\Phi} = 0 , \quad (33.91a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3} (\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3} \Psi = 0 , \quad (33.91b)$$

$$\dot{\mathcal{N}}_\ell + \frac{k}{2\ell+1} [\ell \mathcal{N}_{\ell-1} - (\ell+1) \mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) . \quad (33.91c)$$

The radiative transfer equation for neutrinos can be solved explicitly, equation (34.46). That solution, which involves an integral over the line of sight, provides one way to calculate the multipoles needed in the Einstein equations. However, computer codes that model the CMB commonly calculate the neutrino multipoles $\mathcal{N}_\ell$ from the Boltzmann hierarchy (33.91) suitably truncated at some high harmonic $\ell_{\max}$. Since free streaming allows high neutrino multipoles to become comparable to the monopole and dipole well inside the horizon, it is not a good approximation simply to set neutrino multipoles to zero above some maximum harmonic. A better approximation, which emerges from the radiative transfer solution (34.46), is the approximation (34.49),

$$\mathcal{N}_{\ell_{\max}} \approx - (\mathcal{N}_{\ell_{\max}-2} + \delta_{\ell_{\max}2} \Psi) - \frac{2\ell_{\max}-1}{k\eta} \mathcal{N}_{\ell_{\max}-1} , \quad (33.92)$$

At superhorizon scales, the neutrino distribution was isotropic like any other species. But free-streaming allowed neutrinos to develop significant anisotropy once the scale entered the horizon. Prior to recombination,

neutrinos provided the principal quadrupole pressure that sourced the difference $\Psi - \Phi$ of scalar potentials, Figure 33.4. In Exercise 33.5 you will find that, surprisingly, neutrino anisotropy sourced a finite difference $\Psi - \Phi$ even in the initial superhorizon conditions where the neutrino monopole dominated.

Exercise 33.5 Initial conditions in the presence of neutrinos. Prior to recombination, the neutrino quadrupole pressure is the dominant source for the difference $\Psi - \Phi$ in scalar potentials, Figure 33.4. In this problem you will find that the neutrino quadrupole leads to a finite difference $\Psi - \Phi$ even in the initial conditions at superhorizon scales. Exercise 35.10 considers initial conditions for tensor fluctuations of neutrinos.

- Initially, only the neutrino monopole $\mathcal{N}_0$ is finite. In the Boltzmann hierarchy (33.91) of equations, the lower order multipoles drive the higher multipoles, so that the equations reduce to the form $\dot{\mathcal{N}}_\ell \propto \mathcal{N}_{\ell-1}$. Specifically, the Boltzmann hierarchy (33.91) reduces to, with $y \equiv k\eta$,

$$\frac{d(\mathcal{N}_0 - \Phi)}{dy} = 0, \quad (33.93a)$$

$$\frac{d\mathcal{N}_1}{dy} = -\frac{1}{3}(\mathcal{N}_0 + \Psi), \quad (33.93b)$$

$$\frac{d\mathcal{N}_\ell}{dy} = -\frac{\ell}{2\ell+1}\mathcal{N}_{\ell-1} \quad (\ell \geq 2). \quad (33.93c)$$

Show that the initial ($y \ll 1$) behaviour of the neutrino multipoles is

$$\mathcal{N}_\ell = \frac{(-y)^\ell}{(2\ell+1)!!}(\mathcal{N}_0 + \Psi) \quad (\ell \geq 1). \quad (33.94)$$

- Let f_γ and f_ν be the photon and neutrino fraction of the total radiation density,

$$f_\gamma \equiv \frac{\bar{\rho}_\gamma}{\bar{\rho}_\gamma + \bar{\rho}_\nu} = 1 - f_\nu, \quad f_\nu \equiv \frac{\bar{\rho}_\nu}{\bar{\rho}_\gamma + \bar{\rho}_\nu} = \frac{6 \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}}{2 + 6 \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}} \approx 0.405. \quad (33.95)$$

Show that the Einstein energy equation (33.7a) implies, initially,

$$-\Psi = 2(\Phi + \zeta_r), \quad (33.96)$$

where $\zeta_r \equiv f_\gamma \zeta_\gamma + f_\nu \zeta_\nu$. Assume that the photon quadrupole is negligible (why?). Show that the Einstein quadrupole pressure equation (33.7b) implies, initially,

$$\Psi - \Phi = -\frac{4}{5}f_\nu(\Psi + \Phi + \zeta_\nu). \quad (33.97)$$

- Conclude that, for adiabatic initial conditions $\zeta_\nu = \zeta_\gamma$,

$$\Psi = -\frac{10\zeta_\nu}{15 + 4f_\nu}, \quad \Phi = \left(1 + \frac{2}{5}f_\nu\right)\Psi. \quad (33.98)$$

33.13 Massive neutrinos

Once neutrinos become nonrelativistic, they start to behave like matter, clustering gravitationally like non-baryonic cold dark matter and baryons. Each massive neutrino type defines a free-streaming scale, equal to the characteristic comoving distance that the neutrinos can travel before redshifting to a halt. This free-streaming scale equals approximately the comoving horizon size at the redshift when the neutrino type became nonrelativistic, equation (10.112). Massive neutrinos tend to depress the matter power spectrum at scales smaller than the neutrino free-streaming scale.

The suppression of matter power below the free-streaming scale is substantial (exponential) if massive neutrinos are a dominant component of matter, a scenario termed hot dark matter, or HDM. White, Frenk, and Davis (1983) used the absence of such suppression in the observed galaxy power spectrum to rule out HDM models 30 years ago.

33.13.1 Simplified treatment of massive neutrinos

A full treatment of massive neutrinos, §33.13.2, requires integrating a Boltzmann hierarchy of multipole equations for each of a spectrum of neutrino momenta p_ν . This is more complicated than the massless case, where the fact that massless neutrinos follow the same null worldline regardless of the magnitude of their momentum implies that a single Boltzmann hierarchy covers all momenta.

A simple approximate solution to the additional complexity introduced by mass is to assume an abrupt transition from relativistic to nonrelativistic neutrinos at some time. This was the strategy suggested in Exercise 32.4.

Another possible simplified strategy is to follow the Boltzmann hierarchy (33.100) for just a single representative neutrino momentum p_ν near the peak of the distribution, $p_\nu/T_\nu = 1$.

33.13.2 Full treatment of massive neutrinos

The Boltzmann equation for collisionless neutrinos with mass is equation (33.43) with zero collision term. For massive neutrinos, the neutrino velocity v_ν depends on momentum, so the neutrino temperature fluctuation $\mathcal{N} \equiv \delta T_\nu(\eta, \mathbf{k}, \mathbf{p}_\nu)/T_\nu(\eta)$ depends not only on the direction $\hat{\mathbf{p}}_\nu$ of the neutrino momentum, as in the massless case, but also on its magnitude p_ν . Since the temperature fluctuation $\mathcal{N}$ is already of first order, it suffices to treat the particle velocity v_ν in the Boltzmann equation (33.43) to unperturbed order. To unperturbed order, the momentum of a freely streaming neutrino redshifts as $p_\nu \propto a^{-1}$, and the temperature of the unperturbed distribution redshifts in the same way, $T_\nu \propto a^{-1}$. Thus it is convenient to characterize the neutrino temperature fluctuation as a function of the time-independent ratio p_ν/T_ν ,

$$\mathcal{N}(\eta, \mathbf{k}, \mathbf{p}_\nu) = \mathcal{N}(\eta, \mathbf{k}, p_\nu/T_\nu, \hat{\mathbf{p}}_\nu) . \quad (33.99)$$

The harmonics $\mathcal{N}_\ell(\eta, \mathbf{k}, p_\nu/T_\nu)$ of the temperature fluctuation are functions of the ratio p_ν/T_ν .

At the risk of being repetitious, the Boltzmann hierarchy of equations for a species of massive neutrino is,

equations (33.49),

$$\dot{\mathcal{N}}_0 - v_\nu k \mathcal{N}_1 - \dot{\Phi} = 0 , \quad (33.100a)$$

$$\dot{\mathcal{N}}_1 + \frac{v_\nu k}{3} (\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3v_\nu} \Psi = 0 , \quad (33.100b)$$

$$\dot{\mathcal{N}}_\ell + \frac{v_\nu k}{2\ell + 1} [\ell \mathcal{N}_{\ell-1} - (\ell + 1) \mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) , \quad (33.100c)$$

which differs from the massless neutrino hierarchy (33.91) in that it depends on the velocity $v_\nu \equiv p_\nu/E_\nu$ of the neutrino. As long as neutrinos are relativistic, it suffices to follow a single hierarchy with $v_\nu = 1$. But as neutrinos become nonrelativistic, a full treatment requires following neutrino with different momenta p_ν separately. In due course the neutrinos become nonrelativistic, and the equations re-simplify to the nonrelativistic limit.

Because the massive neutrino multipoles $\mathcal{N}_\ell(p_\nu/T_\nu)$ depend on neutrino momentum p_ν , the perturbed neutrino energy-momentum tensor $\dot{T}_\nu^{kl}$ is more complicated than the massless case, equations (33.53). The perturbed energy density, energy flux, monopole pressure, and quadrupole pressure of massive neutrinos are

$$\dot{T}_\nu^{00} = \int \frac{\partial \dot{f}_\nu}{\partial \ln T_\nu} \mathcal{N}_0(p_\nu/T_\nu) E_\nu \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (33.101a)$$

$$\hat{k}_a \dot{T}_\nu^{0a} = -i \int \frac{\partial \dot{f}_\nu}{\partial \ln T_\nu} \mathcal{N}_1(p_\nu/T_\nu) p_\nu^a \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (33.101b)$$

$$\frac{1}{3} \delta_{ab} \dot{T}_\nu^{ab} = \frac{1}{3} \int \frac{\partial \dot{f}_\nu}{\partial \ln T_\nu} \mathcal{N}_0(p_\nu/T_\nu) \frac{p_\nu^2}{E_\nu} \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (33.101c)$$

$$\left(\frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) \dot{T}_\nu^{ab} = - \int \frac{\partial \dot{f}_\nu}{\partial \ln T_\nu} \mathcal{N}_2(p_\nu/T_\nu) \frac{p_\nu^2}{E_\nu} \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} . \quad (33.101d)$$

33.14 Appendix: Legendre polynomials

The Legendre polynomials $P_\ell(\mu)$ satisfy the orthogonality relations

$$\int_{-1}^1 P_\ell(\mu) P_{\ell'}(\mu) d\mu = \frac{2}{2\ell + 1} \delta_{\ell\ell'} , \quad (33.102)$$

the addition theorem

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}^*(\hat{\mathbf{a}}) Y_{\ell m}(\hat{\mathbf{b}}) = \frac{2\ell + 1}{4\pi} P_\ell(\hat{\mathbf{a}} \cdot \hat{\mathbf{b}}) , \quad (33.103)$$

the recurrence relation

$$\mu P_\ell(\mu) = \frac{1}{2\ell + 1} [\ell P_{\ell-1}(\mu) + (\ell + 1) P_{\ell+1}(\mu)] , \quad (33.104)$$

and the derivative relation

$$\frac{dP_\ell(\mu)}{d\mu} = \frac{\ell+1}{1-\mu^2} [\mu P_{\ell-1}(\mu) - P_{\ell+1}(\mu)] \quad . \quad (33.105)$$

The first few Legendre polynomials are

$$P_0(\mu) = 1 \quad , \quad P_1(\mu) = \mu \quad , \quad P_2(\mu) = -\frac{1}{2} + \frac{3}{2}\mu^2 \quad . \quad (33.106)$$

Fluctuations in the Cosmic Microwave Background

Since the first definitive observation of the amplitude of the first peak of the power spectrum of temperature fluctuations in the CMB by the Boomerang balloon-based experiment (de Bernardis et al., 2000), the observed power spectrum of the CMB has allowed cosmological parameters to be measured with ever-increasing precision, and has provided the primary basis for the Standard Model of Cosmology. It should be emphasized that the CMB power spectrum is by no means the only evidence supporting the Standard Model. What gives confidence in the Standard Model is the fact that a broad range of other astronomical observations are consistent with it, including the Hubble diagram of Type I supernovae, the clustering of matter and of galaxies, Big Bang nucleosynthesis, and the age of the oldest stars.

The power spectrum of the CMB depends on the harmonics $\Theta_\ell(\eta_0, \mathbf{k})$ of the CMB photon distribution at the present time. A fast and elegant approach to calculating these harmonics was pointed out by Seljak and Zaldarriaga (1996).

34.1 Radiative transfer of CMB photons

To determine the harmonics $\Theta_\ell(\eta_0, \mathbf{k})$ of the CMB photon distribution today, return to the Boltzmann equation (33.80) for the temperature fluctuation $\Theta(\eta, \mathbf{k}, \mu)$, where $\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma$ is the cosine of the angle between the wavevector $\mathbf{k}$ and the photon direction $\mathbf{p}_\gamma$. It proves advantageous to rearrange the photon Boltzmann equation as

$$\left(\frac{\partial}{\partial \eta} - ik\mu + |\dot{\tau}| \right) (\Theta + \Psi) = I + |\dot{\tau}| S, \quad (34.1)$$

which in this context is called the **radiative transfer equation**. The terms on the right hand side are source terms. The term I on the right hand side of the radiative transfer equation (34.1) is a monopole term, the Integrated Sachs-Wolfe (ISW) term,

$$I(\eta, \mathbf{k}) \equiv \dot{\Psi}(\eta, \mathbf{k}) + \dot{\Phi}(\eta, \mathbf{k}), \quad (34.2)$$

so-called because, as will be seen in equation (34.17), it contributes a temperature fluctuation that is an integral along the line of sight to the CMB. The term S on the right hand side of equation (34.1) embodies

source terms arising from Thomson scattering, Figure 35.1, a sum of monopole, dipole, and quadrupole harmonics,

$$\begin{aligned} S(\eta, \mathbf{k}, \mu) &\equiv \Theta_0(\eta, \mathbf{k}) + \Psi(\eta, \mathbf{k}) - i\mu v_b(\eta, \mathbf{k}) - \frac{1}{2}\Theta_2(\eta, \mathbf{k})P_2(\mu) \\ &= \sum_{n=0}^2 (-i)^n S_n(\eta, \mathbf{k}) P_n(\mu) , \end{aligned} \quad (34.3)$$

with harmonic coefficients $S_n(\eta, \mathbf{k})$,

$$S_0 \equiv \Theta_0 + \Psi , \quad (34.4a)$$

$$S_1 \equiv v_b , \quad (34.4b)$$

$$S_2 \equiv \frac{1}{2}\Theta_2 . \quad (34.4c)$$

The relatively simple structure of the Thomson scattering source functions (34.4), containing only monopole, dipole, and quadrupole contributions, stems from the simple structure (33.55) of the quantum mechanical amplitude squared $\langle |\mathcal{M}|^2 \rangle$ for nonrelativistic electron-photon scattering, which contains only monopole and quadrupole contributions. The dipole source S_1 is a Doppler term from the motion of the photon-baryon fluid at velocity v_b .

The electron-photon (Thomson) scattering optical depth τ is defined by equation (32.3), the integral along the line sight of the scattering rate $\dot{\tau}$, equation (32.4). The optical depth is zero, $\tau_0 = 0$, at zero redshift, and increases going backwards in time η to higher redshift. The radiative transfer equation (34.1) can be written (note that $\dot{\tau}$ is negative)

$$e^{ik\mu\eta+\tau} \frac{d}{d\eta} [e^{-ik\mu\eta-\tau}(\Theta + \Psi)] = I - \dot{\tau}S . \quad (34.5)$$

The solution for the photon distribution $\Theta(\eta_0, \mathbf{k}, \mu)$ today is obtained by integrating the radiative transfer equation (34.5) over the line of sight from the Big Bang ($\eta = 0$) to the present time ($\eta = \eta_0$),

$$\Theta(\eta_0, \mathbf{k}, \mu) + \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [I(\eta, \mathbf{k}) - \dot{\tau}S(\eta, \mathbf{k}, \mu)] e^{-ik\mu(\eta-\eta_0)-\tau} d\eta . \quad (34.6)$$

Notice that the left hand side of the solution (34.6) of the radiative transfer equation is not the temperature fluctuation $\Theta(\eta_0, \mathbf{k}, \mu)$ by itself, but rather the temperature fluctuation redshifted by the potential, $\Theta(\eta_0, \mathbf{k}, \mu) + \Psi(\eta_0, \mathbf{k})$. The potential $\Psi(\eta_0, \mathbf{k})$ is independent of the photon direction $\hat{\mathbf{p}}_\gamma$, so contributes only to the monopole moment of the photon distribution.

34.1.1 Visibility function

Introduce the **visibility function** $g(\eta)$ defined by

$$g(\eta) \equiv -\dot{\tau}e^{-\tau} . \quad (34.7)$$

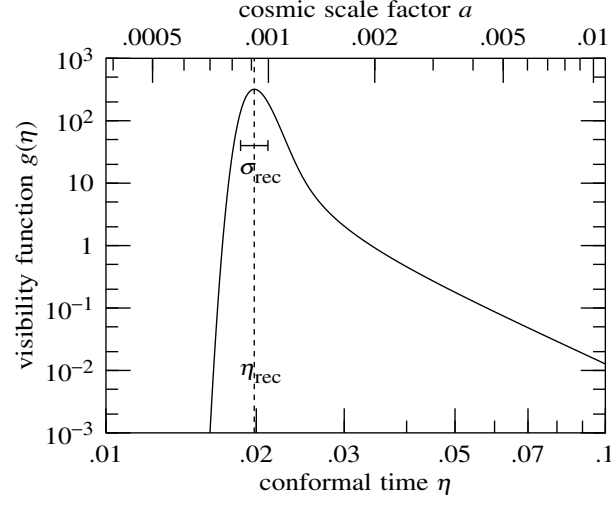


Figure 34.1 Visibility function $g(\eta)$ as a function of conformal time η in units of the conformal time today, $\eta_0 = 1$. The visibility function here is calculated from the Peebles approximation to recombination, Exercise 31.6. The dashed vertical line marks the time η_{rec} of recombination, where the optical depth is one. The width σ_{rec} of recombination is the standard deviation of a Gaussian fit to the core of the visibility function.

The visibility function $g(\eta)$, illustrated in Figure 34.1, acts like a smoothing window over recombination. The visibility function is fairly narrowly peaked around recombination at $\eta = \eta_{\text{rec}}$, and its integral is one,

$$\int_0^{\eta_0} g(\eta) d\eta = \int_{-\infty}^0 -e^{-\tau} d\tau = [e^{-\tau}]_{-\infty}^0 = 1. \quad (34.8)$$

The visibility function $g(\eta)$ has an approximately Gaussian core, and a long tail extending past recombination. The long tail arises because recombination leaves a finite residual electron density.

The solution (34.6) of the radiative transfer equation can be written in terms of the visibility function $g(\eta)$ as

$$\Theta(\eta_0, \mathbf{k}, \mu) + \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [e^{-\tau} I(\eta, \mathbf{k}) + g(\eta) S(\eta, \mathbf{k}, \mu)] e^{-ik\mu(\eta-\eta_0)} d\eta. \quad (34.9)$$

34.2 Harmonics of the CMB photon distribution

If the temperature fluctuations on the CMB sky are statistically isotropic, then the statistical properties of the CMB commute with the rotation operator (the angular momentum operator), which implies that the power spectrum of CMB fluctuations is diagonal in a basis of eigenfunctions of the rotation operator. The eigenfunctions are spherical harmonics. Thus it is natural to expand the temperature fluctuation $\Theta(\eta_0, \mathbf{k}, \mu) + \Psi(\eta_0, \mathbf{k})$ in harmonics, equation (33.47).

The $e^{-ik\mu(\eta-\eta_0)}$ factor in the integrand on the right hand side of equation (34.9) can be expanded in harmonics through the general formula

$$e^{-iy\mu} = \sum_{\ell=0}^{\infty} (-i)^\ell (2\ell+1) P_\ell(\mu) j_\ell(y) , \quad (34.10)$$

where $j_\ell(y) \equiv \sqrt{\pi/(2y)} J_{\ell+1/2}(y)$ are spherical Bessel functions, and here $y \equiv k(\eta - \eta_0)$. The source function S that premultiplies the factor $e^{-ik\mu(\eta-\eta_0)}$ in the integrand of equation (34.9) is a sum of harmonics, equation (34.3). It is useful to introduce modified spherical Bessel functions $j_{\ell n}(y)$ defined by an expansion analogous to (34.10),

$$(-i)^n P_n(\mu) e^{-iy\mu} = \sum_{\ell=0}^{\infty} (-i)^\ell (2\ell+1) P_\ell(\mu) j_{\ell n}(y) . \quad (34.11)$$

The orthogonality relations of the Legendre polynomials, equation (33.102), imply that

$$j_{\ell n}(y) = i^{\ell-n} \int_{-1}^1 e^{-iy\mu} P_\ell(\mu) P_n(\mu) \frac{d\mu}{2} , \quad (34.12)$$

which implies that $j_{\ell n}$ is symmetric or antisymmetric in its indices ℓn as their difference $\ell - n$ is even or odd,

$$j_{\ell n}(y) = (-)^{\ell-n} j_{n\ell}(y) . \quad (34.13)$$

The Legendre functions $P_n(\mu)$ are polynomials in μ , §33.14. Acting on $e^{-iy\mu}$, these polynomials can be replaced by derivatives with respect to y through

$$\mu^n e^{-iy\mu} = \left(i \frac{\partial}{\partial y} \right)^n e^{-iy\mu} . \quad (34.14)$$

The resulting modified spherical Bessel functions $j_{\ell n}(y)$ with $n = 0, 1, 2$ relevant here are

$$j_{\ell 0} = j_\ell , \quad j_{\ell 1} = j'_\ell , \quad j_{\ell 2} = \frac{1}{2} j_\ell + \frac{3}{2} j''_\ell , \quad (34.15)$$

where $'$ denotes the total derivative, $j'_\ell \equiv dj_\ell(y)/dy$. The modified spherical Bessel functions are even or odd as $j_{\ell n}(-y) = (-)^{l+n} j_{\ell n}(y)$. The harmonic expansion of equation (34.9) is thus

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} \left\{ e^{-\tau} I(\eta, \mathbf{k}) j_\ell[k(\eta_0 - \eta)] + g(\eta) \sum_{n=0}^2 S_n(\eta, \mathbf{k}) j_{\ell n}[k(\eta - \eta_0)] \right\} d\eta , \quad (34.16)$$

where $g(\eta)$ is the visibility function defined by equation (34.7). With the ISW and scattering source terms I and S_n written out explicitly, equation (34.16) is an integral from the Big Bang ($\eta = 0$) to the present time

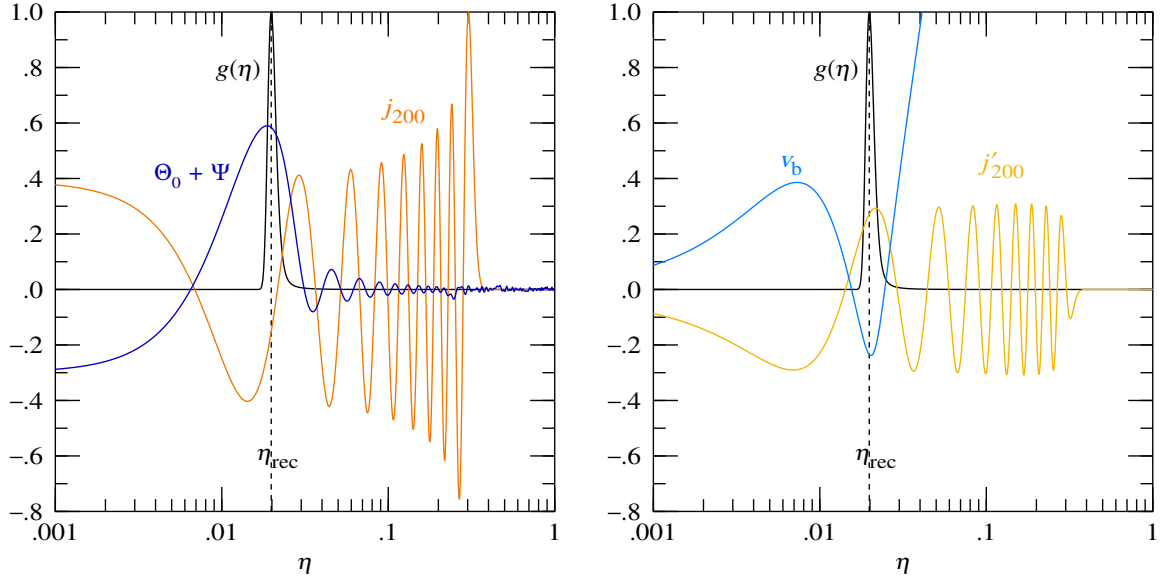


Figure 34.2 Illustrative example of the factors that go into the (left) monopole ($n = 0$) and (right) dipole ($n = 1$) contributions to the integrand of the solution (34.17) of the radiative transfer equation, as a function of conformal time η , in units $\eta_0 = 1$. The example is for a representative wavenumber $k/(a_{\text{eq}}H_{\text{eq}}) = 2$, and harmonic number $\ell = 200$. The factors are the visibility function $g(\eta)$, equation (34.7), scattering source terms $S_n(\eta, \mathbf{k})$, equations (34.4), and modified spherical Bessel functions $j_{\ell n}[k(\eta_0 - \eta)]$, equations (34.15). The visibility function $g(\eta)$ has been scaled to 1 at its peak, and the monopole and dipole spherical Bessel functions j_ℓ and j'_ℓ have been scaled so j_ℓ equals 1 at its (first) peak. The cosmological model is as given in §32.3. The dashed vertical line marks the time η_{rec} of recombination, where the Thomson optical depth is one.

($\eta = \eta_0$),

$$\begin{aligned}
 \Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} \bigg\{ & e^{-\tau} [\dot{\Psi}(\eta, \mathbf{k}) + \dot{\Phi}(\eta, \mathbf{k})] j_\ell[k(\eta - \eta_0)] && \text{ISW} \\
 & + g(\eta) \left\{ [\Theta_0(\eta, \mathbf{k}) + \Psi(\eta, \mathbf{k})] j_\ell[k(\eta - \eta_0)] && \text{monopole} \right. \\
 & + v_b(\eta, \mathbf{k}) j_{\ell 1}[k(\eta - \eta_0)] && \text{dipole} \\
 & \left. + \frac{1}{2} \Theta_2(\eta, \mathbf{k}) j_{\ell 2}[k(\eta - \eta_0)] \right\} && \text{quadrupole} \bigg\} d\eta .
 \end{aligned} \tag{34.17}$$

The term in the first line on the right hand side of equation (34.17) is an integral of the time derivative of the gravitational potential $\Psi + \Phi$ over the line of sight, and is called the **Integrated Sachs-Wolfe (ISW)** effect. The remaining terms are linear combinations of the monopole, dipole, and quadrupole scattering source terms S_n , equations (34.4). Note that the monopole term (on both sides of equation (34.17)) is not the temperature fluctuation Θ_0 by itself, but rather $\Theta_0 + \Psi$, which is the temperature fluctuation redshifted

by the potential Ψ . In the tight-coupling approximation, the baryon velocity on the third line approximates the photon velocity, $v_b \approx 3\Theta_1$.

Figure 34.2 shows an illustrative example of the factors that go into the monopole and dipole contributions to the integrand of the solution (34.17) of the radiative transfer equation.

34.2.1 Harmonics of the CMB with respect to observed photon directions

A final consideration is that the observed direction $\hat{\mathbf{n}}$ of a photon from the CMB is *opposite* to the photon's direction of motion, $\hat{\mathbf{n}} = -\hat{\mathbf{p}}_\gamma$. Photon multipoles Θ_ℓ^{obs} expanded with respect to the direction $\hat{\mathbf{n}}$ of observation are obtained from Θ_ℓ by flipping the sign of the photon direction, $\Theta_\ell^{\text{obs}}(\eta, \mathbf{k}, \mu) = \Theta_\ell(\eta, \mathbf{k}, -\mu)$. Flipping the sign of $\hat{\mathbf{p}}_\gamma$ is equivalent to flipping the sign of odd parity fluctuations,

$$\Theta_\ell^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^\ell \Theta_\ell(\eta_0, \mathbf{k}) . \quad (34.18)$$

Another way to achieve the sign flip is to flip the sign of the argument $k(\eta - \eta_0) \rightarrow k(\eta_0 - \eta)$ of the modified Bessel functions $j_{\ell n}$ in equation (34.17) and simultaneously to flip the sign of the odd source functions S_ℓ , namely $S_1 = v_b \rightarrow -v_b$. The CMB power spectrum involves products of pairs of Θ_ℓ^{obs} with the same ℓ , and is unaffected by the sign flip $\hat{\mathbf{n}} = -\hat{\mathbf{p}}_\gamma$ in the solution (34.17) of the radiative transfer equation.

34.2.2 Integrated Sachs-Wolfe (ISW) effect

The first line of the solution (34.17) of the radiative transfer equation is an integral of the time derivative of the potential $\Psi + \Phi$ along the line of sight to recombination. The contribution is called the Integrated Sachs-Wolfe (ISW) effect. If matter dominates the background, then the potential $\Psi + \Phi$ is constant in time for linear fluctuations, and there is no ISW effect. In practice, there are “early” and “late” ISW effects that arise respectively from the contributions of radiation to the background density shortly after recombination, and of dark energy (and possibly curvature) near the present time. The late time scalar potentials Ψ and Φ (which are equal at late times) evolve in proportion to the growth factor $g(a)$, equation (30.127) (not to be confused with the visibility function $g(\eta)$). The ISW integrand splits accordingly into early and late contributions,

$$e^{-\tau} \frac{d(\Psi + \Phi)}{d\eta} = e^{-\tau} g \frac{d}{d\eta} \left(\frac{\Psi + \Phi}{g} \right) + e^{-\tau} \left(\frac{\Psi + \Phi}{g} \right) \frac{dg}{d\eta} . \quad (34.19)$$

ISW early ISW late ISW

The early and late time contributions to the ISW term are illustrated in Figure 34.3.

In addition to the early and late ISW effects, there is a “nonlinear” ISW effect. Nonlinear gravitational clustering causes the potential Φ to change in time, becoming deeper (more negative) in more highly clustered regions. Photons that travel through a cluster see a slightly deeper potential when they exit the cluster than when they entered it, causing the photon to be slightly redshifted. Figure 34.3 does not include the nonlinear ISW effect.

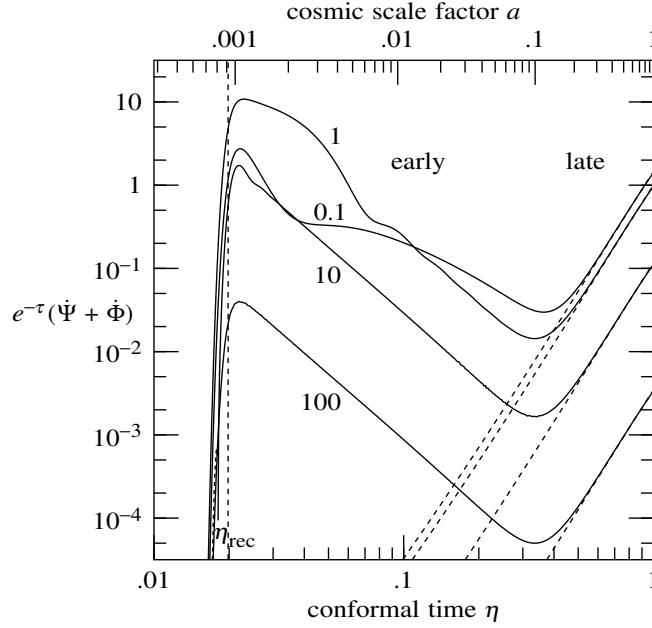


Figure 34.3 ISW integrand $e^{-\tau}(\dot{\Psi} + \dot{\Phi})$ in equation (34.17) as a function of conformal time η , in units $\eta_0 = 1$, for the standard flat Λ CDM model of §32.3. Curves are labelled with the wavenumber $k/(a_{\text{eq}}H_{\text{eq}})$ in units of the Hubble distance at matter-radiation equality. In a matter-dominated cosmology, the gravitational potentials are constant, and the curves would all be zero. The high early values following recombination result from the contribution of radiation to the mean mass-energy density; this is the “early” ISW effect. The turn-up at later times results from the contribution of a cosmological constant to the mean mass-energy density; this is the “late” ISW effect, indicated by the dashed lines. The late time ISW contribution causes a slight turn-up in the CMB power spectrum at the largest scales, Figure 34.8, a characteristic signature of a cosmological constant.

34.2.3 CMB transfer function in Fourier space

As seen in Chapters 30–33, during linear evolution, scalar modes of given comoving wavevector $\mathbf{k}$ evolve with amplitude proportional to the initial curvature fluctuation $\zeta(\mathbf{k})$. The evolution of the amplitude may be encapsulated in a CMB transfer function $T_\ell(\eta, k)$ defined by

$$T_\ell(\eta, k) \equiv \frac{\Theta_\ell(\eta, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta, \mathbf{k})}{\zeta(\mathbf{k})}, \quad (34.20)$$

with $\Theta_\ell(\eta, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta, \mathbf{k})$ computed from equation (34.17). By isotropy, the CMB transfer function $T_\ell(\eta, k)$ is a function only of the magnitude k of the wavevector $\mathbf{k}$.

The square of the transfer function transforms the primordial power spectrum $P_\zeta(k)$ defined by equation (30.131) into the CMB power spectrum $C_\ell(\eta, k)$ in Fourier space, equation (34.27), which is in turn related to the observed CMB power spectrum $C_\ell(\eta_0)$ in real space today by equation (34.34).

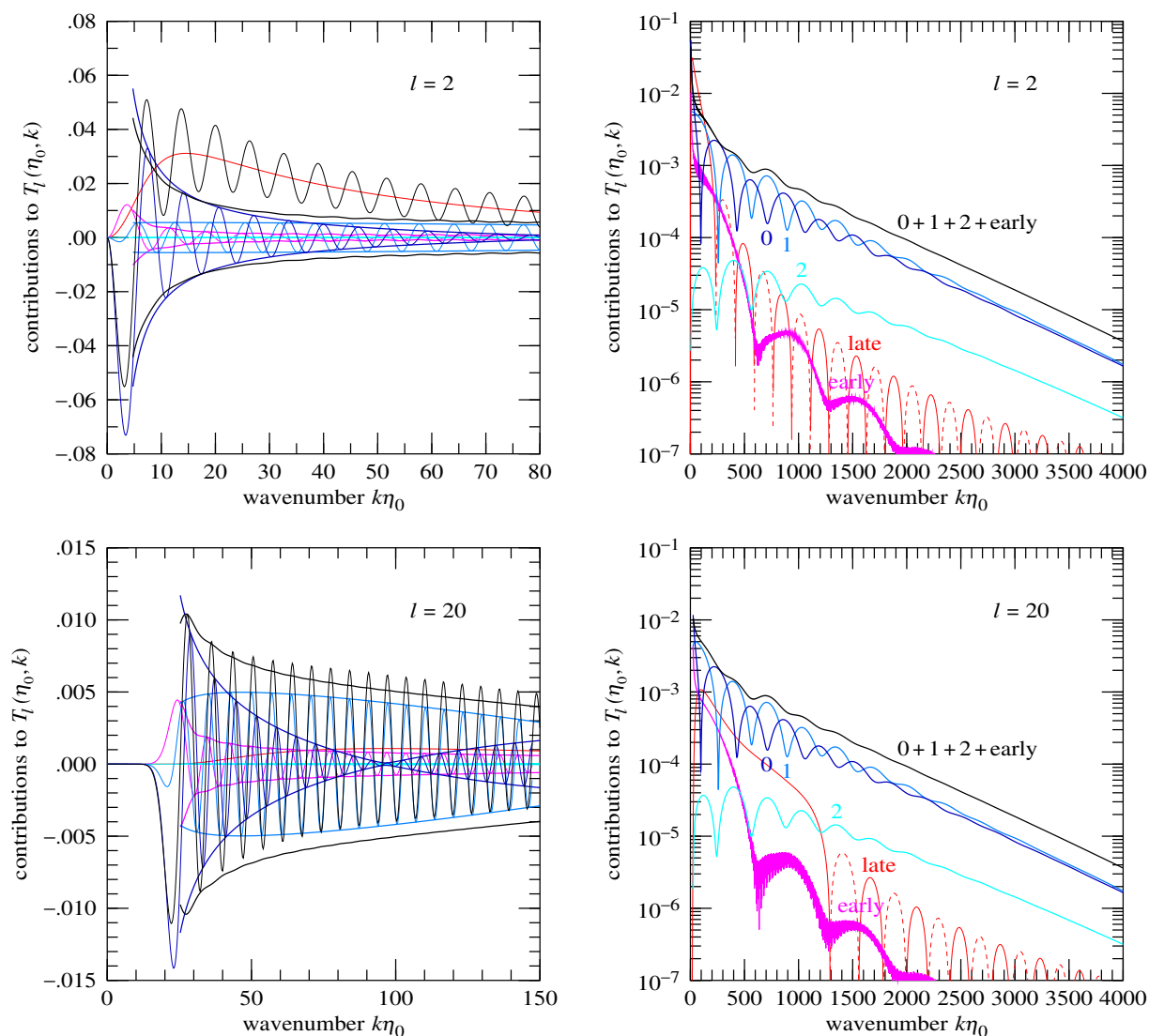


Figure 34.4 (Continued on the next page.) CMB transfer functions $T_\ell(\eta_0, k)$ for a selection of harmonics ℓ , plotted (left) linearly, showing the oscillating functions and their envelopes, and (right) logarithmically over a broader range of wavenumber k , showing only the envelopes of the underlying oscillating functions. The total (black) is a sum of the various contributions in equation (34.17): monopole (dark blue), dipole (light blue), quadrupole (cyan), early ISW (purple), and late ISW (red). The total envelope (black) omits the late ISW contribution, since the late ISW is non-oscillatory where it is important (at small ℓ and small k). The cosmological model is the flat Λ CDM concordance model of §32.3. The computation is a Boltzmann computation including photon and neutrino multipoles up to $\ell_{\max} = 16$.

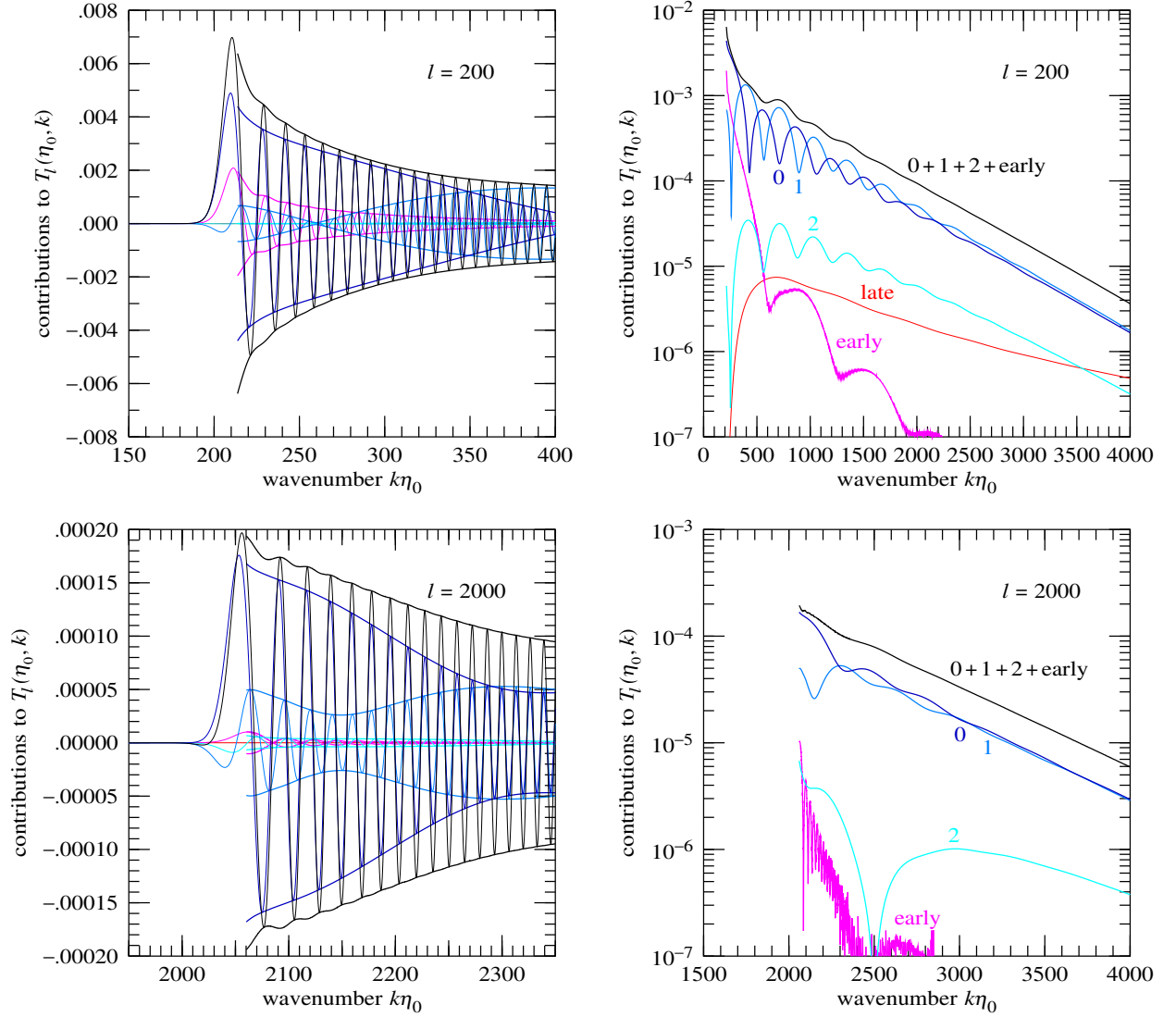


Figure 34.4 continued.

Figure 34.4 shows CMB transfer functions $T_\ell(\eta_0, k)$ at the present time, $\eta = \eta_0$, for a selection of harmonics, $\ell = 2, 20, 200$, and 2000. The CMB transfer functions are calculated by integrating numerically, for each of many wavenumbers k , the solution (34.17) of the radiative transfer equation. The CMB transfer functions

shown in Figure 34.4 are from a Boltzmann computation including photon and neutrino multipoles up to $\ell_{\max} = 16$.

Spherical Bessel functions $j_\ell(y)$ are small for $y \ll \ell$, rise to their first peak at $y \approx \ell + \ell^{1/3}$, and are then oscillating and declining at $y \gg \ell$. This behaviour translates into a similar behaviour in the CMB transfer functions $T_\ell(\eta_0, \mathbf{k})$, Figure 34.4. The transfer functions are small for $k(\eta_0 - \eta_{\text{rec}}) \ll \ell$, peak at

$$k(\eta_0 - \eta_{\text{rec}}) \sim \ell + \ell^{1/3} , \quad (34.21)$$

and then oscillate at $k(\eta_0 - \eta_{\text{rec}}) \gg \ell$ with an exponentially declining envelope, as illustrated in the right panels of Figure 34.4,

$$T_\ell(\eta_0, \mathbf{k})|_{\text{envelope}} \propto \exp(-k\eta_0/600) . \quad (34.22)$$

The exponential decline is caused in part by dissipative processes around the time of recombination, §32.7 and §32.8, and in part by the finite width of recombination, which tends to smooth over oscillating source functions S_ℓ at large wavenumber k , §34.2.4.

Besides the total, Figure 34.4 shows the monopole, dipole, quadrupole, and early and late ISW contributions to the CMB transfer functions. The contributions are, excepting late ISW, highly oscillatory, thanks to the Bessel factors $j_{\ell n}[k(\eta - \eta_0)]$ in the integrand of equation (34.17). The Figure therefore shows also the envelope of the oscillatory contributions. The envelope is computed as an integral in which the Bessel factor $j_{\ell n}(y)$ in the integrand is replaced by the non-oscillatory absolute value of the complex Hankel factor $h_{\ell n}(y)$,

$$h_{\ell n}(y) \equiv j_{\ell n}(y) + \begin{cases} 0 & |y| < \ell + \frac{1}{2} + (\ell + \frac{1}{2})^{1/3} \\ i y_{\ell n}(y) & |y| \geq \ell + \frac{1}{2} + (\ell + \frac{1}{2})^{1/3} \end{cases} , \quad (34.23)$$

with $y_{\ell n}(y)$ the modified spherical Bessel function of the second kind (whereas $j_{\ell n}(y)$ is the modified spherical Bessel function of the first kind). The cut at $y = \ell + \frac{1}{2} + (\ell + \frac{1}{2})^{1/3}$, which is roughly the location of the first zero of $y_{\ell n}(y)$, is introduced to prevent the diverging behaviour of $y_{\ell n}(y)$ as $y \rightarrow 0$ from dominating the integral.

34.2.4 Instantaneous and rapid recombination approximations

At wavelengths much larger than the width of recombination, $k\sigma_{\text{rec}} \ll 1$, recombination can be approximated as instantaneous. In the **instantaneous recombination approximation**, the visibility function $g(\eta)$ is a delta function at $\eta = \eta_{\text{rec}}$, and, without the ISW term, the multipoles $\Theta_\ell(\eta_0, \mathbf{k})$ of the temperature fluctuation today are

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) \approx \sum_{n=0}^2 S_n(\eta_{\text{rec}}, \mathbf{k}) j_{\ell n}[k(\eta_{\text{rec}} - \eta_0)] . \quad (34.24)$$

A better approximation that works also at larger k is the **rapid recombination approximation**, which replaces the source functions S by their averages $\bar{S}$ over recombination. In the rapid approximation, the

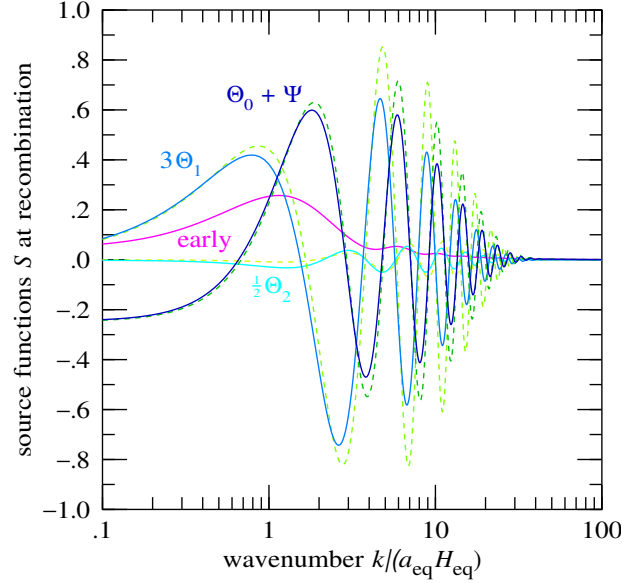


Figure 34.5 Rapid and instantaneous approximations to the Thomson scattering and early ISW source functions at recombination, equations (34.25), as a function of wavenumber $k/(a_{\text{eq}}H_{\text{eq}})$. The solid (bluish) lines are the values $\bar{S}_n(\mathbf{k})$ averaged over the visibility function $g(\eta)$, while the dashed (greenish) lines are the instantaneous values $S_n(\eta_{\text{rec}}, \mathbf{k})$ at recombination, where the Thomson optical depth is one. The purple line is the averaged early ISW source function $\bar{S}_{\text{early}}(\mathbf{k})$. The source functions are normalized to unit primordial curvature, $\zeta(\mathbf{k}) = 1$ (in other words, the plotted source functions are transfer functions). The computation is the hydrodynamic approximation (Boltzmann with $\ell_{\text{max}} = 2$), since this turns out to yield a better rapid recombination approximation to the CMB power spectrum than a full Boltzmann treatment, Figure 34.7. The dipole source term is taken to be $3\Theta_1$ (the tight-coupling limit), not v_b , since this yields a better rapid recombination approximation. The cosmological model is the standard flat Λ CDM model described in §32.3.

temperature multipoles $\Theta_\ell(\eta_0, \mathbf{k})$ today are, including the early ISW, monopole, dipole, and quadrupole contributions,

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) \approx \bar{S}_{\text{early}}(\mathbf{k}) j_\ell[k(\eta_{\text{early}} - \eta_0)] + \sum_{n=0}^2 \bar{S}_n(\mathbf{k}) j_{\ell n}[k(\eta_{\text{rec}} - \eta_0)] , \quad (34.25a)$$

$$\bar{S}_{\text{early}}(\mathbf{k}) \equiv \int_0^{\eta_0} e^{-\tau} g \frac{d}{d\eta} \left(\frac{\Psi(\eta, \mathbf{k}) + \Phi(\eta, \mathbf{k})}{g} \right) d\eta , \quad \bar{S}_n(\mathbf{k}) \equiv \int_0^{\eta_0} g(\eta) S_n(\eta, \mathbf{k}) d\eta , \quad (34.25b)$$

where in the early ISW term g denotes the growth factor (30.127) rather than the visibility function $g(\eta)$. The early ISW effect peaks at a redshift $z_{\text{early}} \approx 900$ slightly after the redshift $z_{\text{rec}} \approx 1100$ of recombination, Figure 34.3, but because the conformal time η_0 today is so much larger than η_{rec} , the difference between $\eta_{\text{early}} - \eta_0$ and $\eta_{\text{rec}} - \eta_0$ is slight enough that it is fair to approximate $\eta_{\text{early}} \approx \eta_{\text{rec}}$.

Figure 34.5 shows the early ISW, monopole, dipole, and quadrupole source terms that go into the instan-

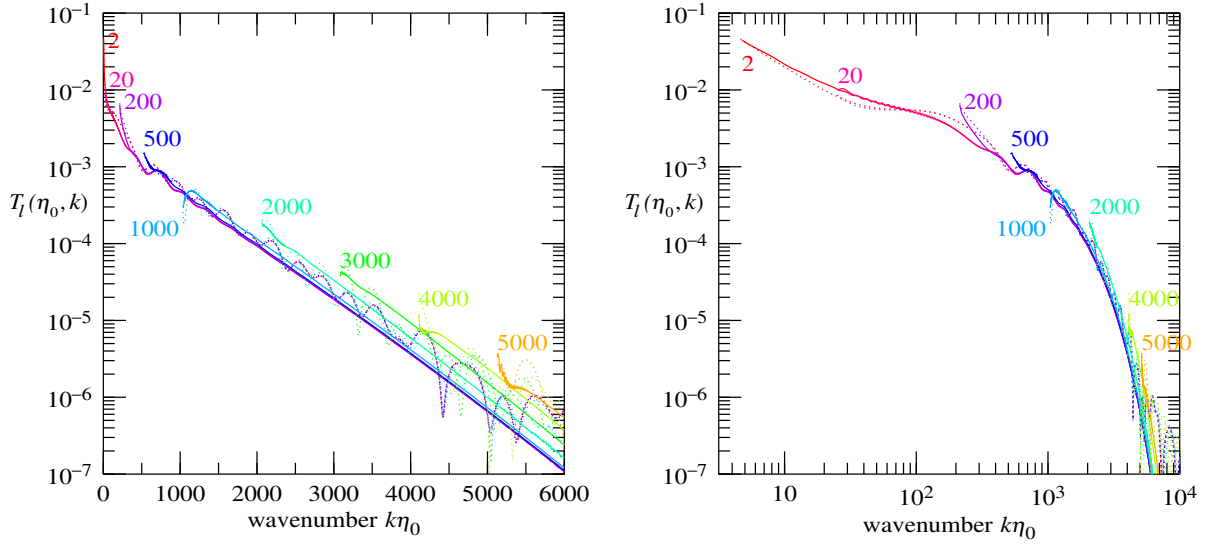


Figure 34.6 CMB transfer functions computed from a Boltzmann treatment (solid lines) including photon and neutrino multipoles up to $\ell_{\max} = 16$, compared to their values in the rapid recombination approximation (dotted lines) with the source functions taken from the hydrodynamic approximation, for a selection of harmonics ℓ , as marked. All lines are the envelopes of the underlying rapidly oscillating transfer functions. The left and right panels are the same, but with wavenumber k plotted linearly on the left, logarithmically on the right. The transfer functions plotted here include monopole, dipole, quadrupole, and early ISW contributions, but exclude the late ISW contribution. The dipole contribution to the rapid recombination approximation is computed from the photon velocity $3\Theta_1$ (the tight-coupling limit), not the baryon velocity v_b . The cosmological model is the standard flat Λ CDM model described in §32.3.

taneous approximation (dashed lines) and the rapid recombination approximation (solid lines). The instantaneous and rapid approximations $S_n(\eta_{\text{rec}}, \mathbf{k})$ and $\bar{S}_n(\mathbf{k})$ to the Thomson scattering source functions agree at small wavenumbers k , where the source terms are slowly varying over the visibility function. The rapid approximation works also at larger wavenumbers, where the source functions $S_n(\eta, \mathbf{k})$ change significantly over the course of recombination. Averaging over recombination tends to reduce the Thomson scattering source functions compared to their instantaneous values at recombination.

The baryon velocity decouples from the photon velocity during recombination, and grows large as baryons fall into the dark matter potential wells, as illustrated in Figures 32.1 or 33.1. As a result, the rapid approximation tends to overestimate the true dipole contribution if the baryon bulk velocity v_b is used for the dipole source term S_1 . A simple empirical fix is to use the bulk photon velocity $v_\gamma \equiv 3\Theta_1$ in place of the baryon velocity to compute S_1 . This fix is adopted in Figures 34.5 and 34.6.

Figure 34.6 compares (envelopes of the rapidly oscillating) CMB transfer functions $T_\ell(\eta, k)$ computed from a Boltzmann treatment to their values in the rapid recombination approximation, equation (34.25), with source functions computed in the hydrodynamic approximation, for a selection of harmonics ℓ . Whereas the envelopes computed in the Boltzmann treatment are rather smooth at large wavenumber k (and exponen-

tially declining, equation (34.22)), the envelopes computed in the rapid recombination approximation remain somewhat oscillatory at large wavenumber k . However, what is important is that the approximation yields approximately the correct overall amplitude of the transfer functions; the CMB power spectrum (34.35) involves integrating over transfer functions, which washes out the residual oscillatory structure. The hydrodynamic approximation (rather than a Boltzmann treatment) is used for the source functions because it (hydro + rapid) turns out to give a yield better approximation (than Boltzmann + rapid) to the CMB power spectrum, as illustrated in the right panel of Figure 34.7.

34.2.5 CMB power spectrum in Fourier space

The power spectrum $C_\ell(\eta, k)$ in Fourier space is defined to be the expectation value of the variance of temperature multipoles $\Theta_\ell(\eta, \mathbf{k})$,

$$\frac{\delta_{\ell'\ell}}{4\pi} (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) C_\ell(\eta, k) \equiv \langle [\Theta_{\ell'}(\eta, \mathbf{k}') + \delta_{\ell 0} \Psi(\eta, \mathbf{k}')] [\Theta_\ell(\eta, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta, \mathbf{k})] \rangle . \quad (34.26)$$

The power spectrum $C_\ell(\eta, k)$ is real-valued. The momentum conserving delta function $(2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k})$ is a consequence of the assumed statistical homogeneity of space, while the angular-momentum conserving delta function $\delta_{\ell'\ell}$ is a consequence of the assumed statistical isotropy of space. By isotropy, the power spectrum $C_\ell(\eta, k)$ is a function only of the magnitude k of the wavevector $\mathbf{k}$. The monopole power $C_0(\eta, k)$ is defined to be the variance of the redshifted monopole $\Theta_0 + \Psi$ because that is what appears in the solution (34.17) of the radiative transfer equation.

In terms of the CMB transfer function (34.20) and the primordial power spectrum $P_\zeta(k)$ defined by equation (30.131), the CMB power spectrum $C_\ell(\eta, k)$ is

$$C_\ell(\eta, k) = 4\pi |T_\ell(\eta, k)|^2 P_\zeta(k) . \quad (34.27)$$

34.3 CMB in real space

34.3.1 CMB harmonics in real space

The solution (34.17) of the radiative transfer equation is in terms of photon multipoles $\Theta_\ell(\eta, \mathbf{k})$ in Fourier space, but astronomers observe the CMB in real space. The real-space temperature fluctuation $\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}})$ at time η and comoving position $\mathbf{x}$ in observed direction $\hat{\mathbf{n}}$ on the sky is related to the Fourier-space temperature fluctuation by

$$\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}}) = \int e^{-i\mathbf{k} \cdot \mathbf{x}} \Theta(\eta, \mathbf{k}, \hat{\mathbf{n}}) \frac{d^3k}{(2\pi)^3} . \quad (34.28)$$

Astronomers observe the temperature fluctuation $\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}})$ now, at time η_0 , and here, at position $\mathbf{x}_0$. Without loss of generality, our position can be taken to be at the origin, $\mathbf{x}_0 = \mathbf{0}$, in which case the phase

factor is unity, $e^{-i\mathbf{k}\cdot\mathbf{x}_0} = 1$, and can be omitted,

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}) = \int \Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}) \frac{d^3k}{(2\pi)^3} . \quad (34.29)$$

The spherical harmonic expansion of the observed real-space temperature fluctuation today is, with a conventional choice of normalization of harmonics $\Theta_{\ell m}$,

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \Theta_{\ell m}(\eta_0, \mathbf{x}_0) Y_{\ell m}^*(\hat{\mathbf{n}}) . \quad (34.30)$$

The sum includes the monopole $\ell = 0$ harmonic because the mean temperature of the observable Universe may differ from the “true” mean temperature of the Universe. From the perspective of statistics, such a difference between the observed and true mean temperature can exist even though it is unobservable to an astronomer confined to position $\mathbf{x}_0$. An astronomer in a cosmologically distant future when the horizon is much larger than today would be able to measure the difference. The spherical harmonic expansion (33.47) of the Fourier-space temperature fluctuation may be written, in view of the relation (33.103) between Legendre polynomials and spherical harmonics

$$\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} (-i)^{\ell} 4\pi \Theta_{\ell}(\eta, \mathbf{k}) Y_{\ell m}(\hat{\mathbf{k}}) Y_{\ell m}^*(\hat{\mathbf{n}}) . \quad (34.31)$$

From equations (34.29)–(34.31) it follows that the real-space photon harmonics are

$$\Theta_{\ell m}(\eta_0, \mathbf{x}_0) = 4\pi (-i)^{\ell} \int \Theta_{\ell}(\eta_0, \mathbf{k}) Y_{\ell m}(\hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3} . \quad (34.32)$$

34.3.2 CMB power spectrum in real space

The CMB power spectrum $C_{\ell}(\eta_0)$ on the sky today is defined to be the expectation value of the variance of temperature multipoles $\Theta_{\ell m}(\eta_0, \mathbf{x}_0)$,

$$\delta_{\ell'\ell} \delta_{m'm} C_{\ell}(\eta_0) \equiv \langle [\Theta_{\ell'm'}^*(\eta_0, \mathbf{x}_0) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{x}_0)] [\Theta_{\ell m}(\eta_0, \mathbf{x}_0) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{x}_0)] \rangle . \quad (34.33)$$

The power spectrum $C_{\ell}(\eta_0)$ is real-valued. By homogeneity, the power spectrum $C_{\ell}(\eta_0)$ is independent of observer position $\mathbf{x}_0$. The real-space monopole harmonic $\Theta_{00}(\eta_0, \mathbf{x}_0) + \Psi(\eta_0, \mathbf{x}_0)$ is the temperature fluctuation gravitationally redshifted by the potential $\Psi(\eta_0, \mathbf{x}_0)$ at our position today. From the perspective of an observer at fixed position $\mathbf{x}_0$, the redshifted monopole is observationally indistinguishable from a rescaling of the mean temperature.

From the expression (34.32) for the real-space harmonics in terms of Fourier-space harmonics, together with the power spectrum (34.26) of the Fourier-space harmonics, it follows that the power spectrum $C_{\ell}(\eta_0)$ of real-space harmonics of the CMB today is

$$C_{\ell}(\eta_0) = \int C_{\ell}(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (34.34)$$

In terms of the CMB transfer function T_ℓ and primordial curvature power spectrum P_ζ or its dimensionless equivalent Δ_ζ^2 , equation (30.133), the power spectrum $C_\ell(\eta_0)$ is, from equation (34.27),

$$C_\ell(\eta_0) = 4\pi \int |T_\ell(\eta_0, k)|^2 P_\zeta(k) \frac{4\pi k^2 dk}{(2\pi)^3} = 4\pi \int |T_\ell(\eta_0, k)|^2 \Delta_\zeta^2(k) \frac{dk}{k}. \quad (34.35)$$

If the primordial power spectrum Δ_ζ^2 is a power-law with tilt n_s , equation (30.136), then the CMB power spectrum today is

$$C_\ell(\eta_0) = 4\pi \Delta_\zeta^2(k_p) \int_0^\infty |T_\ell(\eta_0, k)|^2 \left(\frac{k}{k_p}\right)^{n-1} \frac{dk}{k}. \quad (34.36)$$

As discussed in §34.2.3, the CMB transfer functions $T_\ell(\eta_0, k)$ are small for $k(\eta_0 - \eta_{\text{rec}}) \ll \ell$, peak near $k(\eta_0 - \eta_{\text{rec}}) \sim \ell$ (or more precisely, at harmonics slightly larger than ℓ , equation (34.21)), and then oscillate with an exponentially declining envelope, equation (34.22). Thus the power spectrum $C_\ell(\eta_0)$ (34.35) at harmonic ℓ principally probes comoving scales $1/k$ that are $1/\ell$ times the comoving distance $\eta_0 - \eta_{\text{rec}}$ to recombination today,

$$\frac{1}{k} \sim \frac{\eta_0 - \eta_{\text{rec}}}{\ell}. \quad (34.37)$$

Physically, harmonic number ℓ probes angular scale $\theta \sim \pi/\ell$ on the sky, and the power spectrum at harmonic number ℓ probes comoving scale $\pi/k \sim (\eta_0 - \eta_{\text{rec}})\theta$ on the CMB sky.

34.3.3 Rapid recombination approximation to the CMB power spectrum

Modern, publicly available codes such as CAMB compute an entire model CMB power spectrum $C_\ell(\eta_0)$ in just a few seconds, which is amazingly fast. CAMB is tuned for speed, doing only enough calculations as are needed to achieve a desired accuracy. CAMB is written in a fast language, parallelized fortran 90. If you'd like to write a code that competes with CAMB in speed, expect to invest a substantial time developing it. It's more than just an exercise.

Meanwhile, the rapid recombination approximation, §34.2.4, offers a short-cut to computing the CMB power spectrum that at least captures qualitative features. The rapid recombination approximation effectively sidesteps step 5 of the numerical computation outlined in §30.8.

Exercise 34.1 CMB power spectrum in the instantaneous and rapid recombination approximations. Compute the CMB power spectrum $C_\ell(\eta_0)$ today, equation (34.36), in your choice of the instantaneous and rapid recombination approximations, equations (34.24) or (34.25), with source functions calculated in your choice of level of detail, simple, §30.7, hydrodynamic, §32.2, or full Boltzmann, §33.1). Discuss.

Solution. See Figures 34.7 and 34.8. I used the standard flat Λ CDM cosmological parameters given in §32.3, and the normalization of the power spectrum measured from Planck, equation (30.137). I used Mathematica to solve the evolutionary equations in the simple, hydrodynamic, Boltzmann approaches. But for the integral (34.35), I abandoned fighting Mathematica, and resorted to a publicly available implementation of

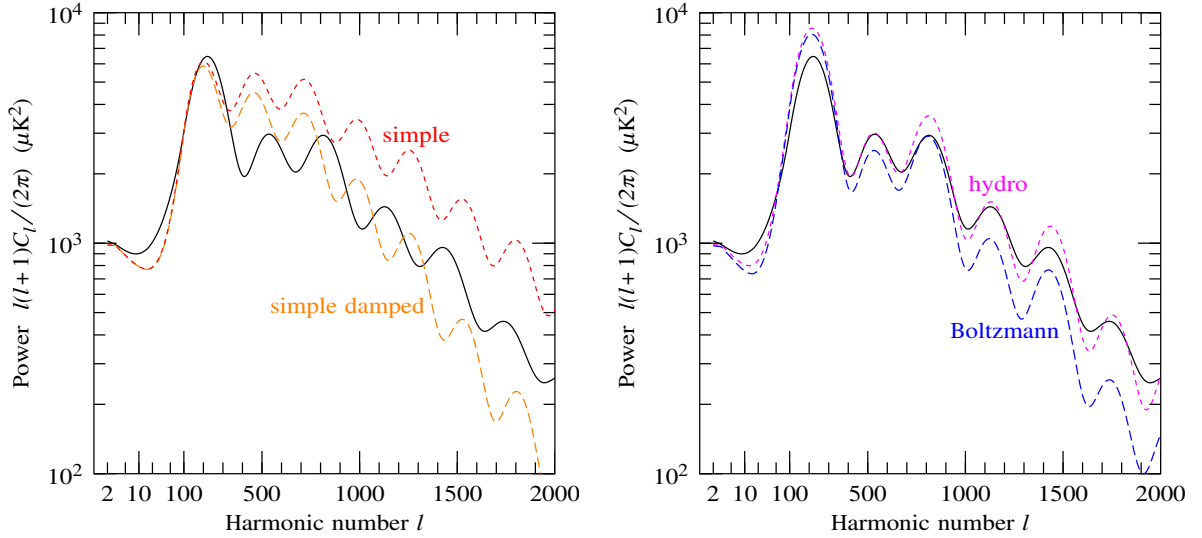


Figure 34.7 Model CMB power spectra computed in the rapid recombination approximation, equation (34.25), with source functions computed in: (left) the simple approximation, §30.7, without (short dashed line) and with (long dashed line) artificial damping, equations (30.58) and (30.59) with $\epsilon = 10^{-3}$; and (right) the hydrodynamic approximation (§32.2, short dashed line), and in a full Boltzmann treatment (§33.1, long dashed line) including photon and neutrino multipoles up to $\ell_{\max} = 16$. The solid (black) lines are a reference model power spectrum computed with CAMB. The CAMB spectrum is similar to that shown in Figure 10.3, but without refinements from reionization and lensing.

Bessel functions (Amos, 1986), and a cubic spline integration implemented in fortran. In Figure 34.8 (but not Figure 34.7) I added the late ISW contribution. The late ISW transfer function is not oscillatory, so its computation from integration of the derivative of the growth function over the line of sight, equations (34.17) and (34.19), is numerically straightforward. Comments:

1. The hydrodynamic and Boltzmann computations get the phasing of peaks more or less right. The phasing of peaks depends on the sound speed in the photon-baryon fluid, which depends on the baryon-to-photon density ratio. The agreement with the hydrodynamic and Boltzmann computations supports the standard model, where the baryonic density begins to become comparable to the photon density near recombination, equation (32.46). The simple approximation gets the phasing slightly wrong because it neglects baryons.
2. The overall angular location of the peaks is correctly reproduced. The overall angular location of peaks depends on geometry, that is, on the apparent angular size of comoving distances at recombination observed by astronomers on Earth today. The geometry depends on various cosmological parameters, notably the curvature Ω_k and the Hubble parameter H_0 today.
3. The power spectrum is roughly constant and dominated by the monopole at the largest scales, $\ell \lesssim 40$. This is the Sachs-Wolfe plateau, §34.5, a signature of a near-scale-invariant primordial power spectrum. The weak minimum at $\ell \approx 20$ results mainly from a cancellation between the monopole $\Theta_0 + \Psi$ and

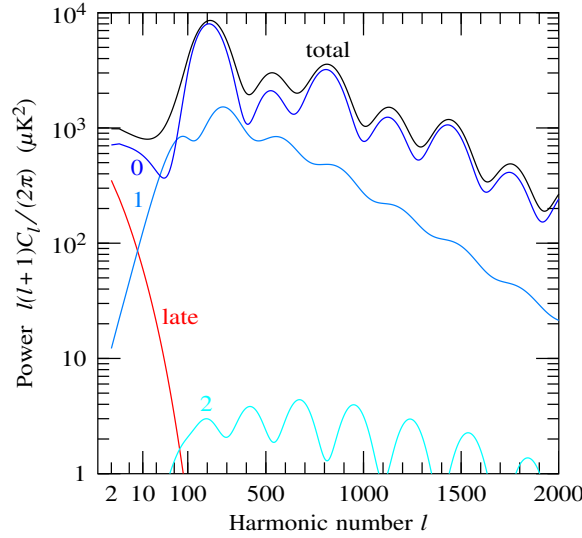


Figure 34.8 Monopole + early ISW (0), dipole (1), quadrupole (2), and late ISW contributions to the CMB power spectrum in the rapid recombination approximation with source functions computed in the hydrodynamic approximation (top model in the right panel of Figure 34.7). The monopole and early ISW are combined into a single curve, labelled 0, since they are highly correlated. To a good approximation, the total power spectrum is an incoherent sum of the monopole and dipole power spectra, the quadrupole contribution being quite small. There are sub-dominant cross-correlations between the various contributions, which are not plotted separately here, but which are included in the total (black line). The late ISW contribution is computed from integration of the derivative of the growth function over the line of sight, equations (34.17) and (34.19).

early ISW contributions, as might be expected from Figure 34.5. The late ISW effect contributes a small enhancement in power in the first several harmonics.

4. The even peaks are stronger than the odd peaks in the hydrodynamic and Boltzmann computations. The difference in strengths between even and odd peaks is caused by baryon loading, §32.10, in which the extra gravity generated by baryons in the oscillating photon-baryon fluid enhances even (compression) peaks and weakens odd (rarefaction) peaks. The simple approximation does not show the even-odd variation because it treats baryons as having negligible density.
5. The power spectrum $C_\ell(\eta_0)$ declines approximately exponentially with harmonic number ℓ . The decline arises partly from dissipative processes around the time of recombination, §32.7 and §32.8, and in part from the finite width of recombination, §34.2.4.

Exercise 34.2 CMB power spectra from CAMB. Compute model CMB power spectra from a publicly available code such as CAMB (google it). Vary the cosmological parameters. Compare to published measurements from Planck or other sources (google it). Formulate a question, and attempt to answer it. For example, what does the observed power spectrum say about:

1. non-baryonic cold dark matter;

2. baryons;
 3. photons;
 4. neutrinos;
 5. dark energy;
 6. curvature;
 7. the origin of fluctuations?
-

34.4 Observing CMB power

The power spectrum $C_\ell(\eta_0)$, equation (34.34), gives an expectation value for the variance (34.33) of CMB temperature fluctuations on the sky, which can be compared to observation. Isotropy predicts that $\Theta_{\ell m}(\eta_0, \mathbf{x}_0)$ with different ℓ or m should be uncorrelated, a prediction that can be tested by observation.

Inflation, which predicts that fluctuations are generated by quantum fluctuations of the scalar inflaton field that supposedly drove inflation, generically predicts a Gaussian distribution of fluctuations in the primordial curvature ζ . This in turn implies a Gaussian distribution of temperature fluctuations Θ as long as the fluctuations remain in the linear regime. The Gaussian distribution of temperature fluctuations $\Theta \equiv \delta T/T$ is characterized entirely by its variance, the power spectrum C_ℓ .

For each harmonic number ℓ , there are $2\ell + 1$ harmonics $\Theta_{\ell m}$ with the same ℓ but different m . Isotropy predicts that the expected variance is the same, C_ℓ , for each m . Thus one way to estimate the variance C_ℓ is to take

$$C_\ell(\text{est}) = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |\Theta_{\ell m}|^2. \quad (34.38)$$

The finite number $2\ell + 1$ of modes at each ℓ places a fundamental fractional uncertainty of $\approx 1/\sqrt{2\ell + 1}$ on the accuracy with which C_ℓ can be determined observationally. This fundamental limit, which arises from the finite size of the observable Universe, is called **cosmic variance**.

In practice there are numerous issues that complicate the measurement of the CMB power spectrum C_ℓ , including incomplete sky coverage, contamination by Earth glow, microwave foregrounds arising from galactic and extragalactic synchrotron radiation, dust, and free-free emission, and observational and detector noise and systematics of one sort or another.

34.5 Large-scale CMB fluctuations (Sachs-Wolfe effect)

The behaviour of the CMB power spectrum at the largest angular scales was first predicted by Sachs and Wolfe (1967), and is therefore called the **Sachs-Wolfe effect**, though why it should be called an effect is mysterious. The Sachs-Wolfe (SW) effect is distinct from the Integrated Sachs-Wolfe (ISW) effect. The ISW effect, ignored in this section, was considered in §34.2.2.

At scales much larger than the sound horizon at recombination, $k\eta_{s,\text{rec}} \ll 1$, the redshifted monopole fluctuation $\Theta_0(\eta_{\text{rec}}, \mathbf{k}) + \Psi(\eta_{\text{rec}}, \mathbf{k})$ at recombination is much larger than the dipole $\Theta_1(\eta_{\text{rec}}, \mathbf{k})$ or quadrupole $\Theta_2(\eta_{\text{rec}}, \mathbf{k})$, so only the monopole contributes materially to the temperature multipoles $\Theta_\ell(\eta_0, \mathbf{k})$ today. The redshifted monopole contribution to the temperature multipoles $\Theta_\ell(\eta_0, \mathbf{k})$ today is, from equation (34.17),

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = [\Theta_0(\eta_{\text{rec}}, \mathbf{k}) + \Psi(\eta_{\text{rec}}, \mathbf{k})] j_\ell[k(\eta_0 - \eta_{\text{rec}})] . \quad (34.39)$$

At superhorizon scales $k\eta_{\text{rec}} \ll 1$, the radiation monopole at the time η_{rec} of recombination is given by the superhorizon solution $\Theta_0 - \Phi = \zeta_\gamma$, equation (30.62b), so

$$\Theta_0(\eta_{\text{rec}}, \mathbf{k}) + \Psi(\eta_{\text{rec}}, \mathbf{k}) = \Psi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \Phi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \zeta_\gamma(\mathbf{k}) . \quad (34.40)$$

The CMB transfer function $T_\ell(\eta, k)$ is conventionally normalized to the primordial curvature fluctuation $\zeta(\mathbf{k})$, equation (34.20). For adiabatic fluctuations ζ is the same for all species; more generally, ζ could be different for different species. For definiteness, take the simple two-component matter plus radiation model of Chapter 30, where the superhorizon potential in the late matter-dominated regime is $\Phi(\text{late}) = -\frac{3}{5}\zeta_c$, equation (30.67), for both adiabatic and isocurvature initial conditions. In the approximation that recombination is in the matter-dominated regime (which is not quite true), and the scalar potentials are equal (which is again not quite true), so $\Psi(\eta_{\text{rec}}) + \Phi(\eta_{\text{rec}}) \approx 2\Phi_{\text{super}}(\text{late}) = -\frac{6}{5}\zeta_c$, the CMB transfer function for the radiation monopole at superhorizon scales, normalized to ζ_c , is

$$T_0(\eta_{\text{rec}}, k) = \frac{\Psi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \Phi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \zeta_\gamma(\mathbf{k})}{\zeta_c(\mathbf{k})} \approx -\frac{6}{5} + \frac{\zeta_\gamma(\mathbf{k})}{\zeta_c(\mathbf{k})} = \begin{cases} -\frac{1}{5} & \text{adiabatic} , \\ -\frac{6}{5} & \text{isocurvature} . \end{cases} \quad (34.41)$$

The monopole transfer function $T_0(\eta_{\text{rec}}, k)$ at recombination is thus approximately constant at superhorizon scales, although the value of the constant depends on the initial conditions.

At superhorizon scales, the CMB transfer function $T_\ell(\eta_0, k)$ in the ℓ th harmonic today is, from equation (34.39),

$$T_\ell(\eta_0, k) = T_0(\eta_{\text{rec}}, k) j_\ell[k(\eta_0 - \eta_{\text{rec}})] . \quad (34.42)$$

The resulting CMB angular power spectrum at superhorizon scales $k\eta_{s,\text{rec}} \ll 1$ is

$$C_\ell(\eta_0) = 4\pi T_0(\eta_{\text{rec}}, k)^2 \int_0^\infty j_\ell[k(\eta_0 - \eta_{\text{rec}})]^2 \Delta_\zeta^2(k) \frac{dk}{k} , \quad (34.43)$$

where $T_0(\eta_{\text{rec}}, k)$, being approximately constant, equation (34.41), has been taken outside the integral. If the primordial curvature power spectrum $\Delta_\zeta^2(k)$ is a power law with tilt n_s , equation (30.136), then the integral over the squared Bessel function can be done analytically, equation (34.56b), yielding

$$C_\ell(\eta_0) = 4\pi T_0(\eta_{\text{rec}}, k)^2 \Delta_\zeta^2[1/(\eta_0 - \eta_{\text{rec}})] U_{\ell;\ell}(n-1) . \quad (34.44)$$

For the particular case of a scale-invariant primordial power spectrum, $n = 1$, the CMB power spectrum C_ℓ at large scales today is given by

$$\frac{\ell(\ell+1)C_\ell(\eta_0)}{2\pi} = T_0(\eta_{\text{rec}}, k)^2 \Delta_\zeta^2[1/(\eta_0 - \eta_{\text{rec}})] \quad \text{if } n = 1 . \quad (34.45)$$

Thus the characteristic feature of a scale-invariant primordial power spectrum, $n = 1$, is that $\ell(\ell + 1)C_\ell$ should be approximately constant at the largest angular scales, $\ell \ll \eta_0/\eta_{\text{rec}}$. The normalization factor $1/(2\pi)$ converts to the power of large scale fluctuations in the potential at recombination. This is the reason that CMB folk routinely plot $\ell(\ell + 1)C_\ell/(2\pi)$ rather than C_ℓ .

34.6 Radiative transfer of neutrinos

Neutrinos decouple not at recombination, but rather after electron-positron annihilation at a redshift $1 + z \sim 10^9$. From that point neutrinos streamed freely. The horizon distance η_ν at neutrino decoupling relative to that at matter-radiation equality was $\eta_\nu/\eta_{\text{eq}} \sim 10^{-5}$. As with radiation, inflation predicts that initially the neutrino distribution was isotropic at superhorizon scales, with only a monopole mode present. But once a mode entered the horizon, without collisions to isotropize their distribution, freely streaming neutrinos could develop appreciable higher multipole moments, Figure 33.2. Prior to recombination, the neutrino quadrupole provided the dominant source for the difference $\Psi - \Phi$ between the scalar potentials, Figure 33.4. In Exercise 33.5 you discovered that the neutrino quadrupole causes a finite difference $\Psi - \Phi$ even in the superhorizon initial conditions, equation (33.97).

Observationally accessible scales in the CMB or in the clustering of matter are large compared to the horizon distance η_ν at neutrino decoupling. At such large scales, $k\eta_\nu \ll 1$, only the neutrino monopole $\mathcal{N}_0$ was present at neutrino decoupling. The neutrino analogue to the solution (34.17) of the radiative transfer equation is then

$$\begin{aligned} \mathcal{N}_\ell(\eta, \mathbf{k}) + \delta_{\ell 0}\Psi(\eta, \mathbf{k}) = & \int_0^\eta [\dot{\Psi}(\eta', \mathbf{k}) + \dot{\Phi}(\eta', \mathbf{k})] j_\ell[k(\eta' - \eta)] d\eta' \quad \text{ISW} \\ & + [\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k})] j_\ell(-k\eta) \quad \text{monopole}, \end{aligned} \quad (34.46)$$

which contains only Integrated Sachs-Wolfe and dipole terms. In equation (34.46), the time η_ν of neutrino decoupling has been replaced by zero, and the optical depth factor $e^{-\tau}$ omitted, since the neutrino decoupling scale is so much smaller than cosmological scales.

Equation (34.46) holds at any time η after neutrino decoupling, as long as the neutrinos remain relativistic. Neutrino oscillation data suggest that at least 2 of the 3 neutrino types are massive, with masses at least 0.01 eV and 0.05 eV (see §10.25). Such neutrinos would have become nonrelativistic at a redshift of $1 + z \approx 60$ and 300 respectively. However, all 3 neutrino types were relativistic prior to and at recombination, when the physics of dark matter and the photon-baryon fluid was imposing its imprint on the CMB.

34.6.1 Truncating the neutrino Boltzmann hierarchy

The integral solution (34.46) provides one way to compute neutrino multipoles of arbitrary order. The solution is equivalent to solving the entire collisionless Boltzmann hierarchy of differential equations for neutrinos. However, it is more common for computer codes to solve for neutrino multipoles using the Boltzmann hierarchy truncated in a suitable fashion. The strategy of setting multipoles above some maximum harmonic

to zero does not work well for neutrinos, because free-streaming allows neutrinos to develop higher order multipoles comparable to the monopole and dipole. An alternative strategy for truncating the neutrino hierarchy, described immediately following, was proposed by Ma and Bertschinger (1995).

Spherical Bessel functions are related by

$$j_\ell(y) - \frac{2\ell+3}{y}j_{\ell+1}(y) + j_{\ell+2}(y) = 0. \quad (34.47)$$

This motivates considering the combination $(\mathcal{N}_\ell + \Psi\delta_{\ell 0}) + (2\ell+3)\mathcal{N}_{\ell+1}/y + \mathcal{N}_{\ell+2}$, with $y = k\eta$, of neutrino multipoles, which has the property that the monopole term from the second line of equation (34.46) vanishes. The ISW term on the first line of equation (34.46) gives a non-vanishing contribution to the combination, which is, with the identity $1/y = y'/[y(y'-y)] - 1/(y'-y)$,

$$\mathcal{N}_\ell + \delta_{\ell 0}\Psi + \frac{2\ell+3}{y}\mathcal{N}_{\ell+1} + \mathcal{N}_{\ell+2} = (2\ell+3) \int_0^y \frac{\partial[\Psi(y') + \Phi(y')]}{\partial y'} \frac{y'}{y} \frac{j_{\ell+1}(y'-y)}{(y'-y)} dy'. \quad (34.48)$$

The integrand on the right hand side of equation (34.48) is everywhere finite, and for $y \gg y'$ is of order y'/y^2 times the integrand of the ISW integral in equation (34.46). In the actual case, $\Psi + \Phi$ varies rapidly at horizon-crossing, $y' \sim 1$, but subsequently varies slowly, Figure 33.2. In this case the integral on the right hand side of equation (34.50) is small compared to $\mathcal{N}_{\ell+2}$ for $y \gg 1$. The integral is also small for $y \ll \ell+1$, since $j_{\ell+1}(y'-y)/(y'-y) \approx (y'-y)^\ell/(2\ell+1)!!$ for $0 \leq y' \leq y \ll \ell+1$. The approximation that the integral is small is better for larger harmonic number ℓ .

If the integral on the right hand side of equation (34.48) is neglected, which becomes an increasingly good approximation at higher ℓ , then

$$\mathcal{N}_{\ell+2} \approx -(\mathcal{N}_\ell + \delta_{\ell 0}\Psi) - \frac{2\ell+3}{k\eta}\mathcal{N}_{\ell+1}. \quad (34.49)$$

Ma and Bertschinger (1995) proposed truncating the neutrino Boltzmann hierarchy by using the approximation (34.49) at some suitably high harmonic number ℓ . The approximation is worst around epochs where $\Psi + \Phi$ varies rapidly, such as around horizon-crossing, $k\eta \sim 1$.

34.6.2 Approximate neutrino quadrupole

The neutrino quadrupole $\mathcal{N}_2$ is of special interest because it is a principal source for the difference $\Psi - \Phi$ in scalar potentials. For the quadrupole, the approximation (34.49) is

$$\mathcal{N}_2 \approx -(\mathcal{N}_0 + \Psi) - \frac{3\mathcal{N}_1}{k\eta}. \quad (34.50)$$

The approximation (34.50) is not adequate for precision modelling, but it provides the basis for the approximation of neutrinos as an imperfect fluid, equation (32.11). It is a better approximation than simply setting the neutrino quadrupole to zero, $\mathcal{N}_2 = 0$. The approximation (34.50) leads to a second order differential equation for the neutrino monopole, equation (32.91), which allows the behaviour of neutrinos to be explored qualitatively, Exercise 32.7.

Exercise 34.3 Cosmic Neutrino Background.

1. Is there a Cosmic Neutrino Background? Think about whether neutrinos are relativistic or nonrelativistic today.
2. Suppose that one neutrino is relativistic. Calculate the power spectrum of the Cosmic Neutrino Background for that neutrino in the approximation that the ISW contribution is negligible.
3. What is the effect of the ISW contribution resulting from the change in the potential when neutrinos entered the horizon in the radiation-dominated regime?

Solution.

1. Neutrinos with the masses of (at least) 0.01 eV and 0.05 eV suggested by neutrino oscillation data (see §10.25) would have become nonrelativistic at a redshift of $1+z \approx 60$ and 300 respectively, whereupon they would start to cluster like dark matter and baryons, rather than continuing to stream like cosmic background photons in more or less straight lines into astronomers' telescopes. There remains the possibility that one of the neutrino types may be light enough, $m_\nu \lesssim 10^{-4}$ eV, to be relativistic today. Such a relativistic neutrino would produce a background today that is an imprint of fluctuations in the Universe at the time of neutrino decoupling.
2. For a light, relativistic neutrino, the multipole moments of the cosmic background today are given by equation (34.46) with $\eta = \eta_0$. Without the ISW term, only the monopole term remains,

$$\mathcal{N}_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = [\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k})] j_\ell(k\eta_0) . \quad (34.51)$$

The initial value is the superhorizon result

$$\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k}) = \Psi(0) + \Phi(0) + \zeta_\nu . \quad (34.52)$$

The neutrino power spectrum is proportional to the photon Sachs-Wolfe power spectrum (34.43), with constant of proportionality

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} = \left(\frac{\Psi(0) + \Phi(0) + \zeta_\nu}{\Psi_{\text{super}}(\eta_{\text{rec}}) + \Phi_{\text{super}}(\eta_{\text{rec}}) + \zeta_\gamma} \right)^2 . \quad (34.53)$$

In the approximation that recombination is in the matter-dominated regime (which is not quite true), and the scalar potentials are equal (which is not quite true thanks to neutrinos), the potentials at recombination are approximately the late time potentials given by equations (30.67), so

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} \approx \left(\frac{-\frac{3}{5}\zeta_r + \zeta_\nu}{-\frac{6}{5}\zeta_c + \zeta_\gamma} \right)^2 . \quad (34.54)$$

Inflation generically predicts adiabatic fluctuations with ζ s of all species the same, in which case

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} \approx \left(\frac{5}{3} \right)^2 . \quad (34.55)$$

3. An ISW effect results from the change in the potential at horizon-crossing for modes that entered the horizon during the radiation-dominated era. The potential $\Psi(y) + \Phi(y)$ is a universal function of $y \equiv k\eta$ during horizon-crossing in the radiation-dominated era, independent of k . The ISW integral yields a result that looks like the spherical Bessel function of the monopole contribution on the second line of equation (34.46), but with a different amplitude and phase. The net result is a power spectrum that again looks like the Sachs-Wolfe power spectrum, but with a (somewhat) different amplitude than the large-scale power spectrum, whose modes entered the horizon in the matter-dominated regime.

34.7 Appendix: Integrals over spherical Bessel functions

Two useful integrals over spherical Bessel functions are

$$U_\ell(z) \equiv \int_0^\infty j_\ell(y) y^z \frac{dy}{y} = \frac{2^{z-2} \sqrt{\pi} \Gamma\left[\frac{1}{2}(\ell+z)\right]}{\Gamma\left[\frac{1}{2}(\ell+3-z)\right]}, \quad (34.56a)$$

$$U_{\ell;\ell'}(z) \equiv \int_0^\infty j_\ell(y) j_{\ell'}(y) y^z \frac{dy}{y} = \frac{2^{z-3} \pi \Gamma(2-z) \Gamma\left[\frac{1}{2}(\ell+\ell'+z)\right]}{\Gamma\left[\frac{1}{2}(\ell+\ell'+4-z)\right] \Gamma\left[\frac{1}{2}(\ell-\ell'+3-z)\right] \Gamma\left[\frac{1}{2}(\ell'-\ell+3-z)\right]}, \quad (34.56b)$$

where $\Gamma(z)$ is the Gamma function. The integrals satisfy the recurrence relations

$$\begin{aligned} U_\ell(z) &= (\ell-2+z) U_{\ell-1}(z-1) \\ &= \frac{\ell-2+z}{\ell+1-z} U_{\ell-2}(z), \end{aligned} \quad (34.57a)$$

$$\begin{aligned} U_{\ell;\ell'}(z) &= \frac{\ell+\ell'-2+z}{\ell+\ell'+2-z} U_{\ell-1;\ell'-1}(z) \\ &= \frac{(\ell+\ell'-2+z)(\ell-\ell'-3+z)}{2(z-2)} U_{\ell-1;\ell'}(z-1) \\ &= \frac{(\ell+\ell'-2+z)(\ell-\ell'-3+z)}{(\ell+\ell'+2-z)(\ell'-\ell-1+z)} U_{\ell-2;\ell'}(z). \end{aligned} \quad (34.57b)$$

Cosmological perturbations including polarization

Well before recombination, frequent collisions drive photons into thermodynamic equilibrium. In thermodynamic equilibrium, the photon distribution is unpolarized. But, as will be seen in §35.10, photons scattering off electrons become linearly polarized. The CMB bears the imprint of polarization generated near the surface of last scattering.

Polarization produces distinct *E*-mode (electric parity) and *B*-mode (magnetic parity) fluctuations, §35.6.2. The *B*-mode fluctuation can be generated only by vector or tensor, not scalar, gravitational potential fluctuations. The *B*-mode polarization has opposite parity to, and can thereby be observationally distinguished from, the much stronger unpolarized and polarized scalar fluctuations. Thus the *B*-mode polarization provides a clean window to gravitational waves generated during inflation in the very early Universe. A detection of *B*-mode polarization was initially claimed by the BICEP2 collaboration (Ade et al., 2014), but subsequent cross-comparison between BICEP2 and Planck data suggests that the detected polarization may have been a galactic foreground from dust aligned by the galactic magnetic field (Planck Collaboration, 2015). If a cosmological signal of *B*-mode polarization is detected in the future, it would present a remarkable observation of physics at near-Planck energies far exceeding those accessible in earthly particle accelerators.

35.1 Photon polarization

Photons have spin one. They have two distinct spin eigenstates, or polarizations, transverse to the photon direction of motion. A general spin eigenstate of a photon is a complex linear combination of the two spin states. Any pair of transverse spin states can be chosen as a basis. If the photon direction of motion is along the 3-direction (*z*-direction), then the two basis spin states can for example be taken to be linear polarizations γ_1 and γ_2 along the 1- and 2-directions (*x*- and *y*-directions) transverse to the 3-direction. An elegant choice of basis spin states are right- and left-circular polarizations $\gamma_+ \equiv (\gamma_1 + i\gamma_2)/\sqrt{2}$ and $\gamma_- \equiv (\gamma_1 - i\gamma_2)/\sqrt{2}$, equations (41.1), in which the spin is respectively aligned (+) and anti-aligned (−) with the photon direction of motion γ_3 . The condition of right- or left-circular polarization, aligned or anti-aligned with the direction of motion, is Lorentz invariant, unchanged by any Lorentz transformation. The general spin eigenstate of a photon is described by a complex **polarization vector** $\mathbf{a}$, a complex linear combination of right- and

left-handed eigenstates,

$$\mathbf{a} \equiv a^a \boldsymbol{\gamma}_a = a^+ \boldsymbol{\gamma}_+ + a^- \boldsymbol{\gamma}_- . \quad (35.1)$$

The polarization vector $\mathbf{a}$ is transverse, that is, it is orthogonal both to the time axis $\boldsymbol{\gamma}_0$ and to the direction $\boldsymbol{\gamma}_3$ of the photon's direction of motion,

$$\mathbf{a} \cdot \boldsymbol{\gamma}_0 = \mathbf{a} \cdot \boldsymbol{\gamma}_3 = 0 . \quad (35.2)$$

According to the rules of quantum mechanics, the squared amplitude is the probability of the photon, which for a single photon is one,

$$|a^+|^2 + |a^-|^2 = 1 . \quad (35.3)$$

The squared individual amplitudes $|a^+|^2$ and $|a^-|^2$ of the polarization vector (35.1) represent the probabilities of observing the photon to have polarization $\boldsymbol{\gamma}_+$ or $\boldsymbol{\gamma}_-$. For example, if a photon with polarization $\mathbf{a}$ is sent through a right-circularly polarized filter, then the photon will be transmitted with probability $|a^+|^2$, and the transmitted photon will then be 100% right-circularly polarized. The total probability of the spin states of the photon is one, equation (35.3). There is a Lorentz-invariant operation of conjugation that leaves orthonormal vectors unchanged, equation (41.115), but flips the spin indices $+\leftrightarrow -$, equation (41.116). The conjugate $\bar{\mathbf{a}}$ of the polarization vector is

$$\bar{\mathbf{a}} = a^{+*} \bar{\boldsymbol{\gamma}}_+ + a^{-*} \bar{\boldsymbol{\gamma}}_- = a^{+*} \boldsymbol{\gamma}_- + a^{-*} \boldsymbol{\gamma}_+ . \quad (35.4)$$

The normalization condition (35.3) can then be written

$$\bar{\mathbf{a}} \cdot \mathbf{a} = 1 . \quad (35.5)$$

More generally, if a photon has polarization $\mathbf{a}$, then the probability $P_{\mathbf{a}'}$ of observing it to have polarization $\mathbf{a}'$ is, by the rules of quantum mechanics,

$$P_{\mathbf{a}'} = |\bar{\mathbf{a}'} \cdot \mathbf{a}|^2 . \quad (35.6)$$

A photon in a pure $\boldsymbol{\gamma}_+$ eigenstate (i.e. with polarization vector $e^{-i\phi} \boldsymbol{\gamma}_+$ where $e^{-i\phi}$ is some arbitrary phase factor) is right-circularly polarized, while a photon in a pure $\boldsymbol{\gamma}_-$ eigenstate is left-circularly polarized. A photon that is a superposition of equal magnitudes of right- and left-circular polarizations is said to be linearly polarized. For example a photon with polarization vector $\mathbf{a}_1 = e^{-i\phi}(\boldsymbol{\gamma}_+ + \boldsymbol{\gamma}_-)/\sqrt{2} = e^{-i\phi} \boldsymbol{\gamma}_1$ is linearly polarized in the 1-direction (x -direction), while a photon with polarization vector $\mathbf{a}_2 = e^{-i\phi}(\boldsymbol{\gamma}_+ - \boldsymbol{\gamma}_-)/(\sqrt{2}i) = e^{-i\phi} \boldsymbol{\gamma}_2$ is linearly polarized in the 2-direction (y -direction). More generally, a photon with polarization vector

$$\mathbf{a}_\chi = e^{-i\phi} \frac{e^{-i\chi} \boldsymbol{\gamma}_+ + e^{i\chi} \boldsymbol{\gamma}_-}{\sqrt{2}} \quad (35.7)$$

is linearly polarized along a direction rotated right-handedly by angle χ from the 1-axis. Polarization angles $\chi = 0$ and $\pi/2$ correspond to photons linearly polarized along respectively the 1- and 2-directions. A polarization angle of $\chi = \pi$ flips the sign of $\mathbf{a}_\chi$, equivalent to changing its phase ϕ , so the polarization angle χ is determined only modulo π .

The most general polarization vector of a photon is elliptically polarized, a superposition of unequal non-zero magnitudes of right and left polarizations,

$$\mathbf{a}_{\chi, \epsilon} = e^{-i\phi} \frac{e^{-i\chi} \cos \epsilon \boldsymbol{\gamma}_+ + e^{i\chi} \sin \epsilon \boldsymbol{\gamma}_-}{\sqrt{2}} . \quad (35.8)$$

The elliptic angle ϵ varies from $\epsilon = 0$ for pure right-circular polarization, to $\epsilon = \pi/4$ for linear polarization, to $\epsilon = \pi/2$ for pure left-circular polarization. The polarization angle χ is the angle by which the polarization ellipse is rotated right-handedly from the 1-axis.

As is usual in quantum mechanics, the phase $e^{-i\phi}$ of a polarization vector $\mathbf{a}$ is by itself unobservable; only probabilities (35.6) are observable.

Concept question 35.1 Relation of the polarization vector to the electromagnetic potential.

How is the polarization vector $\mathbf{a}$ related to the electromagnetic potential $\mathbf{A}$? **Answer.** As discussed in §27.6, the gauge freedom of electromagnetism means that only 3 of the 4 components of the electromagnetic potential $\mathbf{A}$ are gauge-invariant, equations (27.37), and only the 2 vector (i.e. transverse) components $\mathbf{A}_\perp$ of the electromagnetic potential describe propagating waves, equation (27.40). Plane-wave solutions propagating in the $\boldsymbol{\gamma}_3$ -direction (z -direction) are functions of $\mathbf{A}(t - z)$ with $\mathbf{A}$ transverse. The associated electric and magnetic fields are, equations (27.38),

$$\mathbf{E} = -\dot{\mathbf{A}} , \quad \mathbf{B} = \boldsymbol{\gamma}_3 \wedge \mathbf{E} . \quad (35.9)$$

The electric and magnetic fields of a propagating wave are transverse and orthogonal to each other. In quantum field theory, the electromagnetic potential $\mathbf{A}$ is fundamentally complex, §45. Monochromatic waves of positive angular frequency ω propagating forwards in time in the z -direction in Minkowski space are described by a complex potential

$$\mathbf{A} = \mathbf{A}_0 e^{-i\omega(t-z)} , \quad (35.10)$$

where $\mathbf{A}_0$ is a constant complex transverse vector. The mean-squared potential is

$$\langle \bar{\mathbf{A}} \cdot \mathbf{A} \rangle = \langle \bar{\mathbf{A}}_0 \cdot \mathbf{A}_0 \rangle = A^2 . \quad (35.11)$$

The polarization vector $\mathbf{a}$, which has unit magnitude, equation (35.3), equals the constant potential $\mathbf{A}_0$ scaled to unit magnitude, modulo a possible phase factor,

$$\mathbf{a} = \frac{\mathbf{A}_0}{A} \text{ mod phase} . \quad (35.12)$$

In classical electromagnetism it is possible and conventional to work with real quantities only, since the phase of the electromagnetic potential $\mathbf{A}$ is by itself unobservable. That is, although Maxwell's equations admit complex wave solutions such as (35.10), classical electromagnetism does not require them. In classical electromagnetism, a real wave is the real part of a complex wave (35.10), multiplied by $\sqrt{2}$ to get the mean-squared amplitude right. Thus in classical electromagnetism real linearly polarized waves oscillating

in respectively the 1- and 2-directions are

$$\mathbf{A}_{\frac{1}{2}} = \sqrt{2}A\boldsymbol{\gamma}_{\frac{1}{2}} \operatorname{Re} e^{-i\omega(t-z)} = \sqrt{2}A\boldsymbol{\gamma}_{\frac{1}{2}} \cos \omega(t-z) , \quad (35.13)$$

while real right- and left-circularly polarized waves are

$$\mathbf{A}_{\pm} = \sqrt{2}A \operatorname{Re} \boldsymbol{\gamma}_{\pm} e^{-i\omega(t-z)} = A \operatorname{Re} ((\boldsymbol{\gamma}_1 \pm i\boldsymbol{\gamma}_2) e^{-i\omega(t-z)}) = A (\boldsymbol{\gamma}_1 \cos \omega(t-z) \pm \boldsymbol{\gamma}_2 \sin \omega(t-z)) . \quad (35.14)$$

Note that the “complex conjugate” of $\boldsymbol{\gamma}_m$ in this context strictly refers to the Lorentz-invariant conjugate $\bar{\boldsymbol{\gamma}}_m$, equation (41.115), with respect to which $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are both real. For each of the linearly polarized waves (35.13), or circularly polarized waves (35.14), the mean-squared potential is

$$\langle \mathbf{A}_{\frac{1}{2}}^2 \rangle = 2A^2 \langle \cos^2 \omega t \rangle = A^2 , \quad \langle \mathbf{A}_{\pm}^2 \rangle = A^2 . \quad (35.15)$$

Evidently, when dealing with polarization, it is simpler to work with complex (quantum mechanical) waves than with real (classical) waves.

35.2 Photon density matrix

It is necessary to distinguish between photons in mixed states and mixtures of photons in different states. For example, a system consisting of photons all in a linearly polarized state (35.7) is not the same as a mixture of purely right-handed and purely left-handed photons. The systems can be distinguished experimentally by passing the photons through polarizers.

To deal with these distinctions, a statistical ensemble of photons in various polarization states must be described by a **density matrix**. Suppose that the system consists of photons in pure polarization states $\mathbf{a}$ with real occupation numbers $f(\mathbf{a})$. Then the density matrix $\mathbf{f}$ may be defined by the tensor

$$\mathbf{f} \equiv \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) \mathbf{a} \otimes \bar{\mathbf{a}} . \quad (35.16)$$

In any basis $\boldsymbol{\gamma}_a$, the density matrix is

$$\mathbf{f} = \sum_{a\bar{b}} f^{a\bar{b}} \boldsymbol{\gamma}_a \wedge \bar{\boldsymbol{\gamma}}_{\bar{b}} , \quad (35.17)$$

with components

$$f^{a\bar{b}} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) a^a \bar{a}^{b*} . \quad (35.18)$$

A conjugated index $\bar{b}$ signifies the index of the conjugated vector, $\bar{\boldsymbol{\gamma}}_{\bar{b}} = \bar{\boldsymbol{\gamma}}_b$. For orthonormal vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, conjugated vectors are themselves, so $\bar{1} = 1$ and $\bar{2} = 2$; while for chiral vectors $\boldsymbol{\gamma}_+$ and $\boldsymbol{\gamma}_-$, conjugation flips spin, so $\bar{+} = -$ and $\bar{-} = +$. The density matrix is Hermitian,

$$(f^{a\bar{b}})^* = f^{b\bar{a}} . \quad (35.19)$$

Its trace is

$$\sum_a f^{a\bar{a}} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) \sum_a a^a a^{a*} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) , \quad (35.20)$$

which counts the total number of photons.

If the system of photons is measured along the polarization direction $\mathbf{a}$, then the occupation number of photons with that polarization will be found to be, in accordance with equation (35.6),

$$f(\mathbf{a}) = f^{a\bar{b}} a_a a_b^* . \quad (35.21)$$

The conjugate $\bar{\mathbf{f}}$ of the density matrix $\mathbf{f}$ is

$$\bar{\mathbf{f}} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) \bar{\mathbf{a}} \otimes \mathbf{a} = \sum_{\bar{a}b} f^{\bar{a}b} \gamma_{\bar{a}} \wedge \gamma_b , \quad (35.22)$$

whose components are

$$f^{\bar{a}b} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) a^{a*} a^b . \quad (35.23)$$

35.2.1 Physical interpretation of the photon density matrix

Since the complex 2×2 photon density matrix $f^{a\bar{b}}$ is Hermitian, equation (35.19), it is diagonalizable with 2 real eigenvalues. The form (35.16) of the density matrix ensures that the matrix is positive definite, that is, its eigenvalues are both non-negative. If only one eigenvalue is positive, and the other is zero, then the density matrix represents a pure state. The most general pure state consists of photons all in the same (in general elliptically polarized) state. The most general impure state is equivalent to a mixture of photons in two orthogonal (in general elliptically polarized) states. In thermodynamic equilibrium, the two eigenvalues are equal, and the density matrix describes a mixture of equal numbers of photons in any pair of orthogonal states.

With respect to a circularly polarized (Newman-Penrose) basis $\gamma_{\pm}$, the 2×2 density matrix $f^{a\bar{b}}$ is

$$f^{a\bar{b}} = \begin{pmatrix} f^{+-} & f^{++} \\ f^{--} & f^{-+} \end{pmatrix} . \quad (35.24)$$

The components $f^{a\bar{b}}$ comprise two real scalar (spin 0) components f^{+-} and f^{-+} , and a complex tensor (spin 2) component $f^{++} = f^{--*}$. The trace of the density matrix (35.24) counts the total number of photons, equation (35.20). The unpolarized scalar occupation number f defined earlier, equation (31.28), equals one when there is one photon in either of the two polarization states, so the trace equals twice the unpolarized occupation number f ,

$$\sum_a f^{a\bar{a}} = f^{+-} + f^{-+} = \sum_{\text{photons } \mathbf{a}} f(\mathbf{a}) = 2f . \quad (35.25)$$

Conjugation flips chiral indices $+\leftrightarrow -$, so the components of the conjugate density matrix are

$$f^{\bar{a}b} = \begin{pmatrix} f^{-+} & f^{--} \\ f^{++} & f^{+-} \end{pmatrix}. \quad (35.26)$$

Note that conjugation here signifies the Lorentz-invariant operation described in §41.8.4, which is related to, but *not* the same as, complex conjugation. In particular, the components f^{+-} and f^{-+} are conjugates of each other, but they are nevertheless real.

The spin-0 component f^{+-} (the coefficient of $\gamma_+ \otimes \bar{\gamma}_+$) measures the intensity of right-circularly polarized light, while the other spin-0 component f^{-+} (the coefficient of $\gamma_- \otimes \bar{\gamma}_-$) measures the intensity of left-circularly polarized light. The sum $2f = f^{+-} + f^{-+}$ measures the total intensity of light in both polarizations, while the difference $f^{+-} - f^{-+}$ measures the net circularly polarized intensity, the excess of right- over left-circular polarized intensities.

The complex spin-2 component f^{++} measures the degree of linear polarization of the light. A photon linearly polarized in the direction χ , equation (35.7), contributes a density matrix

$$\mathbf{a}_\chi \otimes \bar{\mathbf{a}}_\chi = \frac{1}{2} (\gamma_+ \otimes \gamma_- + \gamma_- \otimes \gamma_+) + \frac{1}{2} e^{-2i\chi} \gamma_+ \otimes \gamma_+ + \frac{1}{2} e^{2i\chi} \gamma_- \otimes \gamma_- , \quad (35.27)$$

whose components are

$$f_\chi^{a\bar{b}} = \frac{1}{2} \begin{pmatrix} 1 & e^{-2i\chi} \\ e^{2i\chi} & 1 \end{pmatrix}. \quad (35.28)$$

The trace of the density matrix is one, which is as it should be for a single photon. Twice the amplitude of f^{++} gives the degree of linear polarization, which here is one (100% linearly polarized), while the phase 2χ of f^{++} measures the angle χ by which the direction of polarization is rotated right-handedly from the 1-axis (x -axis).

In the cosmological case under consideration, Thomson scattering generates linear but not circular polarization. In this case the photon density matrix is

$$f^{a\bar{b}} = \begin{pmatrix} f & f^{++} \\ f^{--} & f \end{pmatrix}, \quad (35.29)$$

with f the unpolarized occupation number. The equality of the diagonal elements of the density matrix, $f^{+-} = f^{-+} = f$, expresses the absence of circular polarization.

Concept question 35.2 Elliptically polarized light. Can a beam of elliptically polarized light be distinguished from a sum of beams of linearly polarized and circularly polarized light? **Answer.** Yes. A beam containing photons all in the same elliptically polarized state is in a pure state, which is not equivalent to any sum of beams of different polarizations. In a beam in a pure state, 100% of the photons will pass through a matched filter, whereas in a beam in a mixed state some photons will be passed through and some will be absorbed by the matched filter.

35.2.2 Relation to Stokes parameters

The 4 components of the polarization density matrix $f^{a\bar{b}}$ are related to the 4 conventional real Stokes parameters I , V , Q , and U by

$$2f = f^{+-} + f^{-+} = I , \quad (35.30a)$$

$$f^{+-} - f^{-+} = V , \quad (35.30b)$$

$$f^{++} + f^{--} = Q , \quad (35.30c)$$

$$f^{++} - f^{--} = iU . \quad (35.30d)$$

The Stokes parameters here are normalized so that the total intensity I measures the total occupation number $2f$, the trace of the density matrix. Stokes parameters can be normalized in other ways, whatever may be convenient. Some of the ways that astronomers normalize intensity are described in the paragraph containing equation (1.80).

35.3 Temperature fluctuation for polarized photons

Previously, the perturbation to the unpolarized scalar occupation number f was expressed in terms of the temperature fluctuation, $\Theta \equiv \delta T/T$, equation (33.38). The temperature fluctuation $\Theta \equiv \Theta^{a\bar{b}} \gamma_a \otimes \bar{\gamma}_b$ including polarization can be defined similarly in terms of the density matrix $f^{a\bar{b}}$, equation (35.18), which is the generalization of the occupation number to include polarization,

$$f^{a\bar{b}} \equiv \frac{\partial f_\gamma}{\partial \ln T} \Theta^{a\bar{b}} . \quad (35.31)$$

The trace of $\Theta^{a\bar{b}}$ is twice the scalar temperature fluctuation, $\sum_a \Theta^{a\bar{a}} = 2\Theta$. The trace-free part of $\Theta^{a\bar{b}}$ describes the polarized temperature fluctuation.

It is conventional in cosmology, and elsewhere in physics, to signify the components of the photon density matrix using a spin index s , positioned to the left of the symbol to distinguish it from harmonic indices ℓm ,

$${}_0\Theta \equiv \Theta^{+-} , \quad {}_{-0}\Theta \equiv \Theta^{-+} , \quad {}_2\Theta \equiv \Theta^{++} , \quad {}_{-2}\Theta \equiv \Theta^{--} . \quad (35.32)$$

In the cosmological situation being considered, Thomson scattering generates linear but not circular polarization, with the consequence that the two spin 0 components ${}_0\Theta$ are equal, and equal to the unpolarized temperature fluctuation Θ ,

$${}_0\Theta = {}_{-0}\Theta = \Theta . \quad (35.33)$$

The spin index s signifies how the polarized temperature fluctuation varies, and is to be distinguished from harmonic indices ℓm . A temperature fluctuation ${}_s\Theta_{\ell m}$, equation (35.37), is the coefficient of an eigenmode that varies as $D_{\ell m s} \propto e^{-is\chi}$ under a right-handed rotation by angle χ about the photon's direction $\hat{\mathbf{p}}$ of motion, and as $D_{\ell m s} \propto e^{-im\phi}$ under a right-handed rotation by angle ϕ about the direction $\hat{\mathbf{k}}$ of the wavevector of the fluctuation.

35.4 Summary of equations including polarization

This section summarizes the coupled Boltzmann and Einstein equations needed to compute linear cosmological fluctuations including photon polarization.

Polarization involves not only scalar ($m = 0$) but also vector ($m = \pm 1$) and tensor ($m = \pm 2$) fluctuations ${}_s\Theta_{\ell m}$. The hierarchy of Boltzmann and Einstein equations for different m are decoupled from each other, so scalar, vector, and tensor equations may be calculated separately. Symmetry between positive and negative m means that in practice equations need be solved only for positive $m = 0, 1$, and 2 . Vector ($m = 1$) modes are commonly treated as being negligible, for the reasons given at the end of this section. Thomson scattering couples unpolarized $\Theta_{\ell m}$ and electric polarized $E_{\ell m}$ photon multipoles, equations (35.68c) and (35.68d). The polarized photon Boltzmann equations (35.45b) and (35.45c) couple the electric $E_{\ell m}$ and magnetic $B_{\ell m}$ parts of the polarized multipoles.

The Boltzmann equations for nonbaryonic dark matter and for baryons are equations (35.49) and (35.48), generalizing the scalar matter equations (33.1) and (33.2).

The Boltzmann equations for polarized photons are given by equations (35.45), with gravitational redshift source terms $G_{\ell m}$ (not to be confused with the Einstein tensor) given by equations (35.41), and Thomson-scattering collision terms $C[\Theta_{\ell m}]$, $C[E_{\ell m}]$, and $C[B_{\ell m}]$ given by equations (35.68). These generalize the scalar Boltzmann equations (33.81) for unpolarized photons.

The Boltzmann equations for neutrinos are equations (35.47), generalizing the scalar neutrino equations (33.91).

The Boltzmann hierarchies for photons and neutrinos may be truncated as described in §35.10.1.

Scalar, vector, and tensor Einstein equations are equations (33.7), (35.52) and (35.53).

Vector and tensor gravitational potentials $W_{\pm}$ and $h_{\pm\pm}$ are in general complex (with $W_- = W_+^*$ and $h_{--} = h_{++}^*$). Linear vector and tensor fluctuations (of all species) are proportional to the initial amplitudes $W_{\pm}(0)$ and $h_{\pm\pm}(0)$. Therefore in numerical calculations the initial amplitudes $W_{\pm}(0)$ and $h_{\pm\pm}(0)$ can be taken to be real, any phase factor being absorbed into a normalization factor. The phase factor cancels in power spectra, equation (36.25). If the initial amplitudes $W_{\pm}(0)$ and $h_{\pm\pm}(0)$ are real, then the coupled Boltzmann and Einstein equations ensure that $W_{\pm}$ and $h_{\pm\pm}$ and the photon multipoles $\Theta_{\ell m}$, $E_{\ell m}$, and $B_{\ell m}$ remain real, as do matter and neutrino multipoles. Since the polarized photon multipoles $E_{\ell m}$ and $B_{\ell m}$ are real, it can be convenient numerically to combine them into the complex polarized multipoles ${}_2\Theta_{\ell m} = E_{\ell m} + iB_{\ell m}$, and to solve a complex polarized Boltzmann equation whose left hand side is the complex expression (35.39). Thomson scattering couples the unpolarized fluctuation $\Theta_{\ell m}$ only to the electric part, that is, the real part, of the polarized fluctuation ${}_2\Theta_{\ell m} = E_{\ell m} + iB_{\ell m}$.

Collisions (before neutrino decoupling in the case of neutrinos, and before recombination in the case of photons) tend to drive initial vector ($|m| = 1$) and tensor ($|m| = 2$) multipoles of all particle species to zero, §35.11. Vector gravitational fluctuations $W_{\pm}$ tend to redshift to zero, equation (29.58), so vector fluctuations of all species are expected to be negligible. On the other hand, tensor gravitational fluctuations $h_{\pm\pm}$ (gravitational waves) generated during inflation survive to low redshift, equation 29.66, and drive tensor fluctuations in collisionless relativistic species, first neutrinos, and then photons near and after recombination.

Exercise 35.3 Boltzmann code including polarization. Upgrade the code you wrote in Exercise 33.1 to implement polarization. Read the summary section 35.4 above for guidance.

35.5 Boltzmann equations for polarized photons

Whereas the unpolarized occupation number f is a scalar, the polarized occupation number f^{ab} is a tensor. The directed derivative ∂_m on the left hand side of the Boltzmann equation (33.8) should therefore be replaced by the covariant derivative $D_m f^{ab}$,

$$D_m f^{ab} = \partial_m f^{ab} + \Gamma_{cm}^a f^{cb} + \Gamma_{cm}^b f^{ac} . \quad (35.34)$$

However, the polarized (trace-free) part of f^{ab} is of linear order, and the tetrad-frame connections Γ_{abm} with a, b both spatial are all of linear order, equation (29.23) (in any gauge), so the connection terms on the right hand side of equation (35.34) are of quadratic order and can be neglected. Consequently no additional terms depending on connections arise on the left hand side of the Boltzmann equation for the polarized photon distribution.

The Boltzmann equation for the unpolarized photon distribution was given previously by equation (33.44) in conformal Newtonian gauge. The gravitational G term in this equation arises, equation (33.21), from the redshifting of photons in the unperturbed photon distribution f . Since the unperturbed photon distribution is unpolarized, the gravitational redshift terms contribute only to the unpolarized Boltzmann equation, not to the polarized Boltzmann equation. The unpolarized (spin-0) and polarized (spin-2) photon Boltzmann equations are thus

$$\boxed{\dot{\Theta} - ik\mu\Theta - G = C[\Theta]} , \quad (35.35a)$$

$$\boxed{{}_2\dot{\Theta} - ik\mu{}_2\Theta = C[{}_2\Theta]} . \quad (35.35b)$$

The collision terms $C[{}_s\Theta]$ that arise from nonrelativistic electron-photon (Thomson) scattering are calculated in §35.10. In conformal Newtonian gauge, and including not only scalar ($m = 0$) but also vector ($|m| = 1$) and tensor ($|m| = 2$) potentials from Exercise 33.3, equation (33.26), the gravitational redshift term G in the unpolarized Boltzmann equation (35.35a) is

$$G(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \dot{\Phi} + ik\mu\Psi + ik\mu\hat{\mathbf{p}} \cdot \mathbf{W} + \hat{p}^a \hat{p}^b \dot{h}_{ab} . \quad (35.36)$$

35.6 Spherical harmonics of the polarized photon distribution

The spin- s component ${}_s\Theta$ of the temperature fluctuation is naturally expanded in spin- s spherical harmonics ${}_sY_{\ell m}$, §35.12 (Seljak and Zaldarriaga, 1997; Zaldarriaga and Seljak, 1997; Hu and White, 1997). With the normalization conventional in CMB studies, the harmonic expansion of the polarized temperature fluctuation

${}_s\Theta$ is, consistent with the expansion (33.47) of the scalar ($m = 0$) fluctuation Θ , with $\hat{\mathbf{k}}$ taken along the 3-direction (z -direction),

$$\begin{aligned} {}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) &= \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} \sqrt{4\pi(2\ell+1)} {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}Y_{\ell m}^*(\hat{\mathbf{p}}, \chi) \\ &= \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} (2\ell+1) {}_s\Theta_{\ell m}(\eta, \mathbf{k}) D_{\ell m s}(\phi, \theta, \chi) , \end{aligned} \quad (35.37)$$

where ${}_sY_{\ell m}(\hat{\mathbf{p}}, \chi)$ are the spin-weighted spherical harmonics defined by equation (35.84), and $D_{\ell m s}(\phi, \theta, \chi)$ are the Wigner rotation matrices discussed in §35.12.1. The angles θ and ϕ are the polar coordinates of the photon direction $\hat{\mathbf{p}}$. Modes ${}_s\Theta_{\ell m}$ with $|m| = 0, 1, 2$ correspond respectively to scalar, vector, and tensor fluctuations. The index on the scalar ($m = 0$) fluctuation is often omitted for brevity, ${}_s\Theta_{\ell 0} = {}_s\Theta_{\ell}$. The orthogonality relation (35.110) implies that the spin harmonics ${}_s\Theta_{\ell m}$ are angular integrals of the temperature fluctuation ${}_s\Theta$ over momentum directions $\hat{\mathbf{p}}$, generalizing equation (33.48),

$${}_s\Theta_{\ell m}(\eta, \mathbf{k}) = i^{\ell+m-s} \int {}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}, 0) D_{\ell m s}^*(\phi, \theta, 0) \frac{d\mathbf{o}_{\mathbf{p}}}{4\pi} . \quad (35.38)$$

The expansion (35.37) differs from the convention of Hu and White (1997) in that (a) the expansion is with respect to ${}_{-s}Y_{\ell m}^*$ as opposed to ${}_sY_{\ell m}$, (b) there is an extra factor of $(-i)^{m-s}$, (c) the spin harmonics ${}_{-s}Y_{\ell m}^*(\hat{\mathbf{p}}, \chi)$ include a factor of $e^{-is\chi}$. The point of expanding with respect to ${}_{-s}Y_{\ell m}^*$ is that ${}_s\Theta_{\ell m}$ is then the coefficient of the spin-weight s and m (rather than s and $-m$) term under rotations about respectively the $\hat{\mathbf{p}}$ and $\hat{\mathbf{k}}$ directions, consistent with the convention in this book that the spin-weight of an object can be read off from its covariant indices. The factor of $(-i)^{m-s}$ is introduced to cancel the factor of $(-)^{m-s}$ between $D_{\ell m s}$ and its complex conjugate, equation (35.95), ensuring reality conditions (35.46) on the harmonic coefficients that match those on the Newman-Penrose components of the gravitational potentials. The factor of $e^{-is\chi}$ in ${}_{-s}Y_{\ell m}^*$ or $D_{\ell m s}$ makes explicit the spin factor that Hu and White (1997) absorb into basis vectors $\gamma_a \otimes \gamma_b$ of the polarization matrix.

35.6.1 Boltzmann equations for spherical harmonics of the polarized photon distribution

The action of $\mu \equiv \cos \theta \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}$ on the spin harmonics follows from the recursion formula (35.112) for the rotation matrices $D_{\ell m n}$. The resulting expression for the terms ${}_s\dot{\Theta} - ik\mu {}_s\Theta$ of the Boltzmann equation (35.35), common to all spins s and all harmonics ℓm , is

$$({}_s\dot{\Theta} - ik\mu {}_s\Theta)_{\ell m} = {}_s\dot{\Theta}_{\ell m} + k \left[\frac{\kappa_{\ell m s}}{2\ell+1} {}_s\Theta_{\ell-1, m} + \frac{ism}{\ell(\ell+1)} {}_s\Theta_{\ell m} - \frac{\kappa_{\ell+1, m s}}{2\ell+1} {}_s\Theta_{\ell+1, m} \right] , \quad (35.39)$$

where the coefficients $\kappa_{\ell mn}$ are given by equation (35.113). The harmonic expansion of the gravitational term G , equation (35.36), in the unpolarized Boltzmann equation is

$$G(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \sum_{\ell=0}^2 \sum_{m=-\ell}^{\ell} (-i)^{\ell+m} (2\ell+1) G_{\ell m}(\eta, \mathbf{k}) D_{\ell m 0}(\phi, \theta) , \quad (35.40)$$

with non-vanishing harmonics (do not confuse $G_{\ell m}$ here with the Einstein tensor)

$$G_{00} = \dot{\Phi} , \quad (35.41a)$$

$$G_{10} = -\frac{k}{3} \Psi , \quad (35.41b)$$

$$G_{2,\pm 1} = -\frac{k}{5\sqrt{3}} W_{\pm} , \quad (35.41c)$$

$$G_{2,\pm 2} = \frac{\sqrt{2}}{5\sqrt{3}} \dot{h}_{\pm\pm} , \quad (35.41d)$$

where $W_{\pm} \equiv \frac{1}{\sqrt{2}}(W_x \pm iW_y)$ are the spin-weight ± 1 components of the vector perturbation W_a , equation (27.22), and $h_{\pm\pm} \equiv h_{xx} \pm i h_{xy}$ are the spin-weight ± 2 components of the tensor perturbation h_{ab} , equation (27.23).

35.6.2 Electric and magnetic parts of the polarized photon distribution

The Wigner rotation matrices $D_{\ell ms}$ transform under a variety of discrete transformations. Of particular relevance here is one that flips the spin index s , which is accomplished by a parity transformation (35.96). Parity eigenstates of the rotation matrices are

$$\begin{aligned} (1 \pm P) D_{\ell ms}(\phi, \theta, \chi) &= D_{\ell ms}(\phi, \theta, \chi) \pm D_{\ell ms}(\phi + \pi, \pi - \theta, -\chi) \\ &= D_{\ell ms}(\phi, \theta, \chi) \pm (-)^{\ell} D_{\ell m, -s}(\phi, \theta, \chi) . \end{aligned} \quad (35.42)$$

The harmonics ${}_{\pm s}\Theta_{\ell m}$ of the spin $\pm s$ fluctuation thus split into an “electric” part ${}_sE_{\ell m}$ of parity $(-)^{\ell}$ and a “magnetic” part ${}_sB_{\ell m}$ of opposite parity $(-)^{\ell+1}$,

$${}_{\pm s}\Theta_{\ell m} = {}_sE_{\ell m} \pm i {}_sB_{\ell m} . \quad (35.43)$$

The names electric and magnetic come from the fact that the parity is the same as that of electric and magnetic multipole radiation; E and B here are unrelated to the electric and magnetic fields of the underlying electromagnetic radiation. There being no ambiguity, the spin index s is dropped on ${}_sE$ and ${}_sB$ for the spin ± 2 fluctuation,

$${}_{\pm 2}\Theta_{\ell m} = E_{\ell m} \pm i B_{\ell m} , \quad (35.44)$$

that is, $E_{\ell m} \equiv {}_2E_{\ell m}$ and $B_{\ell m} \equiv {}_2B_{\ell m}$. The resolution of the polarized fluctuation into parity eigenstates is motivated by the fact that the gravitational redshift term G and Thomson scattering collision terms $C[{}_s\Theta]$ are invariant under a parity transformation, so parity is an eigenstate of evolution of the polarized photon

distribution. As a consequence, the parity components of the temperature fluctuation satisfy the reality conditions (35.46).

Resolved into parity eigenstates, the Boltzmann equations (35.35) for the unpolarized and polarized temperature fluctuations are

$$(\dot{\Theta} - ik\mu\Theta)_{\ell m} = \dot{\Theta}_{\ell m} + k \left(\frac{\kappa_{\ell m 0}}{2\ell + 1} \Theta_{\ell-1, m} - \frac{\kappa_{\ell+1, m 0}}{2\ell + 1} \Theta_{\ell+1, m} \right) = G_{\ell m} + C[\Theta_{\ell m}] , \quad (35.45a)$$

$$(\dot{E} - ik\mu E)_{\ell m} = \dot{E}_{\ell m} + k \left(\frac{\kappa_{\ell m 2}}{2\ell + 1} E_{\ell-1, m} - \frac{2m}{\ell(\ell+1)} B_{\ell m} - \frac{\kappa_{\ell+1, m 2}}{2\ell + 1} E_{\ell+1, m} \right) = C[E_{\ell m}] , \quad (35.45b)$$

$$(\dot{B} - ik\mu B)_{\ell m} = \dot{B}_{\ell m} + k \left(\frac{\kappa_{\ell m 2}}{2\ell + 1} B_{\ell-1, m} + \frac{2m}{\ell(\ell+1)} E_{\ell m} - \frac{\kappa_{\ell+1, m 2}}{2\ell + 1} B_{\ell+1, m} \right) = C[B_{\ell m}] , \quad (35.45c)$$

with coefficients $\kappa_{\ell m n}$ given by equation (35.113). The azimuthal index m runs over scalar ($m = 0$), vector ($m = \pm 1$), and tensor ($m = \pm 2$) modes. Do not confuse the azimuthal index m with spin s : the unpolarized temperature fluctuation $\Theta_{\ell m} \equiv {}_0\Theta_{\ell m}$ is spin 0, while the polarized temperature fluctuations $E_{\ell m} \equiv {}_2E_{\ell m}$ and $B_{\ell m} \equiv {}_2B_{\ell m}$ are spin 2. The harmonic number ℓ must be greater than or equal to both m and s , so ℓ runs from $|m|$ to ∞ for $\Theta_{\ell m}$, and from $\max(|m|, 2)$ to ∞ for $E_{\ell m}$ and $B_{\ell m}$. When combined with the Einstein equations, §35.9, the Boltzmann equations (35.45) imply the reality conditions (35.46), which among other things imply that scalar B -modes vanish identically, $B_{\ell 0} = 0$.

Concept question 35.4 *E* and *B* modes versus Stokes parameters. Since ${}_2\Theta = E + iB$, aren't E and B (up to a factor) the same as the Stokes parameters Q and U in ${}_2\Theta \propto f^{++} \propto Q + iU$, equation (35.30)? **Answer.** No. In ${}_s\Theta = {}_sE + i{}_sB$ it is necessary to distinguish the two spins $s = 2$ and $s = -2$. The two sets of opposite spin $s = \pm 2$ are expansions in eigenfunctions $D_{\ell m s}$ of opposite spin s . In other words, ${}_2E$ is not the same as ${}_{-2}E$ because the eigenfunctions $D_{\ell m 2}$ and $D_{\ell m, -2}$ are not the same, even though the coefficients ${}_sE_{\ell m}$ are the same for $s = \pm 2$.

35.6.3 Reality conditions on the polarized photon distribution

The initial photon distribution well before recombination is in thermodynamic equilibrium and therefore unpolarized. The Einstein scalar (33.7), vector (35.52), and tensor (35.53) equations show that the scalar Ψ and Φ , vector $W_{\pm}$, and tensor $h_{\pm\pm}$ gravitational potentials are sourced by unpolarized ($s = 0$) temperature multipoles $\Theta_{\ell m}$ with respectively $|m| = 0, 1$, and 2 (and $|m| \leq \ell \leq 2$). The unpolarized temperature multipoles $\Theta_{\ell m}$ with $|m| = 0, 1$, and 2 are in turn sourced by gravitational redshift terms $G_{\ell m}$, equations (35.41). Modes with different m ($= -2, -1, 0, 1, 2$) are decoupled: gravitational modes of given m can generate only temperature fluctuations of the same m , and vice versa.

Thomson scattering generates spin 2 electric quadrupole polarization E_{2m} from unpolarized quadrupole multipoles Θ_{2m} , equation (35.68d). The polarized Boltzmann hierarchy (35.45) then feeds $\ell \geq 3$ electric $E_{\ell m}$ and, for $m \neq 0$, magnetic $B_{\ell m}$ multipoles with the same m .

The scalar potentials Ψ and Φ are real, while the Newman-Penrose components $W_{\pm}$ and $h_{\pm\pm}$ components

of the vector and tensor potentials are complex, satisfying $W_+^* = W_-$ and $h_{++}^* = h_{--}$. The Einstein and Boltzmann equations then imply the reality conditions

$$\Theta_{\ell m}^* = \Theta_{\ell, -m}, \quad E_{\ell m}^* = E_{\ell, -m}, \quad B_{\ell m}^* = -B_{\ell, -m}. \quad (35.46)$$

In particular, all scalar ($m = 0$) fluctuations are real. The scalar magnetic fluctuation vanishes, $B_{\ell 0} = 0$. The multipoles $\Theta_{\ell m}$, $E_{\ell m}$, and $B_{\ell m}$ are complex for $m \neq 0$.

As remarked in §35.4, without loss of generality the initial gravitational potentials $W_{\pm}(0)$ and $h_{\pm\pm}(0)$ can be taken to be real by absorbing a complex phase factor into their normalization (the phase factor for negative m is the complex conjugate of the phase factor for positive m ; and the phase factor is different for different m and/or wavevector $\mathbf{k}$). All linear fluctuations are proportional to the same phase factor. The phase factor cancels in power spectra, equation (36.25). If the initial gravitational potentials are real, then the Einstein and Boltzmann equations preserve that reality, so that all multipoles, including the gravitational potentials, the photon multipoles $\Theta_{\ell m}$, $E_{\ell m}$, and $B_{\ell m}$, and matter and neutrino multipoles, are real.

Concept question 35.5 Fluctuations with $|m| \geq 3$? Are there fluctuations with $|m| \geq 3$? **Answer.** No, because there are no gravitational potentials with $|m| \geq 3$. Well before recombination in the case of photons, or well before electron-positron annihilation in the case of neutrinos, collisions drive the distribution into thermodynamic equilibrium, characterized only by its first two moments, the monopole and dipole, or equivalently the density and bulk velocity. The monopole ($\ell = 0$) admits $m = 0$, while the dipole ($\ell = 1$) admits $m = 0$ or $m = \pm 1$. Later, free streaming allows higher multipoles ($\ell \geq 2$) to develop, but symmetry about the wavevector direction $\hat{\mathbf{k}}$ ensures that the azimuthal mode m remains unchanged. Gravity supports scalar ($m = 0$), vector ($m = \pm 1$), and tensor ($m = \pm 2$) modes, and these source photon or neutrino multipoles of the same m , equations (35.41). Thomson scattering sources polarized fluctuations, but leaves the azimuthal mode m unchanged.

35.7 Neutrino Boltzmann equations

Vector and tensor Einstein equations (35.52) and (35.53) are sourced by neutrinos as well as photons. Relativistic neutrinos satisfy a set of Boltzmann equations similar to the unpolarized photon Boltzmann equations (35.45a) but without scattering terms,

$$(\dot{\mathcal{N}} - ik\mu\Theta)_{\ell m} = \dot{\mathcal{N}}_{\ell m} + k \left(\frac{\kappa_{\ell m 0}}{2\ell + 1} \mathcal{N}_{\ell-1, m} - \frac{\kappa_{\ell+1, m 0}}{2\ell + 1} \mathcal{N}_{\ell+1, m} \right) = G_{\ell m}, \quad (35.47)$$

where $\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}$ is the cosine of the angle between the wavevector $\mathbf{k}$ and the neutrino momentum $\mathbf{p}$. Here $G_{\ell m}$ are the harmonics (35.41) of the gravitational term G in the Boltzmann equation, the same as for photons. Equations (35.47) include not only scalar ($m = 0$) but also vector ($m = \pm 1$) and tensor ($m = \pm 2$) equations. The scalar equations are the same as before, equations (33.91).

Concept question 35.6 Are neutrinos polarized? Relativistic neutrinos are purely left-handed, spin antialigned with their direction of motion. If neutrinos are pure left-polarized, should they not be treated using a polarized density matrix? **Answer.** A pure circularly polarized distribution is in a pure state, not a mixed state, and is described by the spin-weight $s = 0$ (not $s = \pm 2$) component f^{+-} of the polarization density matrix, §35.2.1. Gravity (in the present case, the gravitational redshift term G) is invariant under a parity transformation, and affects left- and right-handed spin states the same. The collisionless neutrino Boltzmann equation is a spin 0 equation.

35.8 Matter Boltzmann equations

Matter Boltzmann equations contain vector ($m = \pm 1$) as well as scalar ($m = 0$) parts. Matter sources contribute to the vector Einstein equations (35.52). The scalar equations are the same as before, equations (33.1) and (33.2). The Boltzmann equations for nonbaryonic cold dark matter including scalar and vector parts are

$$\dot{\delta}_c - k v_c = 3 \dot{\Phi} \quad (m = 0) , \quad (35.48a)$$

$$\dot{v}_{c,m} + \frac{\dot{a}}{a} v_{c,m} = 0 \quad (m = 0, \pm 1) . \quad (35.48b)$$

The Boltzmann equations for baryonic matter including scalar and vector parts are

$$\dot{\delta}_b - k v_b = 3 \dot{\Phi} \quad (m = 0) , \quad (35.49a)$$

$$\dot{v}_{b,m} + \frac{\dot{a}}{a} v_{b,m} = -\frac{|\dot{\tau}|}{R} (v_{b,m} - 3\Theta_{1m}) \quad (m = 0, \pm 1) . \quad (35.49b)$$

35.9 Vector and tensor Einstein equations

The photon and neutrino energy-momenta T^{kl} depends only on the unpolarized photon and neutrino distributions Θ and $\mathcal{N}$. The scalar components of the photon energy-momenta were given previously by equations (33.53). The vector components of the photon energy-momenta are given in terms of unpolarized multipole moments $\Theta_{\ell m}$ by, from equation (33.51) with integrals over Θ being converted to harmonics $\Theta_{\ell m}$ using equations (35.38),

$$T^{0\mp} = -T_{0\pm} = 4\bar{\rho}\Theta_{1,\pm 1} , \quad (35.50a)$$

$$T^{3\mp} = T_{3\pm} = -i\frac{4}{\sqrt{3}}\bar{\rho}\Theta_{2,\pm 1} , \quad (35.50b)$$

while the tensor components are

$$T^{\mp\mp} = T_{\pm\pm} = \frac{4\sqrt{2}}{\sqrt{3}}\bar{\rho}\Theta_{2,\pm 2} . \quad (35.51)$$

Massless neutrinos satisfy a similar set of equations.

The scalar Einstein equations were given previously by equations (33.7). The vector Einstein equations are, from equations (29.57),

$$-k^2 W_{\pm} = -16\pi G a^2 (\bar{\rho}_c v_{c,\pm} + \bar{\rho}_b v_{b,\pm} + 4\bar{\rho}_\gamma \Theta_{1,\pm 1} + 4\bar{\rho}_\nu \mathcal{N}_{1,\pm 1}) , \quad (35.52a)$$

$$k \left(\frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) W_{\pm} = \frac{64}{\sqrt{3}} \pi G a^2 (\bar{\rho}_\gamma \Theta_{2,\pm 1} + \bar{\rho}_\nu \mathcal{N}_{2,\pm 1}) , \quad (35.52b)$$

where $v_{\pm} \equiv \frac{1}{\sqrt{2}}(v_x \pm i v_y)$ are the spin-weight ± 1 components of the bulk velocity of a species. Only one of the two equations (35.52) is needed; the other is satisfied automatically as long as total energy-momentum is conserved. The tensor Einstein equations are, from equation (29.65),

$$\left(\frac{\partial^2}{\partial \eta^2} + 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} + k^2 \right) h_{\pm\pm} = -\frac{32\sqrt{2}}{\sqrt{3}} \pi G a^2 (\bar{\rho}_\gamma \Theta_{2,\pm 2} + \bar{\rho}_\nu \mathcal{N}_{2,\pm 2}) . \quad (35.53)$$

35.10 Polarized Thomson scattering

The invariant mean amplitude squared $\langle |\mathcal{M}|^2 \rangle$ for electron-photon scattering by nonrelativistic electrons with random spins, in which the initial photon polarization state is $\mathbf{a}$ (not to be confused with cosmic scale factor a) and the final polarization state $\mathbf{a}'$, is, from equation (47.45), generalizing the unpolarized expression (33.55),

$$\langle |\mathcal{M}|^2 \rangle = (8\pi\alpha)^2 |\bar{\mathbf{a}}' \cdot \mathbf{a}|^2 , \quad (35.54)$$

where $\alpha \equiv e^2/(\hbar c)$ is the fine-structure constant. The mean amplitude squared (35.54) is averaged over initial electron spins but not over initial photon spins, since here the initial photon polarization $\mathbf{a}$ is being specified. The adjective “mean” refers to the averaging over initial electron spins. The differential cross-section $d\sigma_T/d\Omega'$, equation (47.48), for polarized Thomson scattering is related to the mean amplitude squared $\langle |\mathcal{M}|^2 \rangle$ by, in units $c = \hbar = 1$,

$$\frac{d\sigma_T}{d\Omega'} = \frac{\langle |\mathcal{M}|^2 \rangle}{(8\pi m_e)^2} = \frac{\alpha^2}{m_e^2} |\bar{\mathbf{a}}' \cdot \mathbf{a}|^2 = r_e^2 |\bar{\mathbf{a}}' \cdot \mathbf{a}|^2 = \frac{3}{8\pi} \sigma_T |\bar{\mathbf{a}}' \cdot \mathbf{a}|^2 , \quad (35.55)$$

where $r_e = e^2/m_e c^2$ is the classical electron radius, and $\sigma_T = (8\pi/3)r_e^2$ is the total Thomson cross-section.

The collision integral $C[\Theta]$ for unpolarized Thomson scattering was given previously by equation (33.74). The same equation holds for polarized scattering, except that the scalar temperature fluctuation Θ , equation (33.47), is replaced by the polarized temperature fluctuations ${}_s\Theta$, equation (35.37), and the Thomson scattering matrix $\langle |\mathcal{M}|^2 \rangle$ becomes a matrix that couples the different spins s .

The polarized Thomson scattering matrix $\langle |\mathcal{M}|^2 \rangle$ is not diagonal in a spin (circularly polarized) basis $\boldsymbol{\gamma}_{\pm}$ (see equation (35.60)), but it is diagonal with respect to a linearly polarized basis $\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y$ in a frame where the momentum $\mathbf{p}$ of the incoming photon is along the z -direction, and the momentum $\mathbf{p}'$ of the scattered photon is in the xz -plane, as illustrated in Figure 35.1. In this special frame, the polarized Thomson scattering

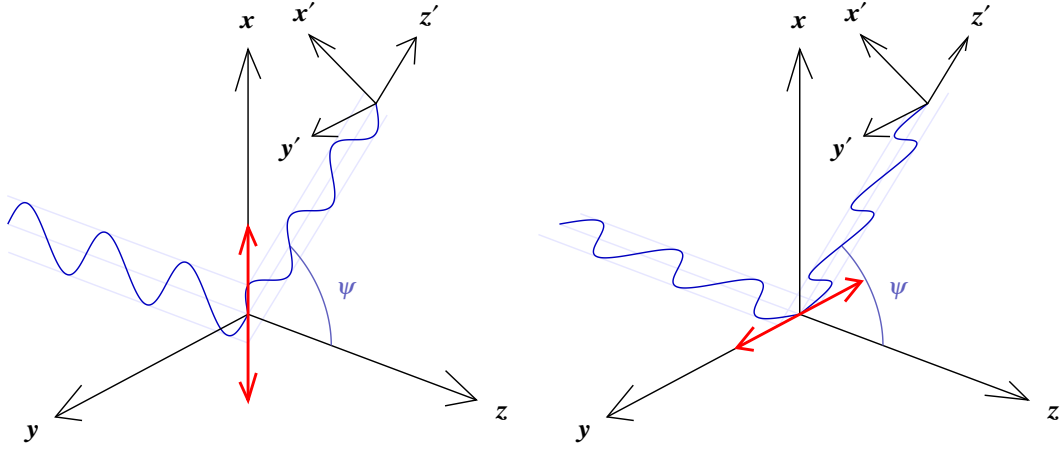


Figure 35.1 Polarized light incident in the z direction (wiggly blue line) on an electron causes the electron to oscillate in the direction $\mathbf{a}$ of polarization (red arrow). The oscillating electron emits scattered light at the same frequency (wiggly blue line). For incident light with polarization vector $\mathbf{a}_x$ in the scattering plane (left), the polarization vector $\mathbf{a}_{x'}$ of the scattered light is rotated by the scattering angle ψ , and reduced in amplitude by a factor $\cos \psi$, so that $\mathbf{a}_x \cdot \mathbf{a}_{x'} = \cos \psi$. On the other hand, for incident light with polarization vector $\mathbf{a}_y$ orthogonal to the scattering plane (right), the polarization vector $\mathbf{a}_{y'}$ of the scattered light is the same as that of the incident light, so that $\mathbf{a}_y \cdot \mathbf{a}_{y'} = 1$.

matrix $\langle |\mathcal{M}|^2 \rangle$ is, equation (47.44),

$$[\langle |\mathcal{M}|^2 \rangle]_{a'b'}^{ab} = (8\pi\alpha)^2 \begin{pmatrix} \cos^2 \psi & 0 & 0 & 0 \\ 0 & \cos \psi & 0 & 0 \\ 0 & 0 & \cos \psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (35.56)$$

where ψ is the scattering angle. The ordering of rows ab and columns $a'b'$ here is xx, xy, yx, yy . Note that up to this point the second of the two polarization indices $a\bar{b}$ has been written as a conjugate $\bar{b}$ (as a reminder that the diagonal components $+\bar{+} = + -$ and $-\bar{-} = - +$ transform as spin 0, while the off-diagonal components $+\bar{-} = ++$ and $-\bar{+} = --$ transform as spin 2), but here the conjugate symbol on the x, y indices can be dropped because the conjugates of orthonormal vectors are themselves, $\bar{\gamma}_x = \gamma_x$ and $\bar{\gamma}_y = \gamma_y$.

The collision integral (33.74) generalizes to

$$C[\Theta^{ab}(\hat{\mathbf{p}}, \chi)] = \frac{\bar{n}_e a}{16\pi m_e^2} \int [\langle |\mathcal{M}|^2 \rangle]_{a'b'}^{ab} \left[(\hat{\mathbf{p}} - \hat{\mathbf{p}}') \cdot \mathbf{v}_b - \Theta^{a'b'}(\hat{\mathbf{p}}, \chi) + \Theta^{a'b'}(\hat{\mathbf{p}}', \chi') \right] \frac{d\sigma'}{4\pi} \frac{d\chi'}{2\pi}. \quad (35.57)$$

The baryon bulk velocity $\mathbf{v}_b$ term in the integrand comes from a difference in the unperturbed, unpolarized photon distribution, the first terms on the right hand side of equation (33.65), so the integral over the baryon velocity term yields the same result as in the unpolarized case. The term $\Theta^{a'b'}(\hat{\mathbf{p}})$ is independent of $\hat{\mathbf{p}}'$, and can be taken outside the integral. The collision integral (35.57) thus reduces by the same manipulations (33.75)–

(33.78) as in the unpolarized case to

$$C[\Theta^{ab}(\hat{\mathbf{p}}, \chi)] = |\dot{\tau}| \left\{ \hat{\mathbf{p}} \cdot \mathbf{v}_b \delta^{ab,0} - \Theta^{ab}(\hat{\mathbf{p}}, \chi) + \int \frac{3}{2} [|\bar{\mathbf{a}}' \cdot \mathbf{a}|^2]_{a'b'}^{ab} \Theta^{a'b'}(\hat{\mathbf{p}}', \chi') \frac{do'}{4\pi} \frac{d\chi'}{2\pi} \right\}, \quad (35.58)$$

generalizing the unpolarized collision integral (33.78). The Kronecker delta $\delta^{ab,0}$ is to be interpreted as equal to 1 for the unpolarized collision term $C[\Theta]$, and zero otherwise. In the special frame aligned with the scattering plane, the integrand on the right hand side of the collision integral (35.58) is, from equation (35.56),

$$\frac{3}{2} [|\bar{\mathbf{a}}' \cdot \mathbf{a}|^2]_{a'b'}^{ab} \Theta^{a'b'}(\hat{\mathbf{p}}') = \frac{3}{2} \begin{pmatrix} \cos^2 \psi & 0 & 0 & 0 \\ 0 & \cos \psi & 0 & 0 \\ 0 & 0 & \cos \psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Theta^{xx} \\ \Theta^{xy} \\ \Theta^{yx} \\ \Theta^{yy} \end{pmatrix}. \quad (35.59)$$

Since Θ^{ab} is Hermitian, Θ^{xx} and Θ^{yy} are real, but Θ^{xy} may be complex, with $\Theta^{xy*} = \Theta^{yx}$. Well before recombination, frequent collisions drive the photons into thermodynamic equilibrium, so the photon distribution is initially unpolarized, with $\Theta^{xx} = \Theta^{yy}$ and $\Theta^{xy} = 0$. Equation (35.59) shows that if light incident in a given direction is initially unpolarized (Θ^{ab} isotropic, proportional to the unit matrix), then the scattered light will be polarized (Θ^{ab} anisotropic). But if Θ^{xy} is initially real, it remains real after a scattering event. Since the imaginary part of Θ^{xy} is associated with circular polarization, Thomson scattering generates linear polarization, but not circular polarization. The reality of Θ^{xy} means that $\Theta^{xy} = \Theta^{yx}$, so Θ^{yx} is redundant, and may be dropped.

In Newman-Penrose components, the absence of circularly polarized light implies that $\Theta^{+-} = \Theta^{-+} = \Theta$, where Θ is the unpolarized temperature fluctuation. In Newman-Penrose components, equation (35.59) becomes

$$\begin{aligned} \frac{3}{2} [|\mathbf{a}' \cdot \mathbf{a}|^2]_{a'b'}^{ab} \Theta^{a'b'}(\hat{\mathbf{p}}') &= \frac{3}{4} \begin{pmatrix} 1 + \cos^2 \psi & -\frac{1}{2} \sin^2 \psi & -\frac{1}{2} \sin^2 \psi \\ -\sin^2 \psi & \frac{1}{2}(1 + \cos \psi) & \frac{1}{2}(1 - \cos \psi) \\ -\sin^2 \psi & \frac{1}{2}(1 - \cos \psi) & \frac{1}{2}(1 + \cos \psi) \end{pmatrix} \begin{pmatrix} \Theta \\ \Theta^{++} \\ \Theta^{--} \end{pmatrix} \\ &= \frac{3}{2} \begin{pmatrix} \frac{2}{3} d_{000} + \frac{1}{3} d_{200} & -\sqrt{\frac{1}{6}} d_{202} & -\sqrt{\frac{1}{6}} d_{20-2} \\ -\sqrt{\frac{2}{3}} d_{220} & d_{222} & d_{22-2} \\ -\sqrt{\frac{2}{3}} d_{2-20} & d_{2-22} & d_{2-2-2} \end{pmatrix} \begin{pmatrix} \Theta \\ \Theta^{++} \\ \Theta^{--} \end{pmatrix}, \end{aligned} \quad (35.60)$$

where the functions $d_{\ell mn}(\psi)$ are the polar part of the Wigner rotation matrix, equation (35.91). The pairs ab and $a'b'$ of indices in equations (35.60) run over $+-$, $++$, and $--$. Equation (35.60) can be written

$$\frac{3}{2} [|\mathbf{a}' \cdot \mathbf{a}|^2]_s^{s'} \Theta = \Theta \delta_{s0} + \sum_{s'} (-i)^{s'-s} c_{ss'} d_{2ss'}(\psi) s' \Theta, \quad (35.61)$$

with s running over $0, 2, -2$ and s' summed over $0, 2, -2$. The first term on the right hand side of equation (35.61) is the unpolarized contribution, while the remainder is the polarized contribution. The coeffi-

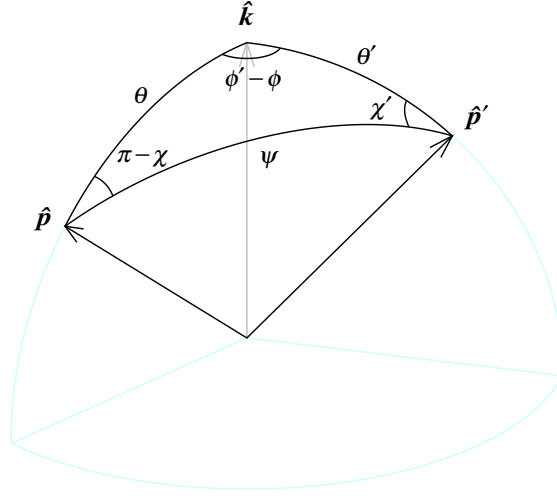


Figure 35.2 Angles between photon momentum $\hat{\mathbf{p}}$, scattered photon momentum $\hat{\mathbf{p}}'$, and wavevector $\hat{\mathbf{k}}$.

cients $c_{ss'}$ encapsulate the polarization structure of Thomson scattering,

$$c_{ss'} = \begin{pmatrix} \frac{1}{2} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} \\ \sqrt{\frac{3}{2}} & \frac{3}{2} & \frac{3}{2} \\ \sqrt{\frac{3}{2}} & \frac{3}{2} & \frac{3}{2} \end{pmatrix}. \quad (35.62)$$

The coefficients depend only on the absolute value of the spins, $c_{ss'} = c_{|s||s'|}$.

The addition theorem (35.119) allows the rotation matrix $d_{2ss'}(\psi)$ from the $\hat{\mathbf{p}}'$ frame into the $\hat{\mathbf{p}}$ frame to be written as a product of rotation matrices from the $\hat{\mathbf{p}}'$ frame into the $\hat{\mathbf{k}}$ frame into the $\hat{\mathbf{p}}$ frame,

$$\frac{3}{2} [|\mathbf{a}' \cdot \mathbf{a}|^2]_s^{s'} \Theta(\hat{\mathbf{p}}', \chi') = \Theta(\hat{\mathbf{p}}') \delta_{s0} + \sum_{s'} (-i)^{s'-s} c_{ss'} \Theta(\hat{\mathbf{p}}', \chi') \sum_{m=-2}^2 D_{2ms}(\phi, \theta, \chi) D_{2ms'}^*(\phi', \theta', \chi'). \quad (35.63)$$

Figure 35.2 illustrates the various angles involved in transforming from the scattering frame to a frame where the wavevector $\hat{\mathbf{k}}$ is along the z -axis.

When ${}_s \Theta(\hat{\mathbf{p}}', \chi')$ in equation (35.63) is expanded in rotation matrices, equation (35.37), the orthogonality of the rotation matrices, equation (35.110), makes the integration over directions $\hat{\mathbf{p}}'$ and χ' straightforward, yielding

$$\int \frac{3}{2} [|\mathbf{a}' \cdot \mathbf{a}|^2]_s^{s'} \Theta(\hat{\mathbf{p}}', \chi') \frac{d\phi'}{4\pi} \frac{d\chi'}{2\pi} = \Theta_{00} \delta_{s0} + \sum_{m=-2}^2 (-i)^{2+m-s} D_{2ms}(\phi, \theta, \chi) \sum_{s'} c_{ss'} \Theta_{2m}. \quad (35.64)$$

Equation (35.64) shows that Thomson scattering changes s (generates polarization), but preserves the scalar-

vector-tensor index m . The sum over $c_{ss'}$ in equation (35.64) is

$$\sum_{s'} c_{ss'} \Theta_{2m} = c_{s0} \Theta_{2m} + 2c_{s2} E_{2m} = c_{s0} (\Theta_{2m} + \sqrt{6} E_{2m}) . \quad (35.65)$$

The collision integral (35.58) is then

$$C[{}_s\Theta(\hat{\mathbf{p}}, \chi)] = |\dot{\tau}| \left[\hat{\mathbf{p}} \cdot \mathbf{v}_b \delta_{s0} - {}_s\Theta(\hat{\mathbf{p}}) + \Theta_{00} \delta_{s0} + c_{s0} \sum_{m=-2}^2 (-i)^{2+m-s} D_{2ms}(\phi, \theta, \chi) (\Theta_{2m} + \sqrt{6} E_{2m}) \right] . \quad (35.66)$$

Expanded in harmonics, the collision integral is

$$C[{}_s\Theta] = \sum_{\ell=|s|}^{\infty} \sum_{m=-\ell}^{\ell} (-i)^{\ell+m-s} (2\ell+1) C[{}_s\Theta_{\ell m}] D_{\ell m s}(\phi, \theta, \chi) , \quad (35.67)$$

with collision terms for the individual harmonics being

$$C[\Theta_{00}] = 0 , \quad (35.68a)$$

$$C[\Theta_{1m}] = -|\dot{\tau}| \left[\Theta_{1m} - \frac{1}{3} v_{b,m} \right] , \quad (35.68b)$$

$$C[\Theta_{2m}] = -|\dot{\tau}| \left[\Theta_{2m} - \frac{1}{10} (\Theta_{2m} + \sqrt{6} E_{2m}) \right] , \quad (35.68c)$$

$$C[E_{2m}] = -|\dot{\tau}| \left[E_{2m} - \frac{\sqrt{6}}{10} (\Theta_{2m} + \sqrt{6} E_{2m}) \right] , \quad (35.68d)$$

$$C[B_{2m}] = -|\dot{\tau}| B_{2m} , \quad (35.68e)$$

$$C[{}_s\Theta_{\ell m}] = -|\dot{\tau}| {}_s\Theta_{\ell m} \quad (\ell \geq 3) . \quad (35.68f)$$

Scalar, vector, and tensor modes correspond to those with respectively $m = 0, \pm 1$, and ± 2 .

Exercise 35.7 Photon diffusion including polarization. A diffusion approximation for the photon quadrupole fluctuation Θ_2 is obtained by neglecting time derivatives, $\dot{\Theta}_2 = 0$, and higher order multipoles, $\Theta_3 = 0$, in the Boltzmann equation for Θ_2 . Without polarization, this led to the quadrupole (32.67) in the unpolarized Boltzmann equation (33.81c). Derive the diffusion approximation for the photon quadrupole Θ_2 taking into account polarization.

Solution. With polarization, the Boltzmann equations for the quadrupole scalar ($\ell m = 20$) unpolarized and polarized fluctuations Θ_2 and E_2 are coupled to each other by Thomson-scattering collision terms, equations (35.68c) and (35.68d). The Boltzmann equations are (the $m = 0$ subscript on $\Theta_{\ell m}$ and $E_{\ell m}$ is dropped in accordance with the standard convention)

$$\dot{\Theta}_2 + \frac{k}{5} (2\Theta_1 - 3\Theta_3) = -|\dot{\tau}| \left[\Theta_2 - \frac{1}{10} (\Theta_2 + \sqrt{6} E_2) \right] , \quad (35.69a)$$

$$\dot{E}_2 - \frac{k}{\sqrt{5}} E_3 = -|\dot{\tau}| \left[E_2 - \frac{\sqrt{6}}{10} (\Theta_2 + \sqrt{6} E_2) \right] . \quad (35.69b)$$

The diffusion approximation amounts to setting time derivatives to zero, $\dot{\Theta}_2 = \dot{E}_2 = 0$, and higher order multipoles to zero, $\Theta_3 = E_3 = 0$, which reduces equations (35.69) to

$$\frac{2k}{5}\Theta_1 = -|\dot{\tau}|\left[\Theta_2 - \frac{1}{10}(\Theta_2 + \sqrt{6}E_2)\right], \quad (35.70a)$$

$$0 = -|\dot{\tau}|\left[E_2 - \frac{\sqrt{6}}{10}(\Theta_2 + \sqrt{6}E_2)\right]. \quad (35.70b)$$

The equation (35.70b) for E_2 implies that

$$E_2 = \sqrt{\frac{3}{8}}\Theta_2. \quad (35.71)$$

Inserting this into the equation (35.70a) for Θ_2 implies

$$\Theta_2 = -\frac{8k}{15|\dot{\tau}|}\Theta_1. \quad (35.72)$$

This looks like the earlier unpolarized estimate (32.67), except that the earlier factor $\frac{4}{9}$ is replaced by the factor $\frac{8}{15}$. The revised diffusion coefficient changes the factor $\frac{8}{9}$ to $\frac{16}{15}$ in the photon-baryon momentum conservation equations (32.74)–(32.76).

35.10.1 Truncating the polarized Boltzmann hierarchy

As in the unpolarized case, §33.10.1, photons are tightly coupled to baryons by scattering well before recombination, and stream freely well after recombination.

Prior to recombination, when $|\dot{\tau}|$ is large, keeping only the dominant ${}_s\Theta_{\ell m}$ term in the Boltzmann hierarchy (35.45) implies the tight-coupling approximation, generalizing the unpolarized equation (33.83),

$${}_s\Theta_{\ell m} \approx -\frac{k\kappa_{\ell ms}}{(2\ell+1)|\dot{\tau}|}{}_s\Theta_{\ell-1,m} \quad (\ell \geq 3), \quad (35.73)$$

which holds for both unpolarized ($s=0$) and polarized ($s=2$) multipoles.

Conversely, multipoles ${}_s\Theta_{\ell m}$ in the free-streaming limit are obtained, similarly to the unpolarized case §34.6.1, from solution of the polarized radiative transfer equations (36.14). The radiative transfer equations (36.14) involve unpolarized and polarized spin spherical Bessel functions $j_{\ell mm}$ and $\epsilon_{\ell 2m} + i\beta_{\ell 2m} = j_{\ell 2m2} = j_{\ell 22m}$. The recurrence (35.127) implies that the unpolarized and polarized spin spherical Bessel functions satisfy, generalizing equation (34.47),

$$\frac{\kappa_{\ell+1,m0}}{\ell+m+1}j_{\ell+1,mm} = \frac{2\ell+1}{y}j_{\ell mm} - \frac{\kappa_{\ell m0}}{\ell-m}j_{\ell-1,mm} \quad (\ell > m \geq 0), \quad (35.74a)$$

$$\frac{\kappa_{\ell+1,m2}}{\ell+3}j_{\ell+1,22m} = (2\ell+1)\left[\frac{1}{y} - \frac{im}{\ell(\ell+1)}\right]j_{\ell 22m} - \frac{\kappa_{\ell m2}}{\ell-2}j_{\ell-1,22m} \quad (\ell \geq 3). \quad (35.74b)$$

Corresponding linear combinations of multipoles in the radiative transfer equations (36.14) yield an integral similar to that on the right hand side of the neutrino equation (34.48); the integral is small in the

free-streaming limit. The result is the free-streaming approximation for unpolarized and polarized photon multipoles (note $y \rightarrow -k\eta$), generalizing equation (33.84),

$$\frac{\kappa_{\ell+1,m0}}{\ell + |m| + 1} \Theta_{\ell+1,m}(\eta, \mathbf{k}) \approx -\frac{2\ell+1}{k\eta} \Theta_{\ell m}(\eta, \mathbf{k}) - \frac{\kappa_{\ell m0}}{\ell - |m|} \Theta_{\ell-1,m}(\eta, \mathbf{k}) , \quad (35.75a)$$

$$\frac{\kappa_{\ell+1,m2}}{\ell + 3} {}_2\Theta_{\ell+1,m}(\eta, \mathbf{k}) \approx -(2\ell+1) \left[\frac{1}{k\eta} + \frac{im}{\ell(\ell+1)} \right] {}_2\Theta_{\ell m}(\eta, \mathbf{k}) - \frac{\kappa_{\ell m2}}{\ell - 2} {}_2\Theta_{\ell-1,m}(\eta, \mathbf{k}) . \quad (35.75b)$$

Normally the Boltzmann equations would be truncated at a suitably large harmonic number ℓ , but if the equations are truncated at small ℓ (for example, $\ell = 1$ for unpolarized scalar fluctuations, $m = s = 0$, yields the hydrodynamic approximation, §32.2), then unpolarized multipoles $\Theta_{\ell m}$ with $\ell = |m|$ and $m = 0$ or ± 1 in equations (35.75a) should be replaced by $\Theta_{00} \rightarrow \Theta_{00} + \Psi$ and $\Theta_{1,\pm 1} \rightarrow \Theta_{1,\pm 1} + \frac{1}{3}W_{\pm}$.

Approximations similar to the unpolarized free-streaming approximation (35.75a) hold also for neutrino multipoles $\mathcal{N}_{\ell m}$, generalizing the scalar ($m = 0$) free-streaming approximation (33.92).

35.11 Initial conditions for vector and tensor fluctuations

Collisions tend to isotropize particle distributions, leaving only the monopole moment $\ell m = 00$ finite. In the particular case of the dipole, $\ell = 1$, the Boltzmann equation (30.11b) contains a redshift term proportional to $(1 - 3w)\dot{a}/a$ that drives the velocity to decay as $v \propto a^{3w-1}$. The redshift term drives the velocity of massive species, $w = 0$, to decay as $v \propto a^{-1}$. The redshift term vanishes for relativistic species, $w = \frac{1}{3}$, but drag from collisions with massive species still causes the velocity of relativistic species to decay. Thanks to collisions, the vector and tensor fluctuations of all particle species were initially close to zero. Although neutrinos are presently collisionless, they were collisional prior to neutrino decoupling, and were isotropized at that time.

In the absence of a vector source, the vector Einstein equation (29.57a) forces the vector potential W_a to vanish,

$$W_a = 0 . \quad (35.76)$$

With no vector gravitational potential, there is no potential to drive vector multipoles of particle species away from their initial zero values. Thus all vector components of all species should remain essentially zero. This conclusion applies only to scales where fluctuations are linear: at nonlinear scales, stream-crossing and collapse generate non-zero vector components (rotations) (Hahn, Angulo, and Abel, 2015). See Exercise 35.9 for more.

In contrast to the vector potential, the tensor gravitational potential $h_{\pm\pm}$ in the absence of sources has a mode that remains constant outside the horizon, equation (29.66). This tensor gravitational potential drives tensor multipoles of collisionless species such as neutrinos, and also photons after recombination. Exercise 35.10 explores the initial evolution of tensor multipoles of neutrinos.

Exercise 35.8 Generic behaviour of scalar, vector, and tensor fluctuations of neutrinos. This exercise generalizes Exercise 32.7 to the case of vector ($m = \pm 1$) and tensor ($m = \pm 2$) fluctuations of massless neutrinos. Start with the two lowest non-vanishing Boltzmann equations (35.47), those for $\mathcal{N}_{\ell, \pm m}$ with $\ell = |m|$ and $|m|+1$, and eliminate the multipole with $\ell = |m|+2$ using the free-streaming approximation (35.75a). Conclude that, generalizing equation (32.91),

$$\left(\frac{d^2}{d\eta^2} + \frac{2}{\eta} \frac{d}{d\eta} + k^2 \right) (\mathcal{N}_0 - \Phi) = -k^2(\Psi + \Phi) , \quad (35.77a)$$

$$\left(\frac{\partial^2}{\partial \eta^2} + \frac{4}{\eta} \frac{\partial}{\partial \eta} + k^2 \right) \mathcal{N}_{1, \pm 1} = -\frac{k^2}{3} W_{\pm} , \quad (35.77b)$$

$$\left(\frac{\partial^2}{\partial \eta^2} + \frac{6}{\eta} \frac{\partial}{\partial \eta} + k^2 \right) \left(\mathcal{N}_{2, \pm 2} - \frac{\sqrt{2}}{5\sqrt{3}} h_{\pm\pm} \right) = -\frac{\sqrt{2} k^2}{5\sqrt{3}} h_{\pm\pm} . \quad (35.77c)$$

Equations (35.77) are forced, damped wave equations with effective sound speed equal to the speed of light. Generically, neutrinos are decaying waves in which:

1. Scalar: $\mathcal{N}_0 - \Phi$ oscillates about $-(\Psi + \Phi)$;
2. Vector: $\mathcal{N}_{1, \pm 1}$ oscillates about $-\frac{1}{3} W_{\pm}$;
3. Tensor: $\mathcal{N}_{2, \pm 2} - \frac{\sqrt{2}}{5\sqrt{3}} h_{\pm\pm}$ oscillates about $-\frac{\sqrt{2}}{5\sqrt{3}} h_{\pm\pm}$.

These conclusions hold for any relativistic, freely streaming particles, so apply also to photons after recombination.

Exercise 35.9 Initial evolution of vector fluctuations of neutrinos. Show that neutrinos do not naturally develop vector fluctuations.

Solution. Vector potentials $W_{\pm}$ are different from scalar or tensor potentials. Scalar and tensor potentials $\Psi + \Phi$ and $h_{\pm\pm}$ can and generically do have non-zero constant initial values well outside the horizon, $k\eta \ll 1$. Scalar potentials can have non-zero initial values because they are sourced by non-zero initial scalar overdensities Θ_0 and $\mathcal{N}_0$, equations (33.98). Tensor potentials can have non-zero initial values even if there are zero initial tensor sources $\Theta_{2, \pm 2}$ and $\mathcal{N}_{2, \pm 2}$, Exercise 35.10. But vector potentials $W_{\pm}$ are constrained by the Einstein equation (35.52a), which in standard cosmology precludes the development of a non-zero vector potential from an initially vanishing vector source. In the radiation-dominated regime following neutrino decoupling, Thomson scattering tends to isotropize radiation, so neutrinos are expected to be the dominant vector source on the right hand side of the Einstein equation (35.52a). With only neutrinos sourcing $W_{\pm}$ in the Einstein equation (35.52a), the approximate neutrino Boltzmann equation (35.77b) becomes

$$\left(\frac{\partial^2}{\partial \eta^2} + \frac{4}{\eta} \frac{\partial}{\partial \eta} + k^2 \right) \mathcal{N}_{1, \pm 1} = -\frac{64\pi G a^2 \bar{\rho}_{\nu}}{3} \mathcal{N}_{1, \pm 1} = -\frac{8f_{\nu}}{\eta^2} \mathcal{N}_{1, \pm 1} , \quad (35.78)$$

in which the final expression holds in the radiation-dominated regime, where $a \propto \eta$. The k^2 term is negligible well outside the horizon, $k\eta \ll 1$. Equation (35.78) then has solutions that are power laws $\mathcal{N}_{1, \pm 1} \propto \eta^q$, but for positive neutrino fraction, $f_{\nu} > 0$, there are no solutions for which the index q has a non-negative real part. So there are no solutions in which $\mathcal{N}_{1, \pm 1}$ is initially zero or finite (as opposed to divergent).

Exercise 35.10 Initial evolution of tensor fluctuations of neutrinos. Derive how tensor ($m = \pm 2$) neutrino multipoles evolve initially in response to gravitational waves from the early Universe, that is, to a tensor gravitational potential $h_{\pm\pm}$. This is a generalization of Exercise 33.5, which addressed the initial evolution of scalar fluctuations of neutrinos.

Solution. The Boltzmann equations for neutrinos are equations (35.47). Prior to neutrino decoupling, collisions drive all tensor multipoles to zero, $\mathcal{N}_{\ell,\pm 2} = 0$. After decoupling, neutrinos stream freely, and the gravitational tensor potential $h_{\pm\pm}$ drives the lowest order tensor multipole, $\ell = 2$, away from zero. Lower order multipoles then drive the higher multipoles, so that the equations reduce to the form $\dot{\mathcal{N}}_{\ell,\pm 2} \propto \mathcal{N}_{\ell-1,\pm 2}$. The Boltzmann hierarchy (35.47) reduces to, with $y \equiv k\eta$,

$$\frac{d\mathcal{N}_{2,\pm 2}}{dy} = \frac{\sqrt{2}}{5\sqrt{3}} \frac{dh_{\pm\pm}}{dy} , \quad (35.79a)$$

$$\frac{d\mathcal{N}_{\ell,\pm 2}}{dy} = -\frac{\kappa_{\ell,\pm 2,0}}{2\ell+1} \mathcal{N}_{\ell-1,\pm 2} \quad (\ell \geq 3) . \quad (35.79b)$$

Well outside the horizon, $y \ll 1$, the gravitational potential $h_{\pm\pm}$ is constant, equation (29.66), while all neutrino multipoles, including the lowest multipole $\mathcal{N}_{2,\pm 2}$, are zero. With the initial condition $\mathcal{N}_{2,\pm 2}(0) = 0$, equation (35.79a) solves to

$$\mathcal{N}_{2,\pm 2} = \frac{\sqrt{2}}{5\sqrt{3}} [h_{\pm\pm}(y) - h_{\pm\pm}(0)] . \quad (35.80)$$

The initial ($y \ll 1$) evolution of the gravitational tensor potential depends on the equation of state w of the background energy-momentum, equation (29.77), and is

$$h_{\pm\pm} \propto y^n J_n(y) \propto 1 - \frac{y^2}{4(1+n)} , \quad (35.81)$$

with n given in terms of w by equation (29.74). Therefore the $\ell = 2$ neutrino moment evolves initially as $\mathcal{N}_{2,\pm 2} \propto y^2$, from equation (35.80),

$$\mathcal{N}_{2,\pm 2} \approx -\mathcal{N}^{(0)} y^2 , \quad \mathcal{N}^{(0)} \equiv \frac{h_{\pm\pm}(0)}{10\sqrt{6}(1+n)} . \quad (35.82)$$

The Boltzmann equations (35.79b) then imply that the initial ($y \ll 1$) behaviour of the neutrino tensor multipoles in general is

$$\mathcal{N}_{\ell,\pm 2} = -\sqrt{\frac{(\ell-2)!(\ell+2)!}{4!}} \frac{2!5!!}{\ell!(2\ell+1)!!} (-y)^\ell \mathcal{N}^{(0)} \quad (\ell \geq 2) . \quad (35.83)$$

35.12 Appendix: Spin-weighted spherical harmonics

Spherical harmonics $Y_{\ell m}(\theta, \phi)$ are simultaneous eigenfunctions of the squared total angular momentum operator L^2 and its component L_z along some direction $\hat{\mathbf{z}}$. They arise as eigenfunctions of the wave operator when separated in spherical coordinates.

Spin contributes to angular momentum. When wave equations for fields of non-zero spin are separated in a spherically symmetric space, the resulting angular eigenfunctions are the **spin-weighted spherical harmonics**, denoted ${}_sY_{\ell m}(\theta, \phi, \chi)$. The spin-weighted spherical harmonics ${}_sY_{\ell m}(\theta, \phi, \chi)$ are defined in terms of the Wigner rotation matrix $D_{\ell mn}(\phi, \theta, \chi)$ discussed in §35.12.1 by

$${}_sY_{\ell m}(\theta, \phi, \chi) \equiv \sqrt{\frac{2\ell+1}{4\pi}} D_{\ell, m, -s}^*(\phi, \theta, \chi) . \quad (35.84)$$

The usual spherical harmonics equal the spin-weighted harmonics with zero spin, $Y_{\ell m} = {}_0Y_{\ell m}$. The reason for complex conjugation and the sign flip of the spin index s on the right hand side of equation (35.84) is that conventionally ${}_sY_{\ell m} \propto e^{im\phi - is\chi}$ whereas $D_{\ell ms} \propto e^{-im\phi - is\chi}$. The convention for the Wigner matrix, which treats the angles ϕ and χ symmetrically, is more natural than the convention for the spin-weighted spherical harmonics. In this book the temperature fluctuations ${}_s\Theta_{\ell m}$ are coefficients of an expansion in Wigner functions $D_{\ell ms}$, equation (35.37), rather than in spin-weighted spherical harmonics.

In the cosmological literature, the spin factor $e^{-is\chi}$ in the spin harmonics is often omitted, being absorbed in the case of photons into the behaviour of the polarization density matrix. The spin harmonics with spin factor suppressed are abbreviated

$${}_sY_{\ell m}(\theta, \phi) \equiv {}_sY_{\ell m}(\theta, \phi, 0) . \quad (35.85)$$

35.12.1 Wigner rotation matrix

The full 3-dimensional rotation group is the special orthogonal group $\text{SO}(3)$, or, when extended to objects of half-integral spin, its covering group $\text{Spin}(3)$, which is isomorphic to $\text{SU}(2)$. The eigenfunctions of $\text{SO}(3)$ or $\text{Spin}(3)$ (or $\text{SU}(2)$) are the elements $D_{\ell m' m}$ of the Wigner rotation matrix.

The **Wigner rotation matrix** $D_{\ell m' m}(\chi', \psi, \chi)$ is defined to be the matrix element between harmonics $Y_{\ell m}(\mathbf{n})$ in one frame and harmonics $Y_{\ell m'}^*(\mathbf{n}')$ in a frame rotated by Euler angles χ', ψ, χ ,

$$\int Y_{\ell m'}^*(\mathbf{n}') Y_{\ell m}(\mathbf{n}) d\mathbf{o} = \int \sum_{m''} (D_{\ell m' m''}(\chi', \psi, \chi) Y_{\ell m''}(\mathbf{n}))^* Y_{\ell m}(\mathbf{n}) d\mathbf{o} = \delta_{\ell' \ell} D_{\ell m' m}^*(\chi', \psi, \chi) . \quad (35.86)$$

The quantum numbers ℓ , m' , and m must be either all integral or all half-integral, and ℓ must exceed both $|m'|$ and $|m|$,

$$\ell \geq |m'|, |m| . \quad (35.87)$$

Equivalently, the spherical harmonics in the unrotated and rotated frames are related by

$$Y_{\ell m'}(\mathbf{n}') \equiv D(\chi', \psi, \chi) Y_{\ell m}(\mathbf{n}) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\chi', \psi, \chi) Y_{\ell m}(\mathbf{n}) , \quad (35.88a)$$

$$Y_{\ell m}(\mathbf{n}) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}^{\dagger}(\chi', \psi, \chi) Y_{\ell m'}(\mathbf{n}') . \quad (35.88b)$$

Notice that spherical harmonics rotate into linear combinations of harmonics of the same harmonic number ℓ , which is true because rotation leaves the total angular momentum L^2 unchanged. The Euler angles χ' , ψ , χ in equation (35.88a) correspond to a right-handed rotation of the unit vector $\mathbf{n}'$ by angle χ' about the z -axis, followed by a right-handed rotation by angle ψ about the y -axis, followed by a right-handed rotation by angle χ about the z -axis,

$$\begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \begin{pmatrix} \cos \chi & \sin \chi & 0 \\ -\sin \chi & \cos \chi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \psi & 0 & -\sin \psi \\ 0 & 1 & 0 \\ \sin \psi & 0 & \cos \psi \end{pmatrix} \begin{pmatrix} \cos \chi' & \sin \chi' & 0 \\ -\sin \chi' & \cos \chi' & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} n'_x \\ n'_y \\ n'_z \end{pmatrix} . \quad (35.89)$$

The generator of an infinitesimal rotation about an axis $\mathbf{n}$ is $-i\mathbf{n} \cdot \mathbf{L}$, and the operator corresponding to a finite rotation by angle χ about direction $\mathbf{n}$ is $\exp(-i\chi \mathbf{n} \cdot \mathbf{L})$. Thus the operator $D(\chi', \psi, \chi)$ that generates a rotation by the 3 Euler angles is

$$D(\chi', \psi, \chi) = e^{-i\chi L_z} e^{-i\psi L_y} e^{-i\chi' L_z} . \quad (35.90)$$

The spherical harmonic components of the rotation operator are correspondingly (no sum over m' , m)

$$D_{\ell m' m}(\chi', \psi, \chi) = e^{-im\chi} d_{\ell m' m}(\psi) e^{-im'\chi'} . \quad (35.91)$$

The matrix $d_{\ell m' m}(\psi)$ is the polar part of the full rotation matrix $D_{\ell m' m}(\chi', \psi, \chi)$. The polar rotation matrix $d_{\ell m' m}(\psi)$ is a real matrix, orthogonal with respect to $m'm$, with matrix inverse

$$d_{\ell m' m}(\psi)^{-1} = d_{\ell m' m}(-\psi) = d_{\ell m m'}(\psi) = d_{\ell, -m', -m}(\psi) = (-)^{m'-m} d_{\ell m' m}(\psi) . \quad (35.92)$$

A parity transformation $\psi \rightarrow \pi - \psi$ flips the sign of one of the indices m' or m and multiplies by $(-)^{\ell-m}$ or $(-)^{\ell+m'}$,

$$d_{\ell m' m}(\pi - \psi) = (-)^{\ell-m} d_{\ell, -m', m}(\psi) = (-)^{\ell+m'} d_{\ell, m', -m}(\psi) . \quad (35.93)$$

The matrix inverse of the Wigner rotation matrix is its Hermitian conjugate, its complex conjugate transpose,

$$D_{\ell m' m}(\chi', \psi, \chi)^{-1} = D_{\ell m' m}(-\chi, -\psi, -\chi') = D_{\ell m m'}^*(\chi', \psi, \chi) . \quad (35.94)$$

Complex conjugation flips the signs of m' and m , and multiplies by $(-)^{m'-m}$,

$$D_{\ell m' m}^*(\chi', \psi, \chi) = (-)^{m'-m} D_{\ell, -m', -m}(\chi', \psi, \chi) . \quad (35.95)$$

A parity transformation $\psi \rightarrow \pi - \psi$, $\chi' \rightarrow \chi' + \pi$, $\chi \rightarrow -\chi$ flips the sign of m and multiplies by $(-)^{\ell}$,

$$D_{\ell m' m}(\chi' + \pi, \pi - \psi, -\chi) = (-)^{\ell} D_{\ell, m', -m}(\chi', \psi, \chi) . \quad (35.96)$$

Particular examples of equation (35.88a), illustrating how the signs work out, are

$$Y_{\ell m}(\theta, \phi) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\chi', 0, \chi) Y_{\ell m'}(\theta, \phi + \chi' + \chi) , \quad (35.97a)$$

$$Y_{\ell m}(\theta, \phi) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\phi', \psi, -\phi) Y_{\ell m'}(\theta + \psi, \phi') . \quad (35.97b)$$

Since $Y_{\ell m}(0, 0) = \sqrt{(2\ell+1)/(4\pi)} \delta_{m0}$, the spherical harmonics $Y_{\ell m}(\theta, \phi)$ themselves can be expressed in terms of Wigner rotation matrices,

$$Y_{\ell m}(\theta, \phi) = \sqrt{\frac{2\ell+1}{4\pi}} D_{\ell 0 m}(0, -\theta, -\phi) = \sqrt{\frac{2\ell+1}{4\pi}} D_{\ell m 0}^*(\phi, \theta, 0) , \quad (35.98)$$

consistent with equation (35.84).

The explicit form of the Wigner rotation matrix elements $D_{\ell m' m}(\chi', \psi, \chi)$ is derived most elegantly from the Newman-Penrose components $L_z, L_{\pm}$ of the total angular momentum operator $\mathbf{L}$, which are (Newman and Penrose, 1962; Goldberg et al., 1967; Geroch, Held, and Penrose, 1973)

$$L_z \equiv -i \frac{\partial}{\partial \chi} , \quad L_{\pm} \equiv \frac{e^{\pm i \chi}}{\sqrt{2}} \left(\pm \frac{\partial}{\partial \psi} + i \frac{1}{\sin \psi} \frac{\partial}{\partial \chi'} + i \frac{\cos \psi}{\sin \psi} \frac{\partial}{\partial \chi} \right) . \quad (35.99)$$

A similar set of equations holds for the total angular momentum operator $\mathbf{L}'$ in the rotated (primed) frame, with $\chi' \leftrightarrow \chi$,

$$L'_z \equiv -i \frac{\partial}{\partial \chi'} , \quad L'_{\pm} \equiv \frac{e^{\pm i \chi'}}{\sqrt{2}} \left(\pm \frac{\partial}{\partial \psi} + i \frac{\cos \psi}{\sin \psi} \frac{\partial}{\partial \chi'} + i \frac{1}{\sin \psi} \frac{\partial}{\partial \chi} \right) . \quad (35.100)$$

The Newman-Penrose components $L_{\pm}$ are Hermitian conjugates with respect to integration over Euler angles,

$$L_{+}^{\dagger} = L_{-} , \quad (35.101)$$

meaning that for any differentiable functions $f(\chi', \psi, \chi)$ and $g(\chi', \psi, \chi)$,

$$\int_0^{2\pi} \int_0^{2\pi} \int_0^{\pi} f (L_{+} g) \sin \psi \, d\psi d\chi' d\chi = \int_0^{2\pi} \int_0^{2\pi} \int_0^{\pi} (L_{-} f) g \sin \psi \, d\psi d\chi' d\chi , \quad (35.102)$$

which follows from an integration by parts, the surface term vanishing when the integration is taken over the full ranges of the Euler angles. The Newman-Penrose components of the angular momentum operator form a Lie algebra, with commutators

$$[L_{+}, L_{-}] = L_z , \quad [L_z, L_{\pm}] = \pm L_{\pm} . \quad (35.103)$$

It follows from the commutation rules (35.103) that the angular momentum operators $L_{\pm}$ raise and lower by one unit the z -component L_z of the angular momentum, and similarly the angular momentum operators

$L'_\pm$ raise and lower by one unit the z -component L'_z of the angular momentum,

$$L_\pm D_{\ell, m', -m}(\chi', \psi, \chi) = \sqrt{\frac{(\ell \pm m)(\ell \mp m + 1)}{2}} D_{\ell, m', -(m \pm 1)}(\chi', \psi, \chi) , \quad (35.104a)$$

$$L'_\pm D_{\ell, m', -m}(\chi', \psi, \chi) = \sqrt{\frac{(\ell \pm m')(\ell \mp m' + 1)}{2}} D_{\ell, m' \pm 1, -m}(\chi', \psi, \chi) . \quad (35.104b)$$

The squared total angular momentum operator is

$$\begin{aligned} L^2 &= L_+ L_- + L_- L_+ + L_z^2 \\ &= L'^2 = L'_+ L'_- + L'_- L'_+ + L_z'^2 . \end{aligned} \quad (35.105)$$

The explicit form of the squared total angular momentum operator is

$$L^2 = -\frac{1}{\sin \psi} \frac{\partial}{\partial \psi} \sin \psi \frac{\partial}{\partial \psi} + \frac{1}{\sin^2 \psi} (L_z'^2 - 2 \cos \psi L'_z L_z + L_z^2) . \quad (35.106)$$

The total angular momentum squared L^2 commutes with all the individual components of the angular momentum,

$$[L^2, L_\pm] = [L^2, L_z] = 0 , \quad [L^2, L'_\pm] = [L^2, L'_z] = 0 . \quad (35.107)$$

The Wigner rotation matrix elements $D_{\ell m' m}(\chi', \psi, \chi)$ are simultaneous eigenfunctions of the total squared angular momentum operator L^2 and of the operators $L'_z \equiv -i \partial / \partial \chi'$, and $L_z \equiv -i \partial / \partial \chi$ with eigenvalues respectively $\ell(\ell + 1)$, $-m'$, and $-m$,

$$L^2 D_{\ell m' m}(\chi', \psi, \chi) = \ell(\ell + 1) D_{\ell m' m}(\chi', \psi, \chi) , \quad (35.108a)$$

$$L'_z D_{\ell m' m}(\chi', \psi, \chi) = -m' D_{\ell m' m}(\chi', \psi, \chi) , \quad (35.108b)$$

$$L_z D_{\ell m' m}(\chi', \psi, \chi) = -m D_{\ell m' m}(\chi', \psi, \chi) . \quad (35.108c)$$

The polar part $d_{\ell m' m}(\psi)$ satisfies

$$L^2 d_{\ell m' m}(\psi) = -\frac{1}{\sin \psi} \frac{\partial}{\partial \psi} \sin \psi \frac{\partial}{\partial \psi} + \frac{1}{\sin^2 \psi} (m'^2 - 2m'm \cos \psi + m^2) d_{\ell m' m}(\psi) = \ell(\ell + 1) d_{\ell m' m}(\psi) . \quad (35.109)$$

The Wigner rotation matrices are orthogonal with respect to integration over Euler angles,

$$\int_0^{2\pi} \int_0^{2\pi} \int_0^\pi D_{\ell' m' n'}^*(\chi', \psi, \chi) D_{\ell m n}(\chi', \psi, \chi) \sin \psi d\psi d\chi' d\chi = \frac{8\pi^2}{2\ell + 1} \delta_{\ell' \ell} \delta_{m' m} \delta_{n' n} . \quad (35.110)$$

The functions $D_{\ell m n}(\chi', \psi, \chi)$ and $d_{\ell m n}(\psi)$ satisfy many recurrence relations. A set of 4 building-block recurrences connecting $D_{\ell m n}$ to $D_{\ell \pm \frac{1}{2}, m \pm \frac{1}{2}, n \pm \frac{1}{2}}$ is

$$\begin{aligned} D_{\frac{1}{2}, \frac{p}{2}, \frac{q}{2}} D_{\ell m n} &= \\ \frac{1}{2\ell + 1} \left(pq \sqrt{(\ell - pm)(\ell - qn)} D_{\ell - \frac{1}{2}, m + \frac{p}{2}, n + \frac{q}{2}} + \sqrt{(\ell + 1 + pm)(\ell + 1 + qn)} D_{\ell + \frac{1}{2}, m + \frac{p}{2}, n + \frac{q}{2}} \right) , \end{aligned} \quad (35.111)$$

with $p = \pm 1$ and $q = \pm 1$. Equation (35.111) remains true with D replaced by d everywhere. Numerically the most useful recurrence relation, stable for increasing ℓ , is, a consequence of equation (35.111),

$$\kappa_{\ell+1,mn} D_{\ell+1,mn} = (2\ell + 1) \left[\cos \psi - \frac{mn}{\ell(\ell+1)} \right] D_{\ell mn} - \kappa_{\ell mn} D_{\ell-1,mn} , \quad (35.112)$$

with

$$\kappa_{\ell mn} \equiv \sqrt{\frac{(\ell^2 - m^2)(\ell^2 - n^2)}{\ell^2}} , \quad (35.113)$$

starting from $D_{\ell mn}(\chi', \psi, \chi) \equiv e^{-in\chi} d_{\ell mn}(\psi) e^{-im\chi'}$ with m or n equal to ℓ , and

$$(-)^{\ell-m} d_{\ell\ell m} = d_{\ell m \ell} = \sqrt{\frac{(2\ell)!}{(\ell+m)!(\ell-m)!}} \cos^{\ell+m} \left(\frac{\psi}{2} \right) \sin^{\ell-m} \left(\frac{\psi}{2} \right) . \quad (35.114)$$

Another useful recurrence is

$$\ell \kappa_{\ell+1,mn} D_{\ell+1,mn} = (2\ell + 1) \left[\sin \psi \frac{\partial}{\partial \psi} + \frac{mn}{\ell(\ell+1)} \right] D_{\ell mn} + (\ell + 1) \kappa_{\ell mn} D_{\ell-1,mn} . \quad (35.115)$$

Again, equations (35.112) and (35.115) remain true with D replaced by d everywhere. The rotation matrices $D_{\ell mn}$ for $m = n = 0$ reduce to Legendre polynomials,

$$D_{\ell 00}(\chi', \psi, \chi) = d_{\ell 00}(\psi) = P_{\ell}(\cos \psi) , \quad (35.116)$$

and those for $n = 0$ are proportional to associated Legendre polynomials,

$$D_{\ell m 0}(\chi', \psi, \chi) = d_{\ell m 0}(\psi) e^{-im\chi'} = \sqrt{\frac{(\ell-m)!}{(\ell+m)!}} P_{\ell}^m(\cos \psi) e^{-im\chi'} . \quad (35.117)$$

For general $\ell m' m$, the rotation matrices $D_{\ell m' m}$ are proportional to Jacobi polynomials,

$$\begin{aligned} D_{\ell m' m}(\chi', \psi, \chi) &= d_{\ell m' m}(\psi) e^{-i(m\chi + m'\chi')} \\ &= \sqrt{\frac{(\ell-m)!(\ell+m)!}{(\ell-m')!(\ell+m')!}} P_{\ell-m}^{(m-m', m+m')}(\cos \psi) \cos^{m+m'} \left(\frac{\psi}{2} \right) \sin^{m-m'} \left(\frac{\psi}{2} \right) e^{-i(m\chi + m'\chi')} . \end{aligned} \quad (35.118)$$

The analysis of polarization in §35.10 involves resolving a rotation from $\hat{\mathbf{p}}'$ to $\hat{\mathbf{p}}$ into the product of a pair of rotations with respect to a frame in which the z -axis lies along $\hat{\mathbf{k}}$. A rotation by angle ψ in the $\hat{\mathbf{p}}' \hat{\mathbf{p}}$ -plane is equivalent to a rotation by Euler angles $-\chi', -\theta', -\phi'$ from the $\hat{\mathbf{p}}'$ frame into the $\hat{\mathbf{k}}$ frame, followed by a rotation by Euler angles ϕ, θ, χ from the $\hat{\mathbf{k}}$ frame into the $\hat{\mathbf{p}}$ frame. The various angles are illustrated in Figure 35.2. The equivalence implies the addition theorem

$$\begin{aligned} d_{\ell m' m}(\psi) &= \sum_{n=-\ell}^{\ell} D_{\ell nm}(\phi, \theta, \chi) D_{\ell m' n}(-\chi', -\theta', -\phi') \\ &= \sum_{n=-\ell}^{\ell} D_{\ell nm}(\phi, \theta, \chi) D_{\ell n m'}^*(\phi', \theta', \chi') . \end{aligned} \quad (35.119)$$

35.12.2 Spin-weighted spherical Bessel functions

Spin-weighted spherical Bessel functions $j_{\ell n m s}(y)$, with $\ell, n \geq \max(|m|, |s|)$, are defined by equation (36.8). The defining equation (36.8) along with the orthogonality relations of the Wigner matrices equation (35.110), imply that

$$j_{\ell n m s}(y) = i^{\ell-n} \int_0^\pi e^{-iy \cos \theta} d_{\ell m s}(\theta) d_{n m s}(\theta) \frac{\sin \theta d\theta}{2} . \quad (35.120)$$

Equation (35.120) implies that the spin spherical Bessel functions $j_{\ell n m s}$ are symmetric or antisymmetric in their first two indices ℓn as their difference $\ell - n$ is even or odd,

$$j_{\ell n m s}(y) = (-)^{\ell-n} j_{n \ell m s}(y) . \quad (35.121)$$

Equation (35.120) also implies that $j_{\ell n m s}$ are symmetric in their last two indices $m s$, because $d_{\ell m s}(\theta) = d_{\ell s m}(-\theta)$, equation (35.92),

$$j_{\ell n m s} = j_{\ell n s m} . \quad (35.122)$$

Equation (35.96) implies that a parity flip transforms $d_{\ell m s}(\pi - \theta) = (-)^{\ell} d_{\ell m, -s}(\theta)$; since a parity flip also flips the sign of $\cos \theta$, the net result is that complex conjugation of $j_{\ell n m s}$ given by equation (35.120) flips the sign of m or s ,

$$j_{\ell n m s}^* = j_{\ell n, -m, s} = j_{\ell n, m, -s} . \quad (35.123)$$

Application of the operator $(-i)^{\frac{1}{2}(1+p-q)} D_{\frac{1}{2}, \frac{p}{2}, \frac{q}{2}}$ with $p = \pm 1$ and $q = \pm 1$ to both sides of the defining equation (36.8) for $j_{\ell n m s}$ implies, from the recurrence (35.111),

$$\begin{aligned} & \frac{1}{2n+1} \left[\sqrt{(n+1+pm)(n+1+qs)} j_{\ell+\frac{1}{2}, n+\frac{1}{2}, m+\frac{p}{2}, s+\frac{q}{2}} - ipq \sqrt{(n-pm)(n-qs)} j_{\ell+\frac{1}{2}, n-\frac{1}{2}, m+\frac{p}{2}, s+\frac{q}{2}} \right] \\ &= \frac{1}{2\ell+2} \left[\sqrt{(\ell+1+pm)(\ell+1+qs)} j_{\ell n m s} - ipq \sqrt{(\ell+1-pm)(\ell+1-qs)} j_{\ell+1, n m s} \right] . \end{aligned} \quad (35.124)$$

For $m = n$ and $p = 1$, the recurrence (35.124) simplifies to

$$\begin{aligned} & \sqrt{\frac{n+1+qs}{2n+1}} j_{\ell+\frac{1}{2}, n+\frac{1}{2}, n+\frac{1}{2}, s+\frac{q}{2}} \\ &= \frac{1}{2\ell+2} \left[\sqrt{(\ell+1+n)(\ell+1+qs)} j_{\ell n n s} - iq \sqrt{(\ell+1-n)(\ell+1-qs)} j_{\ell+1, n n s} \right] . \end{aligned} \quad (35.125)$$

From the recurrence (35.125) it can be shown by induction that $j_{\ell n m, \pm s}(y)$ with $m = n$ and integral $\ell \geq n \geq s \geq 0$ is

$$j_{\ell n n, \pm s}(y) = \sqrt{\frac{(2n)!}{(n+s)!(n-s)!}} \sqrt{\frac{(\ell+n)!(\ell-s)!}{(\ell-n)!(\ell+s)!}} \frac{1}{(2y)^n} \left(\frac{\partial}{\partial y} \mp i \right)^s [y^s j_{\ell}(y)] . \quad (35.126)$$

The $j_{\ell n m s}(y)$ with $m = n$ satisfy the recurrence

$$\frac{\kappa_{\ell+1, ns}}{\ell+n+1} j_{\ell+1, n n s} = (2\ell+1) \left[\frac{1}{y} - \frac{is}{\ell(\ell+1)} \right] j_{\ell n n s} - \frac{\kappa_{\ell ns}}{\ell-n} j_{\ell-1, n n s} , \quad (35.127)$$

with $\kappa_{\ell ns}$ defined by equation (35.113). Applying $\partial/\partial y$ to either side of the defining relation (36.8), and using the recurrence relation (35.112), implies the recurrence

$$\kappa_{n+1,ms} j_{\ell,n+1,ms} = (2n+1) \left[\frac{\partial}{\partial y} + \frac{ims}{n(n+1)} \right] j_{\ell nms} + \kappa_{nms} j_{\ell,n-1,ms} , \quad (35.128)$$

which yields $j_{\ell nms}(y)$ in general. The recurrence (35.128) of $j_{\ell nms}$ with respect to n , along with the symmetry (35.121) of $j_{\ell nms}$ in ℓn , implies a similar recurrence of $j_{\ell nms}$ with respect to ℓ ,

$$\kappa_{\ell+1,ms} j_{\ell+1,nms} = -(2\ell+1) \left[\frac{\partial}{\partial y} + \frac{ims}{\ell(\ell+1)} \right] j_{\ell nms} + \kappa_{\ell ms} j_{\ell-1,nms} . \quad (35.129)$$

Polarization of the Cosmic Microwave Background

36.1 Radiative transfer of the polarized CMB

The Boltzmann, or radiative transfer, equation for unpolarized photons was given previously by equation (34.1). For the polarized photon distribution, the radiative transfer equations are

$$\left(\frac{\partial}{\partial \eta} - ik\mu - \dot{\tau} \right) (\Theta + \Psi + \hat{\mathbf{p}} \cdot \mathbf{W}) = I - \dot{\tau} S , \quad (36.1a)$$

$$\left(\frac{\partial}{\partial \eta} - ik\mu - \dot{\tau} \right) {}_2\Theta = -\dot{\tau} {}_2S . \quad (36.1b)$$

The I in the unpolarized radiative transfer equation (36.1a) is the ISW contribution, a sum of harmonics

$$I(\eta, \mathbf{k}, \hat{\mathbf{p}}) \equiv \dot{\Psi} + \dot{\Phi} + \hat{\mathbf{p}} \cdot \dot{\mathbf{W}} + \hat{p}^a \hat{p}^b \dot{h}_{ab} = \sum_{n=0}^2 \sum_{m=-n}^n (-)^{n-m} I_{nm}(\eta, \mathbf{k}) D_{nm0}(\phi, \theta) , \quad (36.2)$$

with

$$I_{00} \equiv \dot{\Psi} + \dot{\Phi} , \quad (36.3a)$$

$$I_{1,\pm 1} \equiv \dot{W}_{\pm} , \quad (36.3b)$$

$$I_{2,\pm 2} \equiv \sqrt{\frac{2}{3}} \dot{h}_{\pm\pm} . \quad (36.3c)$$

The ${}_sS$ in equations (36.1) are Thomson-scattering (35.65) source terms,

$$S(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \Psi + \hat{\mathbf{p}} \cdot \mathbf{W} + \hat{\mathbf{p}} \cdot \mathbf{v}_b + \Theta_{00} + \frac{1}{2} \sum_{m=-2}^2 (-i)^{2+m} (\Theta_{2m} + \sqrt{6} E_{2m}) D_{2m0} , \quad (36.4a)$$

$${}_2S(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \sqrt{\frac{3}{2}} \sum_{m=-2}^2 (-i)^{2+m-2} (\Theta_{2m} + \sqrt{6} E_{2m}) D_{2m2} . \quad (36.4b)$$

The harmonic components ${}_sS_{nm}$ of ${}_sS$ defined by

$${}_sS(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \sum_{n=|s|}^2 \sum_{m=-n}^n (-i)^{n+m-s} {}_sS_{nm}(\eta, \mathbf{k}) D_{nms}(\phi, \theta, \chi) \quad (36.5)$$

are, generalizing equations (34.4),

$$S_{00} \equiv \Theta_{00} + \Psi, \quad (36.6a)$$

$$S_{10} \equiv v_b, \quad (36.6b)$$

$$S_{1,\pm 1} \equiv v_{b,\pm} + W_{\pm}, \quad (36.6c)$$

$$S_{2m} \equiv \frac{1}{2}(\Theta_{2m} + \sqrt{6} E_{2m}) \quad (-2 \leq m \leq 2), \quad (36.6d)$$

$${}_2S_{2m} \equiv \sqrt{\frac{3}{2}}(\Theta_{2m} + \sqrt{6} E_{2m}) \quad (-2 \leq m \leq 2). \quad (36.6e)$$

The solution of the radiative transfer equations (36.1) is, generalizing the unpolarized solution (34.6),

$$\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{p}}) + \Psi(\eta_0, \mathbf{k}) + \hat{\mathbf{p}} \cdot \mathbf{W}(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [e^{-\tau} I(\eta, \mathbf{k}) + g(\eta) S(\eta, \mathbf{k})] e^{-ik\mu(\eta-\eta_0)} d\eta, \quad (36.7a)$$

$${}_2\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \int_0^{\eta_0} g(\eta) {}_2S(\eta, \mathbf{k}) e^{-ik\mu(\eta-\eta_0)} d\eta, \quad (36.7b)$$

where $g(\eta)$ is the visibility function, equation (34.7).

36.2 Harmonics of the polarized CMB photon distribution

The spherical harmonics of the solution (36.7) can be found, as previously, by expanding the exponential $e^{-iy\mu}$ in spherical Bessel functions, equation (34.10). Spin-weighted spherical Bessel functions $j_{\ell nms}(y)$ with $\ell, n \geq \max(|m|, |s|)$ can be defined by a generalization of equation (34.11),

$$(-i)^{n+m-s} D_{nms}(\phi, \theta, \chi) e^{-iy \cos \theta} = \sum_{\ell=\max(|m|, |s|)}^{\infty} (-i)^{\ell+m-s} (2\ell+1) D_{\ell ms}(\phi, \theta, \chi) j_{\ell nms}(y). \quad (36.8)$$

The spin index is dropped for brevity from the spin 0 modified Bessel functions, $j_{\ell nm0}(y) = j_{\ell nm}(y)$. Properties of the spin-weighted spherical Bessel functions are addressed in Appendix 35.12.2. The spin-weighted spherical Bessel functions are symmetric or antisymmetric in their first two indices ℓn , equation (35.121), and symmetric in their last two indices ms , equation (35.122), and flipping the sign of either m or s transforms them to their complex conjugates, equation (35.123),

$$j_{\ell nms} = (-)^{\ell-n} j_{n\ell ms}, \quad j_{\ell nms} = j_{\ell nsm}, \quad j_{\ell nms}^* = j_{\ell n, -m, s} = j_{\ell n, m, -s}. \quad (36.9)$$

In particular, the spin zero functions $j_{\ell nm}$ are real, and all scalar ($m = 0$) components $j_{\ell n0s}$ are real. The real (electric) and imaginary (magnetic) parts of $j_{\ell nms}$ are conveniently denoted by the real functions ${}_s\epsilon_{\ell nm}$

and ${}_s\beta_{\ell nm}$ defined by

$$j_{\ell nm, \pm s} = {}_s\epsilon_{\ell nm} \pm i {}_s\beta_{\ell nm} . \quad (36.10)$$

The spin zero magnetic part vanishes, ${}_0\beta_{\ell nm} = 0$. The only other spin relevant is $s = \pm 2$, so the spin index is dropped for brevity on the spin two electric and magnetic components,

$$j_{\ell nm, \pm 2} = \epsilon_{\ell nm} \pm i \beta_{\ell nm} . \quad (36.11)$$

Under $m \rightarrow -m$, the electric components are unchanged, while the magnetic components change sign,

$$j_{\ell n, -m} = j_{\ell nm} , \quad \epsilon_{\ell n, -m} = \epsilon_{\ell nm} , \quad \beta_{\ell n, -m} = -\beta_{\ell nm} . \quad (36.12)$$

In all, the spin spherical Bessel functions of relevance are, from equation (35.126) for $j_{\ell nms}$ with $n = m$, and the recurrence (35.128) for $n > m$,

$$j_{\ell 00} = j_\ell , \quad j_{\ell 10} = \frac{dj_\ell}{dy} , \quad j_{\ell 20} = \frac{1}{2} \left(1 + 3 \frac{d^2}{dy^2} \right) j_\ell , \quad (36.13a)$$

$$j_{\ell 11} = \sqrt{\frac{\ell(\ell+1)}{2}} \frac{j_\ell}{y} , \quad j_{\ell 21} = \sqrt{\frac{3\ell(\ell+1)}{2}} \frac{d(j_\ell/y)}{dy} , \quad (36.13b)$$

$$j_{\ell 22} = \sqrt{\frac{3(\ell+2)!}{8(\ell-2)!}} \frac{j_\ell}{y^2} , \quad (36.13c)$$

$$\epsilon_{\ell 20} = \sqrt{\frac{3(\ell+2)!}{8(\ell-2)!}} \frac{j_\ell}{y^2} , \quad \epsilon_{\ell 21} = \frac{\sqrt{(\ell-1)(\ell+2)}}{2} \frac{1}{y^2} \frac{d(y j_\ell)}{dy} , \quad \epsilon_{\ell 22} = \frac{1}{4y^2} \left(\frac{d^2}{dy^2} - 1 \right) (y^2 j_\ell) , \quad (36.13d)$$

$$\beta_{\ell 20} = 0 , \quad \beta_{\ell 21} = -\frac{\sqrt{(\ell-1)(\ell+2)}}{2} \frac{j_\ell}{y} , \quad \beta_{\ell 22} = -\frac{1}{2y^2} \frac{d(y^2 j_\ell)}{dy} . \quad (36.13e)$$

Expanding the solution (36.7) in spherical harmonics using equation (36.8) yields the harmonics of the

CMB photon distribution today including polarization, generalizing equation (34.17),

$$\begin{aligned} \Theta_{\ell 0}(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \left[\dot{\Psi}(\eta_0, \mathbf{k}) + \dot{\Phi}(\eta_0, \mathbf{k}) \right] j_{\ell 00} [k(\eta - \eta_0)] \\ &\quad + g(\eta) \sum_{n=0}^2 S_{n0}(\eta, \mathbf{k}) j_{\ell n0} [k(\eta - \eta_0)] d\eta, \end{aligned} \quad (36.14a)$$

$$\begin{aligned} \Theta_{\ell, \pm 1}(\eta_0, \mathbf{k}) + \frac{1}{3} \delta_{\ell 1} W_{\pm}(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \dot{W}_{\pm}(\eta_0, \mathbf{k}) j_{\ell 11} [k(\eta - \eta_0)] \\ &\quad + g(\eta) \sum_{n=1}^2 S_{n, \pm 1}(\eta, \mathbf{k}) j_{\ell n1} [k(\eta - \eta_0)] d\eta, \end{aligned} \quad (36.14b)$$

$$\begin{aligned} \Theta_{\ell, \pm 2}(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \sqrt{\frac{2}{3}} \dot{h}_{\pm \pm}(\eta_0, \mathbf{k}) j_{\ell 22} [k(\eta - \eta_0)] \\ &\quad + g(\eta) S_{2, \pm 2}(\eta, \mathbf{k}) j_{\ell 22} [k(\eta - \eta_0)] d\eta, \end{aligned} \quad (36.14c)$$

$$E_{\ell m}(\eta_0, \mathbf{k}) = \int_0^{\eta_0} g(\eta) {}_2S_{2m}(\eta, \mathbf{k}) \epsilon_{\ell 2m} [k(\eta - \eta_0)] d\eta \quad (-2 \leq m \leq 2), \quad (36.14d)$$

$$B_{\ell m}(\eta_0, \mathbf{k}) = \int_0^{\eta_0} g(\eta) {}_2S_{2m}(\eta, \mathbf{k}) \beta_{\ell 2m} [k(\eta - \eta_0)] d\eta \quad (-2 \leq m \leq 2). \quad (36.14e)$$

The Thomson-scattering source terms ${}_sS_{nm}$ are given by equations (36.6), and the spin spherical Bessel functions by equations (36.13a).

Exercise 36.1 Neutrino harmonics including vectors and tensors. Equation (34.46) gave the solution to the radiative transfer equation for scalar ($m = 0$) fluctuations of (massless) neutrinos. Generalize this to include vector ($m = \pm 1$) and tensor ($m = \pm 2$) neutrino fluctuations.

Solution. The solution is similar to that (36.14a)–(36.14c) for unpolarized photons, but without the Thomson scattering terms:

$$\begin{aligned} \mathcal{N}_{\ell}(\eta, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta, \mathbf{k}) &= \int_0^{\eta} \left[\dot{\Psi}(\eta', \mathbf{k}) + \dot{\Phi}(\eta', \mathbf{k}) \right] j_{\ell} [k(\eta' - \eta)] d\eta' \\ &\quad + [\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k})] j_{\ell}(-k\eta), \end{aligned} \quad (36.15a)$$

$$\begin{aligned} \mathcal{N}_{\ell, \pm 1}(\eta, \mathbf{k}) + \frac{1}{3} \delta_{\ell 1} W_{\pm}(\eta, \mathbf{k}) &= \int_0^{\eta} \dot{W}_{\pm}(\eta', \mathbf{k}) j_{\ell 11} [k(\eta' - \eta)] d\eta' \\ &\quad + W_{\pm}(0, \mathbf{k}) j_{\ell 11}(-k\eta), \end{aligned} \quad (36.15b)$$

$$\mathcal{N}_{\ell, \pm 2}(\eta, \mathbf{k}) = \int_0^{\eta} \sqrt{\frac{2}{3}} \dot{h}_{\pm \pm}(\eta', \mathbf{k}) j_{\ell 22} [k(\eta' - \eta)] d\eta'. \quad (36.15c)$$

As in equation (34.46), the time η_ν of neutrino decoupling has been replaced by zero, and the optical depth factor omitted, since the neutrino decoupling scale is so much smaller than cosmological scales.

36.2.1 Harmonics of the polarized CMB with respect to observed photon directions

As with the unpolarized power spectrum, §34.2.1, the observed direction of $\hat{\mathbf{n}}$ of a photon from the CMB is opposite to the photon's direction of motion, $\hat{\mathbf{n}} = -\hat{\mathbf{p}}$. Moreover the right-handed direction around $\hat{\mathbf{p}}$ becomes a left-handed direction about $\hat{\mathbf{n}}$, so the spin angle χ also flips in sign. Thus harmonics with respect to the observed direction $\hat{\mathbf{n}}$ are related to those relative to the photon direction $\hat{\mathbf{p}}$ by ${}_s\Theta^{\text{obs}}(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) = {}_s\Theta(\eta, \mathbf{k}, -\hat{\mathbf{p}}, -\chi)$. The reversal of $\hat{\mathbf{p}}$ and χ is equivalent to a parity flip, which changes the spin s temperature multipoles by ${}_s\Theta_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) = (-)^{\ell+s} {}_s\Theta_{\ell m}(\eta_0, \mathbf{k})$. Equivalently, generalizing equation (34.18),

$$\Theta_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^{\ell} \Theta_{\ell m}(\eta_0, \mathbf{k}) , \quad E_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^{\ell} E_{\ell m}(\eta_0, \mathbf{k}) , \quad B_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^{\ell+1} B_{\ell m}(\eta_0, \mathbf{k}) . \quad (36.16)$$

Equivalently, as in the unpolarized equation (34.17), multipoles with respect to the observed direction $\hat{\mathbf{n}}$ to the CMB are obtained from equations (36.14) by flipping the sign of the arguments of the spin spherical Bessel functions $j_{\ell n m s}$ and simultaneously flipping the sign of source terms ${}_sS_{nm}$ with odd n , namely S_{1m} ,

$$k(\eta - \eta_0) \rightarrow k(\eta_0 - \eta) , \quad S_{1m} \rightarrow -S_{1m} . \quad (36.17)$$

The sign flips do not affect power spectra, which involve products of fluctuations with the same ℓ and parity.

36.3 Harmonics of the polarized CMB in real space

The real-space polarized temperature fluctuation ${}_s\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}}, \chi)$ at time η and comoving position $\mathbf{x}$ in observed direction $\hat{\mathbf{n}}$ on the sky is related to the Fourier-space polarized temperature fluctuation ${}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{n}}, \chi)$ by, generalizing the unpolarized expression (34.28),

$${}_s\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}}, \chi) = \int e^{-i\mathbf{k} \cdot \mathbf{x}} {}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{n}}, \chi) \frac{d^3k}{(2\pi)^3} . \quad (36.18)$$

Astronomers observe the temperature fluctuation ${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi)$ now, at time η_0 , and here, at position $\mathbf{x}_0$. Without loss of generality, our position can be taken to be at the origin, $\mathbf{x}_0 = \mathbf{0}$, in which case the phase factor is unity, $e^{-i\mathbf{k} \cdot \mathbf{x}_0} = 1$, and can be omitted,

$${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi) = \int \Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}, \chi) \frac{d^3k}{(2\pi)^3} , \quad (36.19)$$

which generalizes the unpolarized expression (34.29).

The spherical harmonic expansion of the observed real-space temperature fluctuation today is, with conventional normalization of harmonics ${}_s\Theta_{\ell m}$,

$${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi) = \sum_{\ell=|s|}^{\infty} \sum_{m=-\ell}^{\ell} {}_s\Theta_{\ell m}(\eta_0, \mathbf{x}_0) {}_{-s}Y_{\ell m}^*(\hat{\mathbf{n}}, \chi) , \quad (36.20)$$

which generalizes equation (34.30). The reason for the expansion with respect to ${}_{-s}Y_{\ell m}^*$ as opposed to ${}_sY_{\ell m}$ is that, as already remarked after equation (35.37), the coefficient ${}_s\Theta_{\ell m}$ then has spin weight s and m as opposed to s and $-m$. The spherical harmonic expansion (35.37) of the Fourier-space temperature fluctuation may be written

$${}_s\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}, \chi) = \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \sum_{n=-\ell}^{\ell} (-i)^{\ell+m-s} \sqrt{4\pi(2\ell+1)} {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) {}_{-s}Y_{\ell n}^*(\hat{\mathbf{n}}, \chi) , \quad (36.21)$$

where ${}_sD_{\ell m'm}(\hat{\mathbf{n}}', \hat{\mathbf{n}})$ is the matrix that rotates spin harmonics ${}_sY_{\ell m}$, defined by, analogously to the definition (35.86) of Wigner rotation matrices $D_{\ell m'm}(\hat{\mathbf{n}}', \hat{\mathbf{n}})$,

$$\delta_{\ell'\ell} {}_sD_{\ell m'm}(\hat{\mathbf{n}}', \hat{\mathbf{n}}) \equiv \int {}_sY_{\ell' m'}^*(\hat{\mathbf{n}}') {}_sY_{\ell m}(\hat{\mathbf{n}}) d\Omega . \quad (36.22)$$

It is not necessary to know an explicit form for the spin rotation matrices ${}_sD_{\ell m'm}$, because observable power spectra are rotation invariant, and do not depend on the form of ${}_sD_{\ell m'm}$. Whereas the original harmonics ${}_s\Theta_{\ell m}(\mathbf{k})$ are with respect to a frame in which the z -axis is along the wavevector $\mathbf{k}$, the rotated harmonics $\sum_m {}_s\Theta_{\ell m}(\mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}})$ in equation (36.21) are with respect to a frame in which the z -axis is along a direction $\hat{\mathbf{z}}$ fixed in space.

From equations (36.19)–(36.21) it follows that the real-space harmonics are

$${}_s\Theta_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} \int {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3} . \quad (36.23)$$

The factors of $\sqrt{4\pi(2\ell+1)}(-i)^{\ell+m-s}$ arise because of the different choices of normalization of the harmonics (as is the standard cosmological convention) in the harmonic expansions (35.37) and (36.20) of the temperature fluctuation in Fourier and real space.

Rotating the Fourier-space harmonics ${}_s\Theta_{\ell m}(\eta, \mathbf{k})$ from the $\hat{\mathbf{k}}$ frame into the $\hat{\mathbf{z}}$ frame leaves their parity unchanged, so the real-space harmonics inherit their parity from Fourier space. Resolved into parity eigenstates,

the real-space harmonics (36.23) are

$$\Theta_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m} \int \Theta_{\ell m}(\eta, \mathbf{k}) D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}, \quad (36.24a)$$

$$E_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-2}^2 (-i)^{\ell+m-2} \int E_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}, \quad (36.24b)$$

$$B_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-2}^2 (-i)^{\ell+m-2} \int B_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}. \quad (36.24c)$$

36.4 Polarized CMB power spectra

36.4.1 Polarized CMB power spectra in Fourier space

Power spectra $C_\ell^{X'X}(\eta, k)$ with X' and X running over any of Θ , E , and B are defined by, analogously to the power spectrum $C_\ell(\eta, k)$ of unpolarized temperature multipoles, equation (34.26),

$$\frac{\delta_{\ell'\ell}}{4\pi} (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) C_\ell^{X'X}(\eta, k) \equiv \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \langle X_{\ell'm}^*(\eta, \mathbf{k}') X_{\ell m}(\eta, \mathbf{k}) \rangle. \quad (36.25)$$

The reality conditions (35.46) imply that the power spectra are real-valued, and symmetric in $X'X$, $C_\ell^{X'X} = C_\ell^{XX'}$. Strictly, on the right hand side of equation (36.25) the unpolarized monopole Θ_{00} should be replaced by the redshifted monopole $\Theta_{00} + \Psi$, and the unpolarized dipole $\Theta_{1,\pm 1}$ should be replaced by the Doppler-shifted dipole $\Theta_{1,\pm 1} + \frac{1}{3}W_\pm$, in accordance with equations (36.14), but these refinements are omitted here to avoid cluttering the equation.

Polarized CMB transfer functions $T_{\ell m}^X(\eta, k)$ for any of $X = \Theta$, E , or B are defined by, generalizing equation (34.20) (the contributions Ψ and $\frac{1}{3}W_\pm$ to the unpolarized monopole and dipole are again omitted for brevity),

$$T_{\ell m}^X(\eta, k) \equiv \frac{X_{\ell m}(\eta, \mathbf{k})}{\zeta(\mathbf{k})}, \quad (36.26)$$

where $\zeta(\mathbf{k})$ is the primordial curvature fluctuation. In terms of the transfer functions (36.26) and the primordial curvature power spectrum P_ζ , equation (30.131), the power spectrum $C_\ell^{X'X}(\eta, k)$ is

$$C_\ell^{X'X}(\eta, k) = 4\pi \sum_{m=-2}^2 T_{\ell m}^{X'*}(\eta, k) T_{\ell m}^X(\eta, k) P_\zeta(k). \quad (36.27)$$

36.4.2 Conditions on polarized CMB power spectra from parity symmetry

The Universe at large is consistent with being statistically homogeneous and isotropic, and it is reasonable to expect that the statistical properties would similarly be parity symmetric, unchanged under spatial inversion. The prediction of parity symmetry is, like homogeneity and isotropy, testable observationally. The temperature and electric fluctuations $\Theta_{\ell m}$ and $E_{\ell m}$ have the same $(-)^{\ell}$ parity under spatial inversion, while the magnetic fluctuation $B_{\ell m}$ has the opposite $(-)^{\ell+1}$ parity. The assumption of parity symmetry then implies that cross power spectra between fluctuations of opposite parity should vanish, $C_{\ell}^{\Theta B} = C_{\ell}^{EB} = 0$, since these power spectra change sign under parity inversion. Parity symmetry predicts that the non-vanishing power spectra are

$$C_{\ell}^{\Theta\Theta}, \quad C_{\ell}^{\Theta E}, \quad C_{\ell}^{EE}, \quad C_{\ell}^{BB}. \quad (36.28)$$

36.4.3 Polarized CMB power spectra in real space

CMB power spectra $C_{\ell}^{X'X}(\eta_0)$ on the sky today with X' and X any of Θ , E , and B are defined such that, generalizing equation (34.33),

$$\delta_{\ell'\ell} \delta_{m'm} C_{\ell}^{X'X}(\eta_0) \equiv \langle X_{\ell'm'}^{*}(\eta_0, \mathbf{x}_0) X_{\ell m}(\eta_0, \mathbf{x}_0) \rangle. \quad (36.29)$$

Once again, the redshift contribution Ψ to the unpolarized monopole Θ_{00} , and the Doppler-shift contribution $\frac{1}{3}W_m$ to the dipole Θ_{1m} on the right hand side of equation (36.29) have been omitted for brevity. The monopole and dipole are indistinguishable from a rescaling of the mean temperature and from a change in the motion of the observer, so cannot be measured by an observer confined to position $\mathbf{x}_0$.

From the expressions (36.24) for the real-space harmonics in terms of Fourier-space harmonics, together with the power spectra (36.25) of the Fourier-space harmonics, it follows that the power spectra $C_{\ell}^{X'X}(\eta_0)$ of real-space harmonics of the CMB today are, generalizing equation (34.34),

$$C_{\ell}^{X'X}(\eta_0) = \int C_{\ell}^{X'X}(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3}. \quad (36.30)$$

The CMB power spectra $C_{\ell}^{X'X}(\eta_0)$ inherit from $C_{\ell}^{X'X}(\eta_0, k)$ the properties of being real-valued and symmetric in $X'X$.

In terms of the polarized CMB transfer functions $T_{\ell m}^X$ defined by equation (36.26) and the primordial curvature power spectrum P_{ζ} , the power spectra $C_{\ell}^{X'X}(\eta_0)$ are, from equation (36.27),

$$C_{\ell}^{X'X}(\eta_0) = 4\pi \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \int T_{\ell m}^{X'*}(\eta_0, k) T_{\ell m}^X(\eta_0, k) P_{\zeta}(k) \frac{4\pi k^2 dk}{(2\pi)^3}. \quad (36.31)$$

Concept question 36.2 **Scalar, vector, tensor power spectra?** Can power spectra of scalar, vector, and tensor modes be distinguished observationally? **Answer.** No, with an exception. Scalar, vector, and tensor modes are characterized by their transformation properties under rotation about the wavevector $\mathbf{k}$ of

the perturbation. An observed temperature fluctuation in real space is a superposition of fluctuations with many wavevectors $\mathbf{k}$, and thereby becomes a mixture of scalar, vector, and tensor modes. The exception is that the scalar magnetic fluctuation $B_{\ell 0}$ vanishes identically, so the magnetic power spectrum C_ℓ^{BB} measures only vector and tensor modes. Mathematically, the real-space harmonics (36.24) of the temperature fluctuation are sums over scalar, vector, and tensor modes, $|m| = 0, 1, 2$. In Fourier space, power spectra $C_{\ell m}(\eta_0, k)$ with definite m can be defined by equation (36.25) without summing over m . But in real space, the CMB power spectrum $C_\ell(\eta_0)$, equation (36.31), is a sum over the scalar, vector, and tensor Fourier-space power spectra,

$$C_\ell^{X'X}(\eta_0) = \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \int C_{\ell m}^{X'X}(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (36.32)$$

Exercise 36.3 CMB polarized power spectrum. Generalize the CMB code you wrote in Exercise 34.1 to include polarization.

Gravitational lensing of the Cosmic Microwave Background

Galaxies along the line of sight slightly perturb the trajectories of photons emitted at the surface of last scattering (Zaldarriaga and Seljak, 1998, and references therein). The qualitative effect of this gravitational lensing effect is to tend to blur CMB fluctuations at small scales. The gravitational lensing effect has been neglected in this book up to now on the grounds that its magnitude is proportional to a product

$$\frac{d\hat{\mathbf{p}}}{d\lambda} \cdot \frac{\partial f}{\partial \hat{\mathbf{p}}} \quad (37.1)$$

of terms that were both linear in the photon Boltzmann equation (33.8), and therefore of the second order of smallness. The reason the gravitational lensing effect is important despite being of second order is that it feeds B -mode polarization from E -mode polarization. At small angular scales, gravitational lensing proves to dominate the primordial B -mode signal expected from gravitational waves generated at inflation. Fortunately the lensing effect is small at large angular scales, leaving a window where a signal from primordial gravitational waves might be seen in the future. An upside of gravitational lensing is that, because it depends on the clustering of matter well after recombination, it resolves degeneracies in cosmological parameters that would be inferred from the unlensed CMB power spectrum at the surface of last scattering.

The product of terms that was neglected in the photon Boltzmann equation (33.8) is $d\hat{\mathbf{p}}/d\lambda \cdot \partial f/\partial \hat{\mathbf{p}}$, and these must now be restored.

Inflation

As expounded in §10.22, inflation has become part of the Standard Model of Cosmology, in part because it offers a solution to a number of problems, such as the horizon problem, the expansion problem, and where matter came from, and in part because it makes predictions that agree with observations, notably that primordial fluctuations should be adiabatic, Gaussian, and almost scale-free, as seen in the power spectrum of the cosmic microwave background (Planck Collaboration, 2020a; Planck Collaboration, 2020b).

For inflation to solve the horizon problem requires that inflation must have continued over at least ~ 60 e -folds of the cosmic scale factor a ,

$$N \gtrsim 61 + \ln \left(\frac{T_{\text{reheat}}}{10^{16} \text{ GeV}} \right) . \quad (38.1)$$

Figure 10.14 shows inflation extending over 66 e -folds. The leading idea is that inflation was driven by the potential energy $V(\varphi)$ of a scalar field φ , the **inflaton** field, which might perhaps have been the Higgs field associated with grand symmetry breaking at an energy scale of $\sim 10^{16}$ GeV. **Slow-roll inflation** envisages that the scalar field φ started near the top of the potential $V(\varphi)$, and slowly rolled down to the bottom of the potential over the course of many e -folds of expansion. Many e -folds requires the potential energy curve of the scalar field φ to be sufficiently flat. Inverting the power spectrum of the CMB to extract the primordial power spectrum (Planck Collaboration, 2020b) leads to the conclusion that the potential energy curve is indeed remarkably flat.

Inflation predicts that the fluctuations observed in the cosmic microwave background and in the distribution of galaxies today were generated by vacuum fluctuations in the inflaton field that drove inflation. Inflation tends to redshift any pre-existing fluctuations to zero, but the vacuum state of any field sustains intrinsic quantum fluctuations that cannot be eliminated.

Evidently inflation eventually came to an end, whereupon vacuum energy was apparently converted into the energy of matter and radiation in an event called **reheating**, whereupon the Universe transitioned from accelerating to decelerating. What happened at reheating is not well understood theoretically, because the physics at grand unification (or whatever the energy scale of inflation may have been) is not well understood. The general picture is that once the inflaton field reached the bottom of its potential well, it oscillated around that minimum, coupling to other fields, converting vacuum energy into matter and radiation.

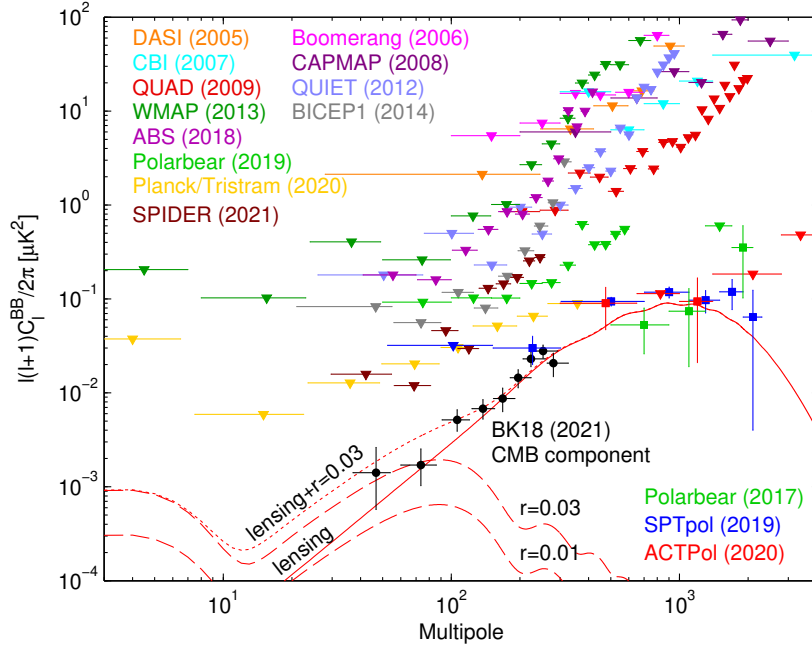


Figure 38.1 CMB B -mode polarization upper limits and detections (BICEP/Keck Collaboration, 2021, Fig. 7). Lines show theoretical predictions for the lensing B -modes (solid red) which peak at arcminute scales $\ell \sim 1000$, and for primordial gravitational wave B -modes (dashed red) for two values of the tensor-to-scalar ratio r , peaking at degree scales $\ell \sim 80$.

Fortunately the details of reheating do not matter greatly to the spectrum of perturbations that is generated. A vacuum fluctuation of the inflaton field becomes frozen as it exits the horizon of the accelerating Universe. Its amplitude is characterized by a curvature fluctuation ζ , equation (29.42), that remains constant while outside the horizon. When the curvature fluctuation comes back inside the horizon, that curvature fluctuation ζ is the seed from which grows the structure seen today in the cosmic microwave background and in the clustering of galaxies.

Inflation predicts that gravitational waves should leave an imprint in the B -mode polarization of fluctuations in the cosmic microwave background. A combination of satellite and ground-based observations have measured the B -mode power spectrum, Figure 38.1, taken from the BICEP/Keck Collaboration (2021, Fig. 7). At smaller angular scales the signal is dominated by foreground lensing, not by the predicted spectrum of primordial waves. At larger angular scales there are at present only upper limits. The Planck Collaboration (2020, eq. (32)) infer an upper limit to the energy scale of inflation, the $\frac{1}{4}$ th power of the potential energy V at wavelengths that exited the horizon during (early) inflation, Concept question 38.2,

$$V^{1/4} \lesssim 1.6 \times 10^{16} \text{ GeV} . \quad (38.2)$$

King et al. (2021) infer a model-dependent lower limit to the energy scale μ_{GUT} of grand unification from the Super-Kamiokande lower limit of $\gtrsim 1.6 \times 10^{34}$ yr on the proton lifetime (Takenaka et. al., 2020),

$$\mu_{\text{GUT}} \gtrsim 4 \times 10^{15} \text{ GeV} . \quad (38.3)$$

The upper (38.2) and lower (38.3) limits are intriguingly close together, encouraging optimism that forthcoming experiments may detect a primordial signal in the B -mode power spectrum of the CMB. If so, it would be a triumphant view of the Universe at its earliest moments, and of physics at GUT-scale energies inaccessible to earthly laboratories.

This chapter starts, §38.1, with a brief review of how quantum field theory predicts that the vacuum state of any field in Minkowski (flat) spacetime contains fluctuations with a nonvanishing Gaussian power spectrum. An FLRW spacetime is not Minkowski, but fluctuations well inside the horizon during inflation may be treated as approximately Minkowski. Inflation predicts that these vacuum fluctuations seeded fluctuations in the cosmic microwave background and in the large scale clustering of galaxies.

The chapter goes on, §38.2, to calculate the power spectrum of gravitational waves (tensor perturbations) generated by inflation. The generation of gravitational waves proves to be more straightforward to calculate than the generation of scalar fluctuations by a scalar field, because the tensor perturbation h_{ab} is gauge invariant, whereas the scalar field perturbation $\delta\varphi$ is not.

The chapter proceeds to slow-roll inflation driven by a scalar inflaton field, §38.3, and the power spectrum of scalar waves generated, §38.4.

Concept question 38.1 Largest and smallest scales of inflation. Vacuum fluctuations of the inflaton field seeded fluctuations seen in the Universe today at any scale that lay inside the horizon at the onset of inflation and exited the horizon during inflation. What are the largest and smallest scales over which inflation generated fluctuations? **Answer.** Solution of the horizon problem requires that the largest observable scale in the Universe, the horizon scale today, about 14.5 Gpc, equation (10.101), lay within the horizon when inflation began. That largest observable scale is just coming inside the horizon today, Figures 10.12 or 10.13. If there was a beginning to inflation, then scales comparable to the horizon size at that beginning were the first to exit the horizon, and would correspond to the largest scale over which the early Universe was in causal contact. That largest scale is larger than the visible Universe today. The largest observable scale today would have exited the horizon somewhat (possibly much) later than the largest causal scale.

The smallest scale of inflation would be the last scale to exit the horizon at reheating. That scale is the horizon size during inflation, expanded by a factor of the redshift since reheating. If inflation occurred at the scale $\approx 10^{16}$ GeV of grand unification, then the horizon size c/H during inflation was about 10^{-29} m, Figure 10.14, and the redshift at reheating was about 4×10^{28} , so the smallest scale of inflation would be approximately $4 \times 10^{28} \times 10^{-29} \text{ m} = 0.4 \text{ m}$.

Concept question 38.2 Did scales of galaxy clustering exit the horizon early or late during inflation? The characteristic scale of galaxy clustering today was set by the sound horizon size at recombination, which the 1° peak in the CMB power spectrum, Figure 10.3, indicates to be about 1/200th of the horizon size today. Thus galaxy clustering scales exited the horizon approximately $\ln(200) = 5$ e -folds after

the horizon size today exited the horizon. Given that inflation continued for at least ~ 60 e -folds after the horizon size today exited the horizon, that is still early during inflation.

38.1 Power spectrum of vacuum fluctuations

Quantum field theory, Chapter 42, promotes fields from classical fields to operators that act on states, creating and destroying particles out of those states. Such fields include: §42 scalar (spin 0); §44 Dirac spinor (spin $\frac{1}{2}$); §45 electromagnetic (spin 1); and §46 gravitational (spin 2) fields.

Fields are defined to be linear combinations of forward-in-time destruction operators a and backward-in-time creation operators $a^\dagger$, such as a real scalar field (42.61), a Dirac spinor field (44.58), the electromagnetic field (45.8), or the gravitational field (46.13). Consider a generic field operator $\hat{\varphi}(t, k)$ in Minkowski space, taken to be a function of time t and spatial wavevector k ,

$$\hat{\varphi}(t, k) \equiv \sqrt{\frac{1}{2\omega_k}} (a_k e^{-i\omega_k t} + a_k^\dagger e^{i\omega_k t}) , \quad (38.4)$$

where $\omega_k = \sqrt{k^2 + m^2}$ for a field of rest mass m . The scalar and gravitational fields considered in this chapter are massless, for which $\omega_k = k$ with k positive. The field $\hat{\varphi}$ has zero vacuum expectation value

$$\langle 0 | \hat{\varphi}(t, k) | 0 \rangle = 0 , \quad (38.5)$$

but at given time t it has a finite power spectrum

$$\begin{aligned} \langle 0 | \hat{\varphi}(t, k') \hat{\varphi}(t, k) | 0 \rangle &= \left\langle 0 \left| \sqrt{\frac{1}{2\omega_{k'}}} (a_{k'} e^{-i\omega_{k'} t} + a_{k'}^\dagger e^{i\omega_{k'} t}) \sqrt{\frac{1}{2\omega_k}} (a_k e^{-i\omega_k t} + a_k^\dagger e^{i\omega_k t}) \right| 0 \right\rangle \\ &= \left\langle 0 \left| \sqrt{\frac{1}{2\omega_{k'}}} \sqrt{\frac{1}{2\omega_k}} a_{k'} a_k^\dagger e^{-i(\omega_{k'} - \omega_k)t} \right| 0 \right\rangle = \frac{(2\pi)^3}{2\omega_k} \delta^3(k' - k) . \end{aligned} \quad (38.6)$$

The Dirac delta function on the rightmost side of equations (38.6) comes from the commutation (or anti-commutation) rules of destruction and creation operators, such as the commutation rules (42.63) for a scalar field or (45.5) for the electromagnetic field or (46.10) for the gravitational field, or the anticommutation rules (44.56) for a Dirac spinor field.

The power spectrum $P(k)$ is defined, equation (30.131), to be the coefficient of the Dirac delta function $(2\pi)^3 \delta^3(k' - k)$. The expression (38.6) shows that the power spectrum $P_{\hat{\varphi}}(k)$ of free fields in flat space is

$$P_{\hat{\varphi}}(k) = \frac{1}{2\omega_k} . \quad (38.7)$$

The power spectra of massive fields with $\omega_k > k$ are suppressed compared to massless fields.

The higher-order moments are Gaussian, §30.22.3. For example, the 4-point function is a product of pairs

of 2-point functions,

$$\begin{aligned}
& \langle 0 | \hat{\varphi}(t, k_1) \hat{\varphi}(t, k_2) \hat{\varphi}(t, k_3) \hat{\varphi}(t, k_4) | 0 \rangle \\
&= \frac{2\pi^3 \delta^3(k_1 - k_2)}{2\omega_{k_1}} \frac{2\pi^3 \delta^3(k_3 - k_4)}{2\omega_{k_3}} + \frac{2\pi^3 \delta^3(k_1 - k_3)}{2\omega_{k_1}} \frac{2\pi^3 \delta^3(k_2 - k_4)}{2\omega_{k_2}} + \frac{2\pi^3 \delta^3(k_1 - k_4)}{2\omega_{k_1}} \frac{2\pi^3 \delta^3(k_2 - k_3)}{2\omega_{k_2}} \\
&= P_{\hat{\varphi}}(k_1 - k_2) P_{\hat{\varphi}}(k_3 - k_4) + P_{\hat{\varphi}}(k_1 - k_3) P_{\hat{\varphi}}(k_2 - k_4) + P_{\hat{\varphi}}(k_1 - k_4) P_{\hat{\varphi}}(k_2 - k_3) .
\end{aligned} \tag{38.8}$$

An FLRW spacetime is not Minkowski, but fluctuations well inside the horizon during inflation may be treated as approximately Minkowski, and these provide the seeds for fluctuations that freeze in as they exit the horizon in the accelerating Universe. In going from quantum field theory to cosmology, delta functions $(2\pi)^3 \delta^3(k' - k)$ in $k' - k$, such as in the vacuum power spectrum (38.6), should be replaced by delta functions $(2\pi)^3 \delta^3(k' + k)$ in $k' + k$. Both delta functions derive from the same underlying reason, spatial translation symmetry, which implies conservation of momentum. In quantum field theory, k' and k denote momenta of outgoing and ingoing states, which conservation requires to be equal, but in cosmology the sum $k' + k$ denotes the total momentum, which conservation requires to be zero.

38.2 Generation of gravitational waves by inflation

This section describes the generation of a potentially observable power spectrum of gravitational waves by inflation. A similar process generates fluctuations in the inflaton field φ , which seed fluctuations in radiation and matter, but calculation of the latter is more involved, and is deferred to §38.4.

Gravitational waves are described to linear order by the gauge-invariant spin 2 perturbation h_{ab} , equation (29.16d). In an FLRW universe, the gravitational wave perturbation h_{ab} satisfies the wave equation (29.65).

As seen in Chapter 43, observers always quantize fields with respect to the frequencies and wavenumbers they perceive in their own frame. In effect, a field to be quantized must satisfy a wave equation that looks like that of a simple harmonic oscillator, with a possibly variable effective mass. The wave equation (29.65) for h_{ab} does not have the desired simple harmonic form, but the wave equation (29.72) for ah_{ab} does,

$$\left(\frac{\partial^2}{\partial \eta^2} - \frac{\ddot{a}}{a} + k^2 \right) (ah_{ab}) = 0 \tag{38.9}$$

(in the absence of a tensor source), so it is ah_{ab} rather than h_{ab} that should be quantized. Furthermore, the gravitational field operator must be divided by $\sqrt{8\pi G}$ as in (46.1), to ensure that the equal-time commutator of the field and its conjugate momentum yields the required 3 dimensional delta function, equation (46.10). In all, the gravitational field operator $\hat{h}_{ab}$ to be quantized is, in terms of the classical perturbation h_{ab} ,

$$\hat{h}_{ab} = \frac{ah_{ab}}{\sqrt{8\pi G}} . \tag{38.10}$$

Beware that the most common convention in the literature is to define the tensor perturbation h_{ab} to

be the metric perturbation (29.22c) rather than the vierbein perturbation (29.10d) adopted here, and the metric perturbation is 2 (or -2 depending on sign convention) times the vierbein perturbation,

$$\underset{\text{metric perturbation}}{h_{ab}} = \underset{\text{vierbein perturbation}}{2h_{ab}} . \quad (38.11)$$

From the perspective of quantum field theory, the vierbein convention is the more natural, because analogously to the (dimensionless) coupling parameter e of electromagnetism in Gaussian units, equation (44.2), the coupling parameter of gravity at low energies is the (dimensionful) quantity $\sqrt{8\pi G}$ as in (46.1) and Exercise 46.1, and not twice that. A consequence of the present convention is that the tensor power spectrum commonly treated in the literature is 4 times that obtained here, equation (38.21).

During inflation, the Hubble parameter H is constant to a good approximation. The definition $H = a^{-2}da/d\eta$ of the Hubble parameter implies that, with constants of integration chosen such that $|\eta| \rightarrow 0$ as $a \rightarrow \infty$ (note that η is negative and increasing to zero),

$$aH = \frac{1}{|\eta|} . \quad (38.12)$$

The wave equation (38.9) becomes, equation (29.76),

$$\left(\frac{\partial^2}{\partial \eta^2} - \frac{2}{\eta^2} + k^2 \right) (ah_{ab}) = 0 . \quad (38.13)$$

Solutions of equation (38.13) are

$$ah_{ab} \propto e^{\mp ik\eta} \left(1 \pm \frac{i}{k|\eta|} \right) . \quad (38.14)$$

Well inside the horizon, $k|\eta| \gg 1$, the perturbation oscillates and redshifts,

$$h_{ab} \sim a^{-1} e^{\mp ik\eta} . \quad (38.15)$$

Inflation tends to smooth out any pre-existing primordial waves, but irreducible vacuum fluctuations remain. In a quantized field, the $\mp$ sign in equation (38.15) can be recognized as pertaining to forward-in-time and backward-in-time modes respectively. The correct normalization is imposed, equation (38.4), by the requirement that the forward-in-time quantized field $\hat{h}_{ab}$ go over to $e^{-ik\eta}/\sqrt{2k}$ well inside the horizon $k|\eta| \gg 1$, where the spacetime approximates Minkowski, thus

$$\hat{h}_{ab} = \frac{e^{-ik\eta}}{\sqrt{2k}} \left(1 + \frac{i}{k|\eta|} \right) . \quad (38.16)$$

So normalized, the perturbation h_{ab} , equation (38.10), is

$$h_{ab} = \frac{\sqrt{8\pi G}}{a} \hat{h}_{ab} = \sqrt{8\pi G} H |\eta| \hat{h}_{ab} = \frac{\sqrt{8\pi G} H}{\sqrt{2k}} e^{-ik\eta} \left(|\eta| + \frac{i}{k} \right) . \quad (38.17)$$

Once the mode disappears over the horizon, $k|\eta| \ll 1$, the perturbation h_{ab} freezes out,

$$h_{ab} \rightarrow i \frac{\sqrt{8\pi G} H}{\sqrt{2k^3}} e^{-ik\eta} . \quad (38.18)$$

The resulting power spectrum $P_{h_{ab}}(k)$ of perturbations h_{ab} is

$$P_{h_{ab}}(k) = \frac{8\pi G H^2}{2k^3} . \quad (38.19)$$

There are two polarizations ab , right-handed $++$ and left-handed $--$, which are uncorrelated with each other. The total tensor power spectrum $P_h(k)$ is twice the power in each polarization,

$$P_h(k) = \frac{8\pi G H^2}{k^3} . \quad (38.20)$$

The corresponding dimensionless power spectrum $\Delta_h^2(k)$, obtained by multiplying by $4\pi k^3/(2\pi)^3$, equation (30.133), is

$$\Delta_h^2(k) = \frac{4GH^2}{\pi} = \frac{V}{6\pi^2 m_{\text{P}}^4} = \frac{\mu^4}{6\pi^2} , \quad (38.21)$$

where $m_{\text{P}} \equiv 1/\sqrt{8\pi G}$ is the reduced Planck mass, and V is the potential energy of the inflaton field, and $\mu \equiv V^{1/4}/m_{\text{P}}$ is the characteristic energy scale of inflation. The detection of an almost scale-free spectrum of tensor modes at large scales would provide a direct measurement of the density V of the inflaton field during early inflation. Beware that, as remarked in equation (38.11), the common convention for the perturbation h_{ab} is the metric perturbation, which is twice the vierbein perturbation adopted here, so the commonly quoted tensor power spectrum is 4 times that derived here,

$$\Delta_h^2(k)_{\text{metric perturbation } h_{ab}} = \frac{4\Delta_h^2(k)}{\text{vierbein perturbation } h_{ab}} . \quad (38.22)$$

The power-spectrum (38.21) is almost but not quite scale-free, because H^2 varies slowly with k . The slow variation is conventionally characterized by a tensor index n_t , or tensor tilt, defined by

$$n_t \equiv \frac{d \ln \Delta_h^2(k)}{d \ln k} \approx \frac{d \ln(H^2)}{d \ln k} . \quad (38.23)$$

A mode of wavenumber k exits the horizon when $k|\eta| \approx 1$, and $|\eta| \approx 1/(aH)$ during early inflation, so $d \ln k \approx -d \ln \eta \approx d \ln a$, and

$$n_t \approx \frac{d \ln(H^2)}{d \ln a} = -2\epsilon_1 , \quad (38.24)$$

the final expression coming from the definition (38.34) of ϵ_1 . The slow-roll parameter ϵ_1 is expected to be slightly positive, so the tensor tilt n_t is expected to be slightly negative. A generic prediction of inflation is that the tensor index should be related to the ratio r of tensor to scalar power, equation (38.77), by

$$n_t \approx -\frac{r}{8} . \quad (38.25)$$

38.3 Slow-roll inflation

For simplicity, this section deals with the case of inflation driven by a single scalar field φ . It is possible to consider inflation driven by multiple scalar fields, but that is more complicated and a priori less likely. If the potential energy curve $V(\varphi)$ of the scalar field is sufficiently flat, then the scalar field will roll slowly down towards the bottom of the potential well, where reheating takes place. The physics of slow-roll inflation is reviewed by Lyth and Riotto (1999) and Baumann (2009). A pedagogical resource is Baumann (2022).

The Lagrangian of a scalar field φ with an unspecified potential energy $V(\varphi)$ is given by equation (42.47). The Euler-Lagrange equations of motion are, similarly to the derivation of equation (42.51),

$$\square\varphi = \frac{\partial V}{\partial\varphi} , \quad (38.26)$$

where $\square \equiv D_m D^m$ is covariant d'Alembertian operator. Equation (38.26) is a wave equation sourced by the slope $\partial V/\partial\varphi$ of the potential energy.

The energy-momentum of the scalar field is given by equation (16.120), yielding

$$T_{mn} = (\partial_m\varphi)(\partial_n\varphi) - \left[\frac{1}{2}(\partial_k\varphi)(\partial^k\varphi) + V\right]\eta_{mn} . \quad (38.27)$$

The energy-momentum (38.27) is that of a perfect fluid, equation (30.2), which may be written

$$T^{mn} = (\rho + p)u^m u^n + \left[\frac{1}{2}(\rho + p) - V\right]\eta^{mn} , \quad (38.28)$$

with proper density plus pressure $\rho + p$ and 4-velocity u^m , satisfying $u_m u^m = -1$, given by

$$\rho + p = -(\partial_m\varphi)(\partial^m\varphi) , \quad u^m = \frac{\partial^m\varphi}{\sqrt{\rho + p}} . \quad (38.29)$$

38.3.1 Unperturbed

Consider first the unperturbed situation of an FLRW universe driven by the scalar field φ . Perturbations to the scalar field will be addressed in §38.4. The scalar field is spatially uniform, and the scalar wave equation (38.26) is

$$\left(\frac{d^2}{dt^2} + 3H\frac{d}{dt}\right)\varphi = -\frac{\partial V}{\partial\varphi} , \quad (38.30)$$

where H is the Hubble parameter. The FLRW time t is the proper time in the rest frame of the scalar field, and the tetrad-frame 4-velocity in the perfect fluid energy-momentum (38.28) is the rest-frame velocity $u^m = \{1, 0, 0, 0\}$. The energy and momentum (38.27) reduce in the rest frame to

$$\rho = \frac{1}{2}\left(\frac{d\varphi}{dt}\right)^2 + V , \quad (38.31a)$$

$$p = \frac{1}{2}\left(\frac{d\varphi}{dt}\right)^2 - V . \quad (38.31b)$$

If the potential energy V dominates, then the equation of state is that of vacuum energy, $p/\rho = -1$, with energy density V . If the kinetic energy $\frac{1}{2}(d\varphi/dt)^2$ dominates, then the equation of state is that of an ultrahard fluid, $p/\rho = 1$, whose sound speed $\sqrt{p/\rho}$ equals the speed of light. It is thought that during reheating, oscillations of the scalar field couple to other species, and the scalar field tends to decay into lighter species of matter and radiation.

The equation (29.34a) of conservation of energy is

$$D_m T^{m0} = \frac{d\rho}{dt} + 3(\rho + p)H = 0, \quad (38.32)$$

where the Hubble parameter H is related to the density ρ in a flat universe by the first Friedmann equation

$$H^2 = \frac{8}{3}\pi G\rho. \quad (38.33)$$

Inserting the expressions (38.31) for ρ and p into the energy conservation equation (38.32) recovers the equation of motion (38.30) for φ .

Slow-roll parameters ϵ_i and η_i are conventionally defined by

$$\epsilon_i \equiv -\frac{d \ln H}{d \ln a}, \quad \eta_i \equiv 2\epsilon_i - \frac{1}{2} \frac{d \ln \epsilon_i}{d \ln a}, \quad (38.34)$$

which slow-roll inflation requires to be small, that is, H changes little over a large number of e -folds of the cosmic scale factor a , and likewise ϵ_i itself changes little over many e -folds. The energy conservation equation (38.32), along with the first Friedmann equation (38.33), implies that the slow-roll parameter ϵ_i satisfies

$$\epsilon_i = -\frac{1}{2} \frac{d \ln \rho}{d \ln a} = -\frac{1}{2H\rho} \frac{d\rho}{dt} = \frac{3}{2} \frac{\rho + p}{\rho} = \frac{3}{2} \frac{(d\varphi/dt)^2}{\rho} = \frac{8\pi G}{2} \left(\frac{1}{H} \frac{d\varphi}{dt} \right)^2. \quad (38.35)$$

The smallness of ϵ_i requires the kinetic energy $\frac{1}{2}(d\varphi/dt)^2$ to be small compared to the potential energy V . The slow-roll parameter η_i satisfies

$$\eta_i = -\frac{d \ln(Hd\varphi/dt)}{d \ln a} = \epsilon_i - \frac{d^2\varphi/dt^2}{Hd\varphi/dt}. \quad (38.36)$$

As will be shown momentarily, during the slow-roll phase of inflation, ϵ_i and η_i approximate ϵ_V and η_V defined in terms of the first and second derivatives of the potential V by

$$\epsilon_V \equiv \frac{1}{16\pi G} \left(\frac{\partial \ln V}{\partial \varphi} \right)^2, \quad \eta_V \equiv \frac{1}{8\pi G V} \frac{\partial^2 V}{\partial \varphi^2}. \quad (38.37)$$

The smallness of η_i , along with that of ϵ_i , implies that $|d^2\varphi/dt^2| \ll Hd\varphi/dt$, which implies from the equation of motion (38.30) that

$$H \frac{d\varphi}{dt} \approx -\frac{1}{3} \frac{\partial V}{\partial \varphi}. \quad (38.38)$$

The approximation (38.38) implies that the slow-roll parameter ϵ_i given by equation (38.35) approximates

$$\epsilon_i \approx \epsilon_V. \quad (38.39)$$

Similarly the slow-roll parameter η_i given by equation (38.36) approximates

$$\eta_i \approx \frac{1}{3H^2 d\varphi/dt} \frac{d}{dt} \frac{\partial V}{\partial \varphi} \approx \eta_V . \quad (38.40)$$

Since ϵ_V defined by equation (38.37) is positive, so also ϵ_i is expected to be positive. Slow roll requires $\epsilon_i \approx \epsilon_V$ and $\eta_i \approx \eta_V$ to be small,

$$\epsilon_V \ll 1 , \quad |\eta_V| \ll 1 . \quad (38.41)$$

Inversion of the CMB power spectrum measured with the Planck satellite yields an upper limit $\epsilon_V \lesssim 0.6$, and a marginal detection $\eta_V = -0.010 \pm 0.006$ (Planck Collaboration, 2020b, p. 15).

Given the slow-roll conditions (38.41) along with $H^2 \approx \frac{8}{3}\pi G V$, equation (38.38) integrates to give an equation that relates the number $N = \Delta \ln a$ of e -folds as the scalar field φ rolls down the hill,

$$N \approx 8\pi G \int \frac{d\varphi}{-\partial \ln V / \partial \varphi} . \quad (38.42)$$

Exercise 38.3 Inflation driven by a scalar field with a Higgs-like potential. Consider inflation driven by an inflaton field φ with a Higgs-like potential. What are the slow-roll parameters ϵ_V and η_V defined by equations (38.34)? Can a Higgs-like potential satisfy the slow-roll conditions (38.41)? Solve equation (38.42) for the number of e -folds produced by the Higgs-like inflaton field. The potential energy of a Higgs-like field, one like that of the electroweak Higgs field, is to leading (tree-level) order an even quartic polynomial in φ ,

$$V(\varphi) = \lambda(\varphi^2 - \varphi_0^2)^2 = \frac{m_H^4}{64\lambda} - \frac{m_H^2}{4}\varphi^2 + \lambda\varphi^4 , \quad \varphi_0 = \frac{m_H}{2\sqrt{2\lambda}} , \quad (38.43)$$

where m_H is the mass of the Higgs boson, and the quartic coefficient λ is dimensionless. The Higgs potential (38.43) takes its minimum value, zero, at the nonzero value $|\varphi| = \varphi_0$ of the field, where the potential $V \approx \frac{1}{2}m_H^2(|\varphi| - \varphi_0)^2$ approximates a simple harmonic potential for a scalar field of mass m_H . Near the beginning of inflation, φ starts out small, and the derivative of the Higgs potential (38.43) is to leading order $\partial V / \partial \varphi = -\frac{1}{2}m_H^2\varphi$.

Solution. The slow-roll parameter η_V defined by equation (38.34) is

$$\eta_V = -\frac{32\lambda m_P^2}{m_H^2} \quad (38.44)$$

where $m_P = 1/\sqrt{8\pi G} = 2.4353 \times 10^{18}$ GeV is the reduced Planck mass. The slow-roll parameter (38.44) is small provided either that the Higgs mass m_H greatly exceeds the reduced Planck mass m_P , or else that the dimensionless quartic coefficient λ is tiny. The slow-roll parameter ϵ_V defined by equation (38.34) is

$$\epsilon_V = 2|\eta_V| \left(\frac{\varphi}{\varphi_0} \right)^2 , \quad (38.45)$$

which, given that φ must be less than its value φ_0 at the minimum of the potential, is small provided that $|\eta_V|$ is small. The integral (38.42) gives

$$N = \frac{1}{2|\eta_V|} [\ln(\varphi_0/\varphi_i)] , \quad (38.46)$$

where φ_i and φ_0 are the values of φ at the beginning and end of inflation. If $|\eta_V|$ is small, then the number N of e -folds is large.

The electroweak Higgs field has $m \approx 125 \text{ GeV}$ and $\lambda \approx 0.13$ (Particle Data Group, 2024, p. 212). The smallness of the electroweak Higgs boson mass relative to the reduced Planck mass, $m/m_P \approx 5 \times 10^{-17}$, makes the electroweak Higgs unpromising as a candidate for the inflaton. A Higgs field at the grand unified scale rather than the electroweak scale would be more promising as the inflaton, because m/m_P would be larger. Interestingly, theoretical calculations of the running of the electroweak quartic coefficient λ with energy suggests that it becomes small as the renormalization energy scale μ grows large, possibly passing from positive (stable) to negative (unstable) at a value that may be of order $\mu \sim 10^{10} \text{ GeV}$, depending on the precise values of the Higgs mass and the top-quark mass (Elias-Miró et al., 2012; Bezrukov et al., 2012; Degraasi et al., 2012; Buttazzo et al., 2013; Bednyakov et al., 2015).

38.4 Generation of scalar waves by inflation

Inflation predicts that fluctuations $\delta\varphi$ in the inflaton field φ generated waves that seeded the scalar fluctuations seen today in the cosmic microwave background and in galaxy clustering. As with the gravitational field, the scalar wave equation for $\delta\varphi$ does not have the simple harmonic form required for quantization, but the wave equation for $a\delta\varphi$ does, the **Mukhanov-Sasaki** wave equation (38.65), similar to the tensor wave equation (38.9) for ah_{ab} . Thus the scalar field operator $\hat{\varphi}$ to be quantized is

$$\hat{\varphi} = a\delta\varphi . \quad (38.47)$$

This is similar to the gravitational field operator $\hat{h}$, equation (38.10), but without the $1/\sqrt{8\pi G}$ factor.

It is helpful to define an auxiliary slow-roll parameter δ_i by

$$\delta_i \equiv \frac{3}{2}(1 + \bar{w}) = -\frac{\partial \ln(\dot{\varphi}/a)}{\partial \ln a} , \quad (38.48)$$

where $\bar{w} \equiv \dot{p}/\dot{\rho}$ is the ‘equation of state’ of the mean scalar field $\bar{\varphi}$. Here, as usual in this part of the book, an overdot denotes differentiation $\partial/\partial\eta$ with respect to conformal time η . The equation of state is expected to exceed slightly the vacuum value, $\bar{w} \gtrsim -1$, so δ_i is expected to be slightly positive. The slow roll parameter δ_i is related to ϵ_i defined by equation (38.34) and η_i given by (38.36) by

$$\delta_i = \epsilon_i - \eta_i = -\frac{\partial \ln(\dot{\varphi}/a)}{\partial \ln a} . \quad (38.49)$$

The slow-roll parameters ϵ_i , η_i , and δ_i are all small during the early slow-roll phase of inflation, but cease to be small as $\bar{\varphi}$ rolls down the hill and inflation comes to an end.

The energy-momentum of the scalar field is given by equation (38.28). The perturbation, $\overset{1}{T}{}^{mn} \equiv T^{mn} - \overset{0}{T}{}^{mn}$, the difference between the energy-momentum and its value in the background FLRW frame where $u^m = \{1, 0, 0, 0\}$, is

$$\delta\rho \equiv \overset{1}{T}{}^{00} = (\bar{\rho} + \bar{p})u_a u^a + \frac{\partial\bar{V}}{\partial\bar{\varphi}}\delta\varphi, \quad (38.50a)$$

$$f_a \equiv \overset{1}{T}{}^{0a} = (\bar{\rho} + \bar{p})u^0 u^a, \quad (38.50b)$$

$$\delta p_{ab} \equiv \overset{1}{T}{}^{ab} = (\bar{\rho} + \bar{p})u^a u^b - \delta_{ab} \frac{\partial\bar{V}}{\partial\bar{\varphi}}\delta\varphi. \quad (38.50c)$$

The spatial velocity u^a is small, and the perturbation $\delta\varphi$ of the scalar field is small, and during the slow-roll phase of inflation the potential derivative $\partial\bar{V}/\partial\bar{\varphi}$ is small, equation (38.39), so the perturbed density and pressure $\delta\rho$ and δp_{ab} are of second order of perturbations, while the energy flux f_a is of first order.

Like any tetrad-frame quantity that vanishes in the background, the energy flux f_a is coordinate gauge-invariant, §26.7, but it is not tetrad gauge-invariant, since it can be adjusted by a tetrad transformation that changes the spatial velocity u^a . Denote the scalar component of the spatial velocity u^a by $u_a = \nabla_a u$. The scalar velocity u , which like the vierbein perturbation w , equation (29.10b), has units of length, is related to the dimensionless scalar velocity v previously defined by equation (30.12) by $u = v/k$. In a completely general gauge, to linear order the velocity $\nabla_a u$ is related to the scalar perturbation $\delta\varphi$ by

$$\nabla_a u \equiv \frac{f_a}{\bar{\rho} + \bar{p}} = \frac{a\partial_a\varphi}{\dot{\bar{\varphi}}} = \nabla_a \left(\frac{\delta\varphi}{\dot{\bar{\varphi}}} - \tilde{w} \right). \quad (38.51)$$

Equation (38.51) can be arranged using the expression (29.16b) for Φ into a form that is coordinate and tetrad gauge-invariant on both sides of the equation,

$$\frac{\dot{a}}{a} \frac{\delta\varphi}{\dot{\bar{\varphi}}} + \phi = \frac{\dot{a}}{a} (u + \tilde{w}) + \phi = \frac{\dot{a}}{a} (u - w + \dot{h}) + \Phi. \quad (38.52)$$

The rightmost side of equations (38.52) is coordinate and tetrad gauge-invariant because u , $w - \dot{h}$ (one of the combinations (29.15)), and Φ are all coordinate gauge-invariant, and $u - w$, $\dot{h}$, and Φ are all tetrad gauge-invariant. The quantity $u - w$ is tetrad gauge-invariant because when the tetrad is boosted to another frame, the velocity $\nabla_a u$ and the vierbein perturbation $\nabla_a w$, equation (29.10b), change by the same velocity. Since the right hand side of equation (38.52) is coordinate and tetrad gauge-invariant, so also is the left hand side. The right hand side of equation (38.52) can be interpreted as minus the primordial curvature fluctuation $\zeta \equiv \delta - \Phi$ defined by equation (29.42). Thus the coordinate and tetrad gauge-invariant density perturbation δ defined by equation (29.40) is

$$\delta = -\frac{\dot{a}}{a} (u - w + \dot{h}), \quad (38.53)$$

and the primordial curvature fluctuation ζ itself equals minus the left or right hand side of equation (38.52),

$$\zeta = -\frac{\dot{a}}{a} \frac{\delta\varphi}{\dot{\bar{\varphi}}} + \phi = \delta - \Phi. \quad (38.54)$$

Equations (38.53) and (38.54) are valid in any gauge, but a particularly advantageous choice of gauge is spatially flat gauge, §29.10, in which $\phi = h = 0$ and $\Phi = (\dot{a}/a)w$, and the primordial curvature fluctuation ζ simplifies to

$$\zeta = -\frac{\dot{a}}{a} \frac{\delta\varphi}{\dot{\bar{\varphi}}} = \frac{\dot{a}}{a} u . \quad (38.55)$$

The primordial curvature fluctuation ζ can also be expressed in terms of the slow-roll parameter ϵ_1 given by the final part of equations (38.35),

$$\zeta = -\sqrt{\frac{4\pi G}{\epsilon_1}} \delta\varphi . \quad (38.56)$$

To linear order, the scalar wave equation for a scalar field φ with potential energy $V(\varphi)$ is, in spatially flat gauge,

$$\begin{aligned} 0 &= D^\mu \partial_\mu \varphi - \frac{\partial V}{\partial \varphi} \\ &= -\frac{1-2\psi}{a^2} \left[\left(\frac{\partial^2}{\partial \eta^2} + 2\frac{\dot{a}}{a} \frac{\partial}{\partial \eta} - \nabla^2 \right) \varphi - \left(\frac{\partial \psi}{\partial \eta} + \nabla^2 w \right) \frac{\partial \bar{\varphi}}{\partial \eta} + a^2(1+2\psi) \frac{\partial V}{\partial \varphi} \right] . \end{aligned} \quad (38.57)$$

The wave equation (38.57) involves the two scalar potentials ψ and w of spatially flat gauge. The two potentials can be eliminated using the energy and scalar momentum Einstein equations. To do so, it is advantageous to use the Einstein equations of the background FLRW spacetime,

$$\overset{0}{G}_{00} + \overset{0}{G}_{aa} = 2\epsilon_1 H^2 = 8\pi G(\bar{\rho} + \bar{p}) = 8\pi G \dot{\bar{\varphi}}^2 , \quad (38.58)$$

to replace $8\pi G = 2\epsilon_1 H^2 / \dot{\bar{\varphi}}^2$ in the perturbed Einstein equations. The perturbed Einstein energy and scalar momentum equations are

$$\frac{a^2}{2} (\overset{1}{G}_{00} - 8\pi G \overset{1}{T}_{00}) = \frac{\dot{a}}{a} \left(-3\frac{\dot{a}}{a} \psi + \nabla^2 w \right) - \epsilon_1 \frac{\dot{a}^2}{a^2} \left(\psi - \frac{\delta\varphi}{\dot{\bar{\varphi}}} - \frac{a^2}{\dot{\bar{\varphi}}^2} \frac{\partial V}{\partial \varphi} \delta\varphi \right) , \quad (38.59a)$$

$$\frac{a^2}{2} (\overset{1}{G}_{0a} - 8\pi G \overset{1}{T}_{0a}) = \frac{\dot{a}}{a} \nabla_a \psi - \epsilon_1 \frac{\dot{a}^2}{a^2} \frac{1}{\dot{\bar{\varphi}}} \nabla_a \varphi , \quad (38.59b)$$

which imply that the scalar potentials ψ and w are

$$\nabla^2 w = 3\frac{\dot{a}}{a} \psi + \epsilon_1 \frac{\dot{a}}{a} \left(\psi - \frac{\delta\varphi}{\dot{\bar{\varphi}}} - \frac{a^2}{\dot{\bar{\varphi}}^2} \frac{\partial V}{\partial \varphi} \delta\varphi \right) , \quad (38.60a)$$

$$\psi = \epsilon_1 \frac{\dot{a}}{a} \frac{\delta\varphi}{\dot{\bar{\varphi}}} . \quad (38.60b)$$

To bring the wave equation into a simple harmonic form suitable for quantization, recast the wave equation (38.57) for φ into a wave equation for $a\varphi$,

$$0 = \left(\frac{\partial^2}{\partial \eta^2} - \frac{\ddot{a}}{a} - \nabla^2 \right) (a\varphi) - \epsilon_1 \frac{\dot{a}}{a} (a\delta\varphi) + a^3(1+2\psi) \frac{\partial V}{\partial \varphi} . \quad (38.61)$$

The derivative $\partial V/\partial\varphi$ of the potential can be replaced by its expansion to linear order,

$$\frac{\partial V}{\partial\varphi} = \frac{\partial\bar{V}}{\partial\bar{\varphi}} + \frac{\partial^2\bar{V}}{\partial\bar{\varphi}^2}\delta\varphi, \quad (38.62)$$

where the first derivative $\partial\bar{V}/\partial\bar{\varphi}$ comes from the unperturbed scalar wave equation (38.30), and the second derivative $\partial^2\bar{V}/\partial\bar{\varphi}^2$ from differentiating that,

$$a^2 \frac{\partial\bar{V}}{\partial\bar{\varphi}} = -\frac{\partial^2\bar{\varphi}}{\partial\eta^2} - 2\frac{\dot{a}}{a} \frac{\partial\bar{\varphi}}{\partial\eta} = -H^2(3 - \delta_i) \frac{\partial\bar{\varphi}}{\partial\ln a}, \quad (38.63)$$

$$a^2 \frac{\partial^2\bar{V}}{\partial\bar{\varphi}^2} = H^2 \left((3 - \delta_i)(\epsilon_i + \delta_i) + \frac{\partial\delta_i}{\partial\ln a} \right). \quad (38.64)$$

Subtracting the unperturbed part $(\partial^2/\partial\eta^2 - \ddot{a}/a)(a\bar{\varphi}) + a^3\partial\bar{V}/\partial\bar{\varphi}$ from the wave equation (38.61) yields (after some algebra) the Mukhanov-Sasaki equation,

$$0 = \left(\frac{\partial^2}{\partial\eta^2} - \frac{\ddot{b}}{b} + k^2 \right) (a\delta\varphi), \quad b \equiv a \frac{\partial\bar{\varphi}}{\partial\ln a}, \quad (38.65)$$

where the comoving Laplacian has been replaced by the squared comoving Fourier wavenumber, $\nabla^2 = -k^2$. The Mukhanov-Sasaki equation (38.65) is valid to linear order in the scalar perturbation $a\delta\varphi$ and in the perturbations ψ and w to the gravitational field, but it does *not* assume that the mean field $\bar{\varphi}$ is small, or that the slow-roll parameters ϵ_i and δ_i are small (as they are during early inflation). The Mukhanov-Sasaki equation (38.65) is valid throughout inflation, not just in the slow-roll limit.

Modes that formed large scale structure seen today exited the horizon early during inflation, Concept question 38.2. During early inflation,

$$\frac{\ddot{b}}{b} \approx \frac{\ddot{a}}{a} \approx \frac{2}{\eta^2}. \quad (38.66)$$

Well inside the horizon, $k|\eta| \gg 1$, the perturbation $\delta\varphi$ oscillates and redshifts, similarly to the perturbations (38.15) of the gravitational field h_{ab} ,

$$\delta\phi \sim a^{-1} e^{\mp ik\eta}. \quad (38.67)$$

Inflation tends to smooth out any pre-existing primordial waves, but irreducible vacuum fluctuations remain. The correct normalization is imposed, equation (38.4), by the requirement that the forward-in-time quantized field $\hat{\varphi}$ go over to $e^{-ik\eta}/\sqrt{2k}$ well inside the horizon $k|\eta| \gg 1$, where the spacetime approximates Minkowski, thus

$$a\delta\varphi \rightarrow \frac{e^{-ik\eta}}{\sqrt{2k}}. \quad (38.68)$$

Once outside the horizon, $k|\eta| \ll 1$, the Mukhanov-Sasaki equation (38.65) goes over to

$$\frac{(a\ddot{\delta\varphi})}{a\delta\varphi} = \frac{\ddot{b}}{b}. \quad (38.69)$$

The solution to equation (38.69) is

$$a\delta\varphi \propto b . \quad (38.70)$$

The ratio $a\delta\varphi/b$,

$$\frac{a\delta\varphi}{b} = \frac{\dot{a}}{a} \frac{\delta\varphi}{\dot{\varphi}} = -\zeta , \quad (38.71)$$

equals minus the primordial curvature fluctuation. Thus the primordial curvature fluctuation ζ freezes out once the mode exits the horizon, and this remains true throughout the epoch of inflation, not just during the early slow-roll phase. Once the scalar field φ has rolled down the hill, it will oscillate around the minimum of the potential V , coupling to other fields and generating radiation and matter in the poorly understood phase of reheating. But the primordial curvature fluctuation ζ remains constant as long as the mode is outside the horizon, setting the initial conditions for the radiation and matter fields when they come back inside the horizon.

The large-scale modes that generated the large-scale structure of the Universe today exited the horizon early during inflation, when b was approximately proportional to a , equation (38.66). During this phase, the forward-in-time solution of the Mukhanov-Sasaki wave equation (38.65) is, similar to the solution (38.16) for gravitational waves, correctly normalized,

$$a\delta\varphi \approx \frac{e^{-ik\eta}}{\sqrt{2k}} \left(1 + \frac{i}{k|\eta|} \right) . \quad (38.72)$$

Once the mode disappears over the horizon, $k|\eta| \ll 1$, the perturbation φ freezes out,

$$\delta\varphi \rightarrow i \frac{H}{\sqrt{2k^{3/2}}} e^{-ik\eta} , \quad (38.73)$$

which is similar to the gravitational wave solution (38.18), but without the $\sqrt{8\pi G}$ factor. The resulting power spectrum $P_\varphi(k)$ of scalar perturbations $\delta\varphi$ is

$$P_\varphi(k) = \frac{H^2}{2k^3} . \quad (38.74)$$

The initial conditions for the cosmic microwave background and large scale structure are set by the primordial curvature fluctuation ζ , which is related to $\delta\varphi$ by equation (38.56),

$$P_\zeta(k) = \frac{2\pi G H^2}{\epsilon_1 k^3} . \quad (38.75)$$

The corresponding dimensionless power spectra $\Delta_\varphi^2(k)$ and $\Delta_\zeta^2(k)$, obtained by multiplying by $4\pi k^3/(2\pi)^3$, equation (30.133), are

$$\boxed{\Delta_\varphi^2(k) = \frac{H^2}{4\pi^2} , \quad \Delta_\zeta^2(k) = \frac{G H^2}{\pi \epsilon_1} .} \quad (38.76)$$

The ratio of gravitational wave power (38.20) (multiplied by 4 to accord to the common convention, equation (38.22)), to primordial curvature power (38.75) defines the tensor-to-scalar ratio r ,

$$r \equiv \frac{4P_h(k)}{P_\zeta(k)} = 16\epsilon_i . \quad (38.77)$$

The power-spectrum (38.76) is almost but not quite scale-free, because H^2/ϵ_i varies slowly with k . The slow variation is conventionally characterized by a scalar index n_s defined by

$$n_s - 1 \equiv \frac{d \ln \Delta_\zeta^2(k)}{d \ln k} \approx \frac{d \ln(H^2/\epsilon_i)}{d \ln k} . \quad (38.78)$$

A mode of wavenumber k exits the horizon when $k|\eta| \approx 1$, and $|\eta| \approx 1/(aH)$ during early inflation, so $d \ln k \approx -d \ln \eta \approx d \ln a$, and

$$n_s - 1 \approx \frac{d \ln(H^2/\epsilon_i)}{d \ln a} \approx -2\delta_i , \quad (38.79)$$

the final expression coming from the definitions (38.34) of ϵ_i and η_i and the definition (38.48) of δ_i . The slow-roll parameter δ_i is expected to be slightly positive, so the scalar index n_s is expected to be slightly less than 1. The value of n_s deduced from observations is measured around a certain pivot scale, equation (30.136). The corresponding value of δ_i should approximate that when those modes exited the horizon.

PART THREE

SPINORS, GEOMETRIC ALGEBRAS, THE STANDARD
MODEL, QUANTUM FIELD THEORY, AND BEYOND

Representations of Lie groups

The gauge groups of physics are Lie groups. The fields that the Lie groups act on constitute representations. For example, fermions transform in the spinor representation of the Lie group, while gauge bosons transform in the adjoint representation of the group. This chapter describes Lie groups and their representations. The account in this chapter is compressed; several results are quoted without proof. See Slansky (1981) for a pedagogical review in the context of unification.

39.1 Lie group and Lie algebra

The gauge groups of physics are continuous groups G that act linearly on fields, preserving inner products. That means symmetry transformations are unitary, and generators of symmetry transformations are Hermitian or skew-Hermitian. The symmetry groups are then Lie groups, whose D generators S_A , $A = 1, \dots, D$, satisfy commutation relations of the form

$$[S_A, S_B] = f_{ABC} S_C . \quad (39.1)$$

The complex coefficients f_{ABC} are called the structure coefficients of the group. The dimension of a Lie group G equals the number D of its generators,

$$\dim(G) = D . \quad (39.2)$$

The algebra consisting of linear combinations of generators S_A with multiplication rule defined by equation (39.1) is called the Lie algebra of the group. Elements of the Lie group G are exponentials $\exp(\theta_A S_A)$ of the Lie algebra (cf. §13.8). The Lie algebra is in a sense the logarithm of the Lie group.

39.2 Classification of finitely generated simple Lie groups

The classification of all finitely generated simple Lie groups was completed by Cartan in 1894 (see Hawkins (2000) for a historical review), and made transparent by Dynkin in the 1950s (Dynkin, 1957b; Dynkin,

1957a). A comprehensive reference is Yamatsu (2017). A simple group is one that cannot be expressed as a non-trivial product of groups. There are four infinite sequences of finitely generated simple Lie groups, denoted A_n , B_n , C_n , and D_n by Cartan, with $n = 1, 2, \dots, \infty$ the rank of the group. The **rank** of a group is the number of mutually commuting generators of the group. Cartan's groups A_n , B_n , C_n , and D_n are related to the traditional special unitary (SU), spin (Spin), and symplectic (Sp) groups by

$$A_n = \text{SU}(n+1) , \quad B_n = \text{Spin}(2n+1) , \quad C_n = \text{Sp}(n) , \quad D_n = \text{Spin}(2n) . \quad (39.3)$$

In addition, there are 5 exceptional groups, denoted

$$G_2 , F_4 , E_6 , E_7 , \text{ and } E_8 \quad (39.4)$$

where the subscript denotes the rank, 2, 4, 6, 7, and 8.

Beware that there is no universal convention for denoting the symplectic group; some authors use $\text{Sp}(2n)$, and others $\text{USp}(2n)$, in place of $\text{Sp}(n)$. The present convention follows Cornwell (1997, p. 150).

Let S_A , $A = 1, \dots, D$ be the generators of a continuous group G , and let v_i , $i = 1, \dots, d$ be a set of d linearly independent vectors that transform linearly into each other under the action of the group,

$$S_A : v_i \rightarrow (S_A)_{ij} v_j . \quad (39.5)$$

The d vectors v_i and the accompanying set of $d \times d$ matrices $(S_A)_{ij}$ define a d -dimensional representation of the group. The dimension of the representation is defined to be the dimension d of the vector space,

$$\dim(\text{rep}) = d . \quad (39.6)$$

A representation is said to be irreducible if the vector space contains no proper non-trivial subset of vectors that transform exclusively among each other under the action of the group. Physicists often refer to a representation by its dimension. For example, the spinor grade p representations of $\text{SU}(5)$, the columns of the $\text{Spin}(10)$ chart (48.15), are 1, 5, 10, and their conjugates $\bar{10}$, $\bar{5}$, $\bar{1}$.

The **adjoint representation** of a group is the special representation where the vectors upon which the group acts are the group generators themselves. The generators of a Lie group act on each other by commutation, and the adjoint representation is the set of matrices $(S_A)_{BC}$ satisfying

$$[S_A, S_B] = (S_A)_{BC} S_C . \quad (39.7)$$

Evidently the matrices of the adjoint representation are equal to the structure coefficients, $(S_A)_{BC} = f_{ABC}$. The dimension of the adjoint representation of a Lie group equals the dimension of the group itself, the number of its generators,

$$\dim(\text{adj}) = \dim(G) = \begin{cases} n(n+2) & \text{SU}(n+1) \\ n(2n+1) & \text{Spin}(2n+1) \\ n(2n+1) & \text{Sp}(n) \\ n(2n-1) & \text{Spin}(2n) \end{cases} . \quad (39.8)$$

Given a group, it is always possible to choose a basis of group generators that is orthonormal in the sense that the trace of matrix products of generators S_A and S_B is proportional to the unit $D \times D$ matrix δ_{AB}

in any representation of the group (this assertion is not self-evidently true; but it is manifestly true in the examples based on $\text{Spin}(N)$, equation (39.12), and its subgroups considered below),

$$\text{Tr}(S_A S_B) \equiv (S_A)_{ij} (S_B)_{ji} = S_2(\text{rep}) \delta_{AB} . \quad (39.9)$$

The constant of proportionality defines the Dynkin index $S_2(\text{rep})$, a real number whose value depends on the representation. The structure coefficients f_{ABC} in an orthonormal basis are totally antisymmetric.

In an orthonormal basis, the antisymmetry of the structure coefficients implies that the sum $\sum_A S_A^2$ of matrix products of generators commutes with all generators, and is therefore proportional to the unit $d \times d$ matrix,

$$\sum_{A=1}^D S_A^2 \equiv \sum_{A=1}^D (S_A)_{ij} (S_A)_{jk} = C_2(\text{rep}) \delta_{ik} . \quad (39.10)$$

The coefficient $C_2(\text{rep})$ is called the quadratic Casimir invariant of the representation. For $\text{Spin}(N)$, the quadratic Casimir invariant is half the total angular momentum squared of the representation. Equating the trace of equation (39.9) over D generator indices A with the trace of equation (39.10) over d vector indices i implies that the Dynkin index S_2 is related to the Casimir invariant C_2 by

$$S_2(\text{rep}) = \frac{\dim(\text{rep})}{\dim(G)} C_2(\text{rep}) . \quad (39.11)$$

The orthonormal generators of $\text{Spin}(N)$ are its $N(N-1)/2$ bivectors

$$\frac{1}{2} \gamma_{[ab]} \equiv \frac{1}{2} \gamma_a \wedge \gamma_b \quad (39.12)$$

with distinct indices a and b running over 1 to N . The normalization factor of $\frac{1}{2}$ is introduced so that charges of eigenvectors upon which the generators act differ by integer increments, for example equation (39.20). The non-vanishing commutators of the orthonormal bivectors are

$$[\frac{1}{2} \gamma_{[ab]}, \frac{1}{2} \gamma_{[bc]}] = \frac{1}{2} \gamma_{[ac]} . \quad (39.13)$$

The commutators (39.13) imply that the non-vanishing structure coefficients are

$$f_{[ab][bc][ca]} = -1 \quad (39.14)$$

for any $a \neq b \neq c \neq a$. The structure coefficients $f_{[ab][bc][ca]}$ are totally antisymmetric in their indices $[ab]$, $[bc]$, and $[ca]$.

The bivector generators of $\text{Spin}(N)$ are operators that act on the vectors of a representation. To characterize and construct representations, it is advantageous to work with the chiral representation of the bivector generators, since these provide raising and lowering operators that connect the vectors of a representation. $\text{Spin}(N)$ has $n \equiv [N/2]$ mutually commuting bivector generators, whose eigenvalues are n conserved charges. In place of orthonormal indices $a = 1, \dots, N$, the chiral bivectors of $\text{Spin}(N)$ use chiral indices i and $\bar{i}$ with

$i = 1, \dots, n$, plus a final index N when N is odd. In a chiral basis, the $N(N-1)/2$ bivectors collect into:

$$\frac{1}{2}\gamma_{[i\bar{i}]} \quad n \text{ diagonal bivectors that measure charge } i, \quad (39.15a)$$

$$\frac{1}{2}\gamma_{[ij]}, \frac{1}{2}\gamma_{[i\bar{j}]}, \frac{1}{2}\gamma_{[\bar{i}j]}, \frac{1}{2}\gamma_{[\bar{i}\bar{j}]} \quad 2n(n-1) \text{ bivectors that raise and/or lower 2 charges } i \text{ and } j, \quad (39.15b)$$

$$\frac{1}{2}\gamma_{[iN]}, \frac{1}{2}\gamma_{[\bar{i}N]} \quad \text{if } N \text{ is odd, } 2n \text{ bivectors that raise or lower 1 charge } i. \quad (39.15c)$$

Mathematicians call the chiral generators (39.15) of the Lie algebra the **roots** of the Lie algebra, because all the properties of the Lie algebra sprout from its roots. The non-vanishing commutators of the chiral bivectors are

$$[\frac{1}{2}\gamma_{[i\bar{j}]}, \frac{1}{2}\gamma_{[j\bar{k}]}] = \frac{1}{2}\gamma_{[i\bar{k}]}, \quad (39.16)$$

and the same with $i \leftrightarrow \bar{i}$ and/or $j \leftrightarrow \bar{j}$ and/or $k \leftrightarrow \bar{k}$, and allowing $i = j$ or $j = k$ or $k = i$, but excluding $i = j = k$. The commutators (39.16) imply that the non-vanishing chiral structure coefficients are

$$f_{[i\bar{j}][j\bar{k}][i\bar{k}]} = 1 \quad (39.17)$$

for any $i \neq j \neq k \neq i$. The chiral structure coefficients $f_{[i\bar{j}][j\bar{k}][i\bar{k}]}$ are *not* totally antisymmetric in their indices.

Suppose that $|\mathbf{m}\rangle$ is an eigenvector with i -charge m_i ,

$$\frac{1}{2}\gamma_{[i\bar{i}]}|\mathbf{m}\rangle = m_i|\mathbf{m}\rangle. \quad (39.18)$$

Suppose further that the operator $\frac{1}{2}\gamma_{[ij]}$ acting on the eigenvector $|\mathbf{m}\rangle$ yields another vector $|\mathbf{m}'\rangle$,

$$\frac{1}{2}\gamma_{[ij]}|\mathbf{m}\rangle = |\mathbf{m}'\rangle. \quad (39.19)$$

Then the i -charge of $|\mathbf{m}'\rangle$ is

$$\frac{1}{2}\gamma_{[i\bar{i}]}|\mathbf{m}'\rangle = \frac{1}{2}\gamma_{[i\bar{i}]} \frac{1}{2}\gamma_{[ij]}|\mathbf{m}\rangle = (\frac{1}{2}\gamma_{[ij]} \frac{1}{2}\gamma_{[i\bar{i}]} + [\frac{1}{2}\gamma_{[i\bar{i}], \frac{1}{2}\gamma_{[ij]}])|\mathbf{m}\rangle = (m_i + 1)|\mathbf{m}'\rangle, \quad (39.20)$$

that is, the raising operator $\frac{1}{2}\gamma_{[ij]}$ increases the i -charge of whatever it operates on by 1. The same raising operator $\frac{1}{2}\gamma_{[ij]}$ similarly increases the j -charge of whatever it operates on by 1. If i is changed to $\bar{i}$, the operator lowers i -charge by 1, and if j is changed to $\bar{j}$, the operator lowers j -charge by 1. The change of charge by increments of 1 explains the factors of $\frac{1}{2}$ in the choice of normalization of bivector generators. The eigenvectors $|\mathbf{m}\rangle$ and $|\mathbf{m}'\rangle$ must be orthogonal because they have different charges. The normalization of eigenvectors can be deduced from recurrence relations of the form

$$0 = \langle \mathbf{m} | [\frac{1}{2}\gamma_{[i\bar{j}]}, \frac{1}{2}\gamma_{[j\bar{i}]}] - \frac{1}{2}\gamma_{[i\bar{i}]} | \mathbf{m} \rangle = \langle \mathbf{m} | \frac{1}{2}\gamma_{[i\bar{j}]} | \frac{1}{2}\gamma_{[j\bar{i}]} \mathbf{m} \rangle - \langle \mathbf{m} | \frac{1}{2}\gamma_{[j\bar{i}]} | \frac{1}{2}\gamma_{[i\bar{j}]} \mathbf{m} \rangle - \langle \mathbf{m} | \frac{1}{2}\gamma_{[i\bar{i}]} | \mathbf{m} \rangle. \quad (39.21)$$

The vectors of a representation may be obtained by starting at one vector in the lattice of charges, and successively applying raising and lowering operators until all vectors of the representation are found. The Cartan-Weyl-Dynkin approach is to use a judiciously chosen minimal subset of n raising operators and n complementary lowering operators. The group $SU(n)$ is the subgroup of $Spin(2n)$ that preserves the sum of all charges, so only $n-1$ raising operators are needed to fill out a representation of $SU(n)$; equivalently n

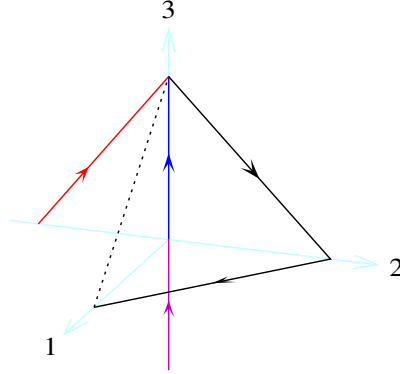


Figure 39.1 Cartan-Weyl-Dynkin raising operators for the groups $SU(3)$, $Spin(6)$, $Spin(7)$, and $Sp(3)$ with 3 conserved charges, from equation (39.23). The 3 charges point along orthogonal axes labelled 1, 2, and 3. All four groups share raising operators along the $1\bar{2}$ and $2\bar{3}$ directions, drawn in black. The dashed black line is added to bring out the fact that the black lines form edges of an equilateral triangle. The groups are distinguished by their final raising operator. $SU(3)$ conserves the total of the 3 charges, so only the 2 black raising operators are needed. The red, blue, and purple+blue lines are final raising operators for respectively $Spin(6)$, $Spin(7)$, and $Sp(3)$.

raising operators are needed by $SU(n+1)$. For the groups $SU(n+1)$, $Spin(2n+1)$, and $Spin(2n)$, the Cartan-Weyl-Dynkin basis of n raising operators is

$$\frac{1}{2}\gamma_{[1\bar{2}]}, \frac{1}{2}\gamma_{[2\bar{3}]}, \dots, \frac{1}{2}\gamma_{[n-1\bar{n}]}, \begin{cases} \frac{1}{2}\gamma_{[n\bar{n+1}]} & SU(n+1) \\ \frac{1}{2}\gamma_{[nN]} & Spin(2n+1) \\ \frac{1}{2}\gamma_{[n-1n]} & Spin(2n) \end{cases}. \quad (39.22)$$

The lowering operators are their complements; for example the complement of $\frac{1}{2}\gamma_{[1\bar{2}]}$ is $\frac{1}{2}\gamma_{[2\bar{1}]}$. The Cartan-Weyl-Dynkin raising operators (39.22) can be regarded as vectors α_i which shift the charge of a vector through an integrally-spaced lattice of n (or $n+1$ for $SU(n+1)$) charges,

$$\alpha_1, \dots, \alpha_n \equiv \{1, -1, \dots, 0, \dots\}, \{0, 1, -1, \dots, 0, \dots\}, \dots, \begin{cases} \{\dots, 0, \dots, 1, -1\} & SU(n+1) \\ \{\dots, 0, \dots, 1\} & Spin(2n+1) \\ \{\dots, 0, \dots, 2\} & Sp(n) \\ \{\dots, 0, \dots, 1, 1\} & Spin(2n) \end{cases}, \quad (39.23)$$

where $\dots, 0, \dots$ denotes a (possibly empty) sequence of zeros. The corresponding lowering vectors are $-\alpha_i$. The raising vectors for the case where there are 3 conserved charges are illustrated in Figure 39.1. Notice that the vectors α_i for $SU(n+1)$ are in an $(n+1)$ -dimensional space of charges, whereas the vectors for the other groups are in an n -dimensional space of charges. For $SU(n+1)$, the sum of all $n+1$ charges can be taken without loss of generality to be zero, since the total charge separates into a commuting $U(1)$ generator characterized by a charge $\{1, 1, \dots, 1\}$. $Spin(2)$ has 1 conserved charge, and zero raising operators; it is

isomorphic to $U(1)$. The lengths squared of the simple roots α_i (39.23) are

$$\alpha_i \cdot \alpha_i = c_i , \quad (39.24)$$

where the dot $\cdot$ denotes the Euclidean scalar product on the n -dimensional (or $(n+1)$ -dimensional for $SU(n+1)$) lattice of charges. The lengths squared c_i are

$$c_i = \begin{cases} 1 & \text{Spin}(2n+1) \text{ and } i = n , \\ 4 & \text{Sp}(n) \text{ and } i = n , \\ 2 & \text{otherwise} . \end{cases} \quad (39.25)$$

The representation is built up by applying vectors α_i successively. The charges m_j of a vector in the charge lattice may be expressed in terms of the net number $2\lambda_i/c_i$ of applications of α_i needed to reach the vector from the origin,

$$m_j = \sum_{i=1}^n \frac{2\lambda_i}{c_i} \alpha_{ij} , \quad j = 1, \dots, \begin{cases} n+1 & SU(n+1) \\ n & \text{otherwise} \end{cases} , \quad (39.26)$$

where α_{ij} are the components of α_i . The scaling factor $2/c_i$, which equals 1 for all but at most one of the vectors α_i in the set (39.23), is introduced to simplify the subsequent definition (39.27) of Dynkin coordinates. For any representation, there is always a lowest vector λ_i with the smallest possible charge, and a highest vector λ_i with the largest possible charge. The vector with the smallest charge is annihilated by all lowering operators; the vector with the highest charge is annihilated by all raising operators. The vectors of a representation can be obtained by starting with the highest (or lowest) vector and successively applying lowering and raising operators in all possible ways until the lowest (or highest) vector is reached.

Suppose that $\lambda_i = \{\lambda_1, \dots, \lambda_n\}$ is the highest vector of a representation. Only some choices of highest vector λ_i yield viable representations. Dynkin's key trick is to introduce Dynkin coordinates λ^i dual to the components λ_i (implicit sum over paired indices, one up and one down),

$$\lambda_i = g_{ij} \lambda^j , \quad \lambda^i = g^{ij} \lambda_j , \quad (39.27)$$

where the symmetric $n \times n$ Dynkin metric g_{ij} and its inverse g^{ij} are defined by

$$g_{ij} \equiv \frac{c_i}{2} (\alpha_i \cdot \alpha_j)^{-1} \frac{c_j}{2} , \quad g^{ij} \equiv \frac{2}{c_i} \alpha_i \cdot \alpha_j \frac{2}{c_j} \quad (\text{no sum over } i \text{ or } j) , \quad (39.28)$$

with the dot $\cdot$ again denoting the Euclidean scalar product in the charge lattice. The key result of Dynkin theory is that the Dynkin coordinates λ^i of every vector of a representation are integers. Every sequence of non-negative integers $\lambda^i = \{\lambda^1, \dots, \lambda^n\}$ defines a highest vector that gives rise to a distinct representation, and every representation is characterized by such a sequence of non-negative integers. A single step α_i changes the Dynkin coordinates λ^j of a vector by

$$C_{ij} \equiv (\Delta \lambda^j)_i = g^{ji} \frac{c_i}{2} = \alpha_i \cdot \alpha_j \frac{2}{c_j} \quad (\text{no sum over } i \text{ or } j) , \quad (39.29)$$

which is called the Cartan-Weyl matrix C_{ij} , or simply the Cartan matrix. The $n \times n$ Cartan matrix has all

integer entries, consistent with the fact that the Dynkin coordinates λ^i of any vector of a representation are always integers. The charge of a vector with Dynkin coordinates λ^i is, from equations (39.26) and (39.27),

$$m_k = \sum_{i,j=1}^n \lambda^i g_{ij} \frac{2\alpha_{jk}}{c_j}, \quad k = 1, \dots, \begin{cases} n+1 & \text{SU}(n+1) \\ n & \text{otherwise} \end{cases}. \quad (39.30)$$

The quadratic Casimir invariant, equation (39.10), of a representation whose highest vector has Dynkin coordinates λ^i is

$$C_2(\lambda^i) = \frac{1}{2} \sum_{i,j=1}^n \lambda^i g_{ij} (\lambda^j + 2^j), \quad (39.31)$$

where 2^j denotes the vector $\{2, 2, \dots, 2\}$. Equation (39.31) holds for a simple Lie group; the generalization to a product of groups is obtained in Exercise 39.4, equation (39.60).

The dimension of a representation with highest vector λ^i is given by Weyl's formula (39.36) below. The formula depends not only on the Cartan-Weyl-Dynkin basis set of n raising operators (39.22), but on the full set of raising operators. The raising and lowering operators of a group divide into two equal sets, raising operators, and their complements, lowering operators. For $\text{Spin}(N)$, the raising operators from the list (39.15) are $\frac{1}{2}\gamma_{[ij]}$ with $i < j$, and $\frac{1}{2}\gamma_{[ij]}$ also with $i < j$ without loss of generality. For N odd, the raising operators include $\frac{1}{2}\gamma_{[iN]}$. The number n_{raise} of raising operators is

$$n_{\text{raise}} = \begin{cases} \frac{1}{2}n(n+1) & \text{SU}(n+1) \\ n^2 & \text{Spin}(2n+1) \\ n^2 & \text{Sp}(n) \\ n(n-1) & \text{Spin}(2n) \end{cases}. \quad (39.32)$$

The raising operators are characterized by n_{raise} charge vectors α_i whose components in the Euclidean space of charges are

$$\alpha_i = \begin{cases} \{\dots 0\dots, 1, \dots 0\dots, -1, \dots 0\dots\} & \frac{1}{2}n(n+1) \text{ vectors} & \text{SU}(n+1), \\ \{\dots 0\dots, 1, \dots 0\dots, -1, \dots 0\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \text{Spin}(2n+1), \text{Sp}(n), \text{Spin}(2n), \\ \{\dots 0\dots, 1, \dots 0\dots, 1, \dots 0\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \text{Spin}(2n+1), \text{Sp}(n), \text{Spin}(2n), \\ \{\dots 0\dots, 1, \dots 0\dots\} & n \text{ vectors} & \text{Spin}(2n+1), \\ \{\dots 0\dots, 2, \dots 0\dots\} & n \text{ vectors} & \text{Sp}(n). \end{cases} \quad (39.33)$$

Express the charges α_{ij} of the n_{raise} raising operators α_i as linear combinations of the charges of the subset of n Cartan-Weyl-Dynkin basis operators α_{kj} given by equation (39.23),

$$\alpha_{ij} = \sum_{k=1}^n \frac{2a_{ik}}{c_k} \alpha_{kj}, \quad (39.34)$$

which defines the $n_{\text{raise}} \times n$ (or $n_{\text{raise}} \times (n+1)$ for $\text{SU}(n+1)$) matrix a_{ik} . The n_{raise} rows α_i of the matrix a_{ik}

in equation (39.34) are

$$\mathbf{a}_i = \left\{ \begin{array}{lll} \{\dots 0\dots, 1\dots, 0\dots\} & \frac{1}{2}n(n+1) \text{ vectors} & \text{SU}(n+1) , \\ \{\dots 0\dots, 1\dots, 0\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \\ \{\dots 0\dots, 1\dots, \dots 2\dots, 1\} & \frac{1}{2}n(n-1) \text{ vectors} & \left. \vphantom{\begin{array}{l} \{\dots 0\dots, 1\dots, 0\dots\} \\ \{\dots 0\dots, 1\dots, 0\dots\} \end{array}} \right\} \text{Spin}(2n+1) , \\ \{\dots 0\dots, \dots 1\dots, \frac{1}{2}\} & n \text{ vectors} & \\ \\ \{\dots 0\dots, 1\dots, 0\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \\ \{\dots 0\dots, 1\dots, 2\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \left. \vphantom{\begin{array}{l} \{\dots 0\dots, 1\dots, 0\dots\} \\ \{\dots 0\dots, 1\dots, 0\dots\} \end{array}} \right\} \text{Sp}(n) , \\ \{\dots 0\dots, 2\dots\} & n \text{ vectors} & \\ \\ \{\dots 0\dots, 1\dots, 0\dots\} & \frac{1}{2}n(n-1) \text{ vectors} & \\ \{\dots 0\dots, 1\dots, 2\dots, 1, 1\} & \frac{1}{2}(n-2)(n-3) \text{ vectors} & \left. \vphantom{\begin{array}{l} \{\dots 0\dots, 1\dots, 0\dots\} \\ \{\dots 0\dots, 1\dots, 0\dots\} \end{array}} \right\} \text{Spin}(2n) , \\ \{\dots 0\dots, \dots 1\dots, 0, 1\} & n-1 \text{ vectors} & \\ \{\dots 0\dots, 1\dots, 1, 1\} & n-2 \text{ vectors} & \end{array} \right. \quad (39.35)$$

where $\dots i\dots$ (dots on both sides) denotes a possibly empty sequence of i 's, while $i\dots$ (dots only on the right) denote a non-empty sequence of i 's. Weyl's formula for the dimension of a representation with highest vector λ^j is

$$\dim(\lambda^j) = \prod_{i=1}^{n_{\text{raise}}} \frac{\sum_{j=1}^n a_{ij}(\lambda^j + 1^j)}{\sum_{j=1}^n a_{ij}}, \quad (39.36)$$

where 1^j denotes the vector $\{1, \dots, 1\}$.

Two powerful features of Cartan-Weyl-Dynkin theory are that (1) a Lie group can be visualized in terms of its diagram, Figure 39.1 for example, and (2) a Lie group is characterized by its Dynkin metric (39.28). Two groups are isomorphic if and only if they have the same Dynkin metric (after a possible permutation of Dynkin coordinates). An example is the isomorphism between $\text{SU}(4)$ and $\text{Spin}(6)$, Exercise 39.3.

The Dynkin metric g_{ij} of a Lie group that is a direct product of Lie groups with metrics $g_{ij}(1)$ and $g_{ij}(2)$ is the block diagonal metric

$$g_{ij} = \begin{pmatrix} g_{ij}(1) & 0 \\ 0 & g_{ij}(2) \end{pmatrix}. \quad (39.37)$$

If the Dynkin metric of a Lie group is block diagonal, then the group is a direct product. An example is the isomorphism between $\text{Spin}(4)$ and $\text{SU}(2) \times \text{SU}(2)$, Exercise 39.2.

The question of whether a Lie group is a direct product of groups can be determined by inspection from its diagram. If the Cartan-Weyl-Dynkin raising operators of the group split into two sets that are orthogonal to each other (which is the same as the condition that the Dynkin metric is block diagonal), then the group is a direct product. For example, the diagram of $\text{Spin}(4)$ consists of the red line and the one black line orthogonal to it in Figure 39.1. Therefore $\text{Spin}(4)$ is isomorphic to the product of two groups each of whose diagrams consist of a single line, namely $\text{SU}(2)$.

Table 39.1 lists all spinor and multivector representations of $\text{SU}(N)$ and $\text{Spin}(N)$, along with the Dynkin coordinates λ^i of the highest vector (which defines the representation), the dimension, quadratic Casimir in-

Table 39.1: Spinor and multivector representations of Lie groups

Group	Dynkin coordinate $\{\lambda^1, \dots, \lambda^n\}$	Spinor?	Grade	Dimension	Casimir invariant C_2	Dynkin index S_2
Any	$\{0\dots\}$	$\times$	0	1	0	0
$SU(n+1)$	$\{\dots 0\dots, \overset{p}{1}, \dots 0\dots\}$	$\checkmark$	$p \leq n$	$\binom{n+1}{p}$	$\frac{p(n+1-p)(n+2)}{2(n+1)}$	$\frac{1}{2} \binom{n-1}{p-1}$
	$\{\dots 0\dots, \overset{p}{1}, \dots 0\dots, \overset{n+1-p}{1}, \dots 0\dots\}$	$\times$	$2p < n+1$	$\frac{n+2-2p}{n+2} \binom{n+2}{p}^2$	$p(n+2-p)$	$\frac{(n+2-2p)}{n} \binom{n+1}{p-1} \binom{n+1}{p}$
	$\{\dots 0\dots, \overset{p}{2}, \dots 0\dots\}$	$\times$	$2p = n+1$	$\frac{1}{n+2} \binom{n+2}{p}^2$	$p(n+2-p)$	$\frac{1}{n} \binom{n+1}{p-1} \binom{n+1}{p}$
$Spin(2n+1)$	$\{\dots 0\dots, 1\}$	$\checkmark$	—	2^n	$\frac{1}{8} n(2n+1)$	2^{n-3}
	$\{\dots 0\dots, \overset{p}{1}, \dots 0\dots\}$	$\times$	$p \leq n-1$	$\binom{2n+1}{p}$	$\frac{1}{2} p(2n+1-p)$	$\binom{2(n-1)}{p-1}$
	$\{\dots 0\dots, 2\}$	$\times$	n	$\binom{2n+1}{n}$	$\frac{1}{2} n(n+1)$	$\binom{2(n-1)}{n-1}$
$Sp(n)$	$\{\dots 0\dots, \overset{p}{1}, \dots 0\dots\}$	$\times$	—	$\frac{2(n+1-p)}{2n+2-p} \binom{2n+1}{p}$	$\frac{1}{2} p(2n+2-p)$	$\frac{n+1-p}{n} \binom{2n}{p-1}$
	$\{2, \dots 0\dots\}$	$\times$	—	$n(2n+1)$	$2(n+1)$	$2(n+1)$
$Spin(2n)$	$\{\dots 0\dots, 1, 0\}$ or $\{\dots 0\dots, 0, 1\}$	$\checkmark$	—	2^{n-1}	$\frac{1}{8} n(2n-1)$	2^{n-4}
	$\{\dots 0\dots, \overset{p}{1}, \dots 0\dots\}$	$\times$	$p \leq n-2$	$\binom{2n}{p}$	$\frac{1}{2} p(2n-p)$	$\binom{2(n-1)}{p-1}$
	$\{\dots 0\dots, 1, 1\}$	$\times$	$n-1$	$\binom{2n}{n-1}$	$\frac{1}{2} (n-1)(n+1)$	$\binom{2(n-1)}{n-2}$
	$\{\dots 0\dots, 2, 0\}$ or $\{\dots 0\dots, 0, 2\}$	$\times$	n	$\frac{1}{2} \binom{2n}{n}$	$\frac{1}{2} n^2$	$\frac{1}{2} \binom{2(n-1)}{n-1}$

variant C_2 , and Dynkin index S_2 of the representation. Other representations, not listed, are representations of irreducible components of tensor products of spinors and/or multivectors. Beware that the Casimir invariant and Dynkin index are proportional to charge squared, and their numerical values depend on the units of charge adopted, which may vary between authors. The Casimir invariant and Dynkin index in Table 39.1 are in charge units such that the separation of adjacent charges on the charge lattice is unity, Figure 39.1.

$Spin(2n+1)$ has one spinor representation, while $Spin(2n)$ has two, which are conjugates of each other. $Spin(N)$ has a representation for multivectors of grade p , and the same representation holds for their pseudo partners, multivectors of grade $N-p$; except that for $Spin(2n)$ and grade $p = n$ there are two representations, each containing half of the grade- n multivectors, one representation being the pseudoscalar times the other. The adjoint representation is the bivector representation, grade $p = 2$.

$SU(N)$ is the subgroup of $Spin(2N)$ (or of $Spin(2N+1)$) that preserves the total charge (total spin weight, or number of up bits, in the language of the supergeometric algebra, §40.2). The spinor representations of $SU(N)$ are characterized by their spinor grade, the number of up bits of the spinor. For example, $SU(5)$ has spinor representations of spinor grades 0 to 5, as listed in the $Spin(10)$ chart (48.15), with dimensions 1,

5, 10, 10, 5, 1. $SU(N)$ has a multivector representation for each even grade $2p$, consisting of the subset of $Spin(2N)$ multivectors of grade $2p$ that have zero charge (zero spin weight). The adjoint representation is the bivector representation, grade $2p = 2$.

$Sp(n)$ does not have spinor or multivector representations, but Table 39.1 includes data on its fundamental representations, those whose Dynkin coordinate is $\{\dots 0\dots, 1, \dots 0\dots\}$, and its adjoint representation, whose Dynkin coordinate is $\{2, \dots 0\dots\}$.

Not included in Table 39.1 is the simplest of all Lie groups, the group $U(1)$ of dimension 1. Whereas generators of groups of dimension 2 or more are normalized naturally by setting the separation between adjacent charges on the charge lattice to 1 (Figure 39.1), the group $U(1)$, having only 1 charge, has no such natural normalization. Yet the charges of the $U(1)_Y$ hypercharge and $U(1)_{em}$ electromagnetic groups do come in discrete increments, leading to the commonly adopted empirical normalizations of hypercharge and electric charge listed in the Table 48.1 of SM charges. A “natural” normalization of a $U(1)$ charge may emerge if it is embedded in a larger unifying group such as $Spin(10)$. Regardless of the choice of units of charge, the Casimir invariant C_2 and Dynkin index S_2 of $U(1)$ are dimensionful quantities equal to the square of the $U(1)$ charge, equations (39.10) and (39.11).

Exercise 39.1 Representations of $Spin(3)$. The group $Spin(3)$ of rotations in 3 spatial dimensions is the simplest simple Lie group with non-vanishing commutators.

1. Use the Cartan-Weyl-Dynkin approach to find all representations of $Spin(3)$.
2. Given that $Spin(3)$ is isomorphic to $SU(2)$, is there any difference between their representations?

Solution.

1. The group $Spin(3)$ has 3 generators, which in an orthonormal basis are $\frac{1}{2}\gamma_{[ab]}$ with indices a and b drawn from 1, 2, 3. The orthonormal generators are commonly denoted by angular momentum operators L_a with $a = 1, 2, 3$ (or $a = x, y, z$),

$$iL_1 \equiv \frac{1}{2}\gamma_{[23]} , \quad iL_2 \equiv \frac{1}{2}\gamma_{[31]} , \quad iL_3 \equiv \frac{1}{2}\gamma_{[12]} . \quad (39.38)$$

The commutators of the orthonormal generators L_a are

$$[L_a, L_b] = i\varepsilon_{abc}L_c , \quad (39.39)$$

with ε_{abc} the totally antisymmetric symbol. In a chiral basis, the generators are the diagonal generator $\frac{1}{2}\gamma_{[1\bar{1}]}$ and the raising and lower generators $\frac{1}{2}\gamma_{[13]}$ and $\frac{1}{2}\gamma_{[\bar{1}3]}$. The chiral generators are

$$L_3 = \frac{1}{2}\gamma_{[1\bar{1}]} , \quad L_+ \equiv \frac{1}{\sqrt{2}}(L_1 + iL_2) = -\frac{1}{2}\gamma_{[13]} , \quad L_- \equiv \frac{1}{\sqrt{2}}(L_1 - iL_2) = \frac{1}{2}\gamma_{[\bar{1}3]} . \quad (39.40)$$

The commutation rules of the chiral generators are

$$[L_+, L_-] = L_3 , \quad [L_3, L_{\pm}] = \pm L_{\pm} , \quad (39.41)$$

in agreement with equations (35.103). $Spin(3)$ has a single conserved charge, the eigenvalue m of L_3 , the component of angular momentum about the 3-axis. A representation of $Spin(3)$ is labelled by the

single Dynkin integer coordinate of its highest vector, $\{\lambda^1\} = \{2\ell\}$. The Casimir invariant (39.31) is

$$C_2(\{2\ell\}) = \frac{1}{2}\ell(\ell+1) , \quad (39.42)$$

half the total angular momentum squared of the representation. The dimension of a representation of Dynkin coordinate $\{2\ell\}$ is

$$\dim(\{2\ell\}) = 2\ell + 1 . \quad (39.43)$$

The Dynkin index (39.11) of the representation is

$$S_2(\{2\ell\}) = \frac{\dim(\{2\ell\})}{\dim(\text{Spin}(3))} C_2(\{2\ell\}) = \frac{(2\ell+1)\ell(\ell+1)}{6} . \quad (39.44)$$

The smallest non-trivial representation is the spinor representation $\ell = \frac{1}{2}$, which has dimension $2\ell+1 = 2$. The representation of the angular momentum operators in that case are $\frac{1}{2}$ the Pauli matrices,

$$(L_3)_{ij} = \frac{1}{2}\sigma_3 , \quad (L_{\pm})_{ij} = \frac{1}{2}\sigma_{\pm} . \quad (39.45)$$

2. Spin(3) is isomorphic not only to SU(2) but also to Sp(1). The isomorphism is evident from the fact that the diagrams for all three groups are the same, a line joining two points, Figure 39.1. So yes, their representations are the same. However, it is necessary to worry about units. The Dynkin metric is a 1×1 matrix, but with different normalizations,

$$g_{ij} = \begin{pmatrix} 1 \end{pmatrix} \times \begin{cases} \frac{1}{2} & \text{SU}(2) \\ \frac{1}{4} & \text{Spin}(3) \\ 1 & \text{Sp}(1) \end{cases} . \quad (39.46)$$

The quadratic Casimir invariants (39.31) differ correspondingly,

$$C_2(\{2\ell\}) = \ell(\ell+1) \times \begin{cases} 1 & \text{SU}(2) \\ \frac{1}{2} & \text{Spin}(3) \\ 2 & \text{Sp}(1) \end{cases} . \quad (39.47)$$

The reason for the difference is that the Dynkin metric and the quadratic Casimir invariant are proportional to the square of the separation between adjacent charges, which, as can be seen in Figure 39.1, is 2 (black line) for SU(2), 1 (blue line) for Spin(3), and 4 (purple+blue line) for Sp(1). The separation of charges is a matter of units. In rotations in 3 dimensions, there is a single conserved charge, the angular momentum L_3 about the 3-axis. It is natural to adopt units such that a change by one unit in the charge lattice (generated by the raising operator $\frac{1}{2}\gamma_{[13]}$, the vertical blue line in Figure 39.1), as in equation (39.20), corresponds to one unit of angular momentum; indeed this is so if angular momentum is measured in natural units $\hbar$. In weak interactions on the other hand, there are two weak charges, y and z , and it is natural (conventional) to choose the charge separation such that y and z change by 1 in a weak interaction (the 45° black line in Figure 39.1). In those units, the correct normalization of the Casimir invariant (39.47) for weak interactions is the SU(2) normalization.

Exercise 39.2 Spin(4) is isomorphic to $\text{SU}(2) \times \text{SU}(2)$. Prove that the groups $\text{Spin}(4)$ and $\text{SU}(2) \times \text{SU}(2)$ are isomorphic. How are the Dynkin indices and Casimir invariants of $\text{Spin}(4)$ related to those of $\text{SU}(2)$?

Solution. The Dynkin metric of $\text{SU}(2)$ is the 1×1 matrix

$$g_{ij} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} . \quad (39.48)$$

The Dynkin metric of $\text{Spin}(4)$ is the 2×2 matrix

$$g_{ij} = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} , \quad (39.49)$$

which is the block diagonal composition of two $\text{SU}(2)$ Dynkin metrics. Therefore $\text{Spin}(4)$ is isomorphic to $\text{SU}(2) \times \text{SU}(2)$.

The representations of $\text{Spin}(4)$ are labelled by an integral Dynkin vector $\{\lambda_1, \lambda_2\} = \{2\ell_1, 2\ell_2\}$ with 2 coordinates. The Dynkin index and Casimir invariant of a representation of $\text{Spin}(4)$ are related to those of the $\text{SU}(2)$ representations by, equation (39.59),

$$\begin{aligned} S_2(\text{Spin}(4), \{2\ell_1, 2\ell_2\}) &= \frac{S_2(\text{SU}(2), \{2\ell_1\}) + S_2(\text{SU}(2), \{2\ell_2\})}{2} \\ C_2(\text{Spin}(4), \{2\ell_1, 2\ell_2\}) &= \frac{(2\ell_1 + 1)C_2(\text{SU}(2), \{2\ell_1\}) + (2\ell_2 + 1)C_2(\text{SU}(2), \{2\ell_2\})}{2(\ell_1 + \ell_2 + 1)} . \end{aligned} \quad (39.50)$$

Exercise 39.3 $\text{SU}(4)$ is isomorphic to $\text{Spin}(6)$. Prove that the groups $\text{SU}(4)$ and $\text{Spin}(6)$ are isomorphic. How are the Dynkin index and Casimir invariant of a representation of $\text{SU}(4)$ and $\text{Spin}(6)$ related?

Solution. The Dynkin metrics of $\text{SU}(4)$ and $\text{Spin}(6)$ are

$$g_{ij} = \begin{cases} \frac{1}{4} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix} & \text{SU}(4) \\ \frac{1}{4} \begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} & \text{Spin}(6) \end{cases} , \quad (39.51)$$

which are related by permuting the first two rows and columns, $1 \leftrightarrow 2$. Therefore $\text{SU}(4)$ and $\text{Spin}(6)$ are isomorphic.

The Dynkin index S_2 and Casimir invariant C_2 of a representation of $\text{SU}(4)$ and $\text{Spin}(6)$ are related by swapping the first two Dynkin coordinates,

$$\begin{aligned} S_2(\text{SU}(4), \{\lambda_1, \lambda_2, \lambda_3\}) &= S_2(\text{Spin}(6), \{\lambda_2, \lambda_1, \lambda_3\}) \\ C_2(\text{SU}(4), \{\lambda_1, \lambda_2, \lambda_3\}) &= C_2(\text{Spin}(6), \{\lambda_2, \lambda_1, \lambda_3\}) . \end{aligned} \quad (39.52)$$

Exercise 39.4 Representations of products of simple Lie groups. If a Lie group is a direct product $G \times H$ of Lie groups, then the corresponding Lie algebra is a direct sum of the Lie algebras of G and H . Describe the representations of the product $G \times H$ of simple Lie groups in terms of the representations of

G and H . How are the Dynkin index and Casimir invariant of a representation of $G \times H$ related to those of the representations of G and H ?

Solution. The dimension of the product Lie group $G \times H$ is the sum of the dimensions of the groups,

$$\dim(G \times H) = \dim(G) + \dim(H) . \quad (39.53)$$

The representations of $G \times H$ are direct sums of individual representations of G and H ,

$$\text{rep}_{G \times H} = \text{rep}_G \oplus \text{rep}_H . \quad (39.54)$$

The dimension of a direct sum representation is the sum of the dimensions of the individual representations,

$$\dim(\text{rep}_{G \times H}) = \dim(\text{rep}_G \oplus \text{rep}_H) = \dim(\text{rep}_G) + \dim(\text{rep}_H) . \quad (39.55)$$

The fact that $G \times H$ is a direct product implies that the vectors of G are orthogonal to vectors of H in the direct sum Lie algebra of $G \times H$ (it follows from the Baker-Campbell-Hausdorff formula (13.59) that elements of the Lie algebra of G all commute with elements of the Lie algebra of H). With respect to a basis orthogonal between G and H , the elements of the Lie group and Lie algebra of a direct sum representation are block diagonal $\dim(\text{rep}_{G \times H}) \times \dim(\text{rep}_{G \times H})$ matrices, which act on $\dim(\text{rep}_{G \times H})$ -dimensional vectors, as in

$$\begin{pmatrix} g & 0 \\ 0 & h \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} . \quad (39.56)$$

In an orthogonal basis, the $\dim(\text{rep}_{G \times H})$ generators of a direct sum representation comprise

$$\dim(\text{rep}_G) \text{ generators } \begin{pmatrix} g & 0 \\ 0 & 0 \end{pmatrix} \text{ together with } \dim(\text{rep}_H) \text{ generators } \begin{pmatrix} 0 & 0 \\ 0 & h \end{pmatrix} . \quad (39.57)$$

The Dynkin index $S_2(\text{rep})$ of an irreducible representation of a group was defined by equation (39.9) in terms of a trace of products of orthonormal generators. The corresponding trace of products of orthonormal generators of a direct sum representation is

$$\text{Tr}(S_A S_B) = \begin{pmatrix} S_2(\text{rep}_G) 1_{\dim(G)} & 0 \\ 0 & S_2(\text{rep}_H) 1_{\dim(H)} \end{pmatrix} , \quad (39.58)$$

where 1_D denotes the $D \times D$ unit matrix. The Casimir invariant $C_2(\text{rep})$ of an irreducible representation of a simple group was defined by equation (39.10) in terms of a sum of squares of orthonormal generators. The corresponding sum of squares of orthonormal generators of a product group $G \times H$ is

$$\sum_{A=1}^{\dim(G \times H)} S_A^2 = \begin{pmatrix} C_2(\text{rep}_G) 1_{\dim(\text{rep}_G)} & 0 \\ 0 & C_2(\text{rep}_H) 1_{\dim(\text{rep}_H)} \end{pmatrix} . \quad (39.59)$$

The Dynkin index S_2 and Casimir invariant C_2 of a direct sum representation $\text{rep}_{G \times H}$ may be defined by

$$\begin{aligned} \dim(G \times H) S_2(\text{rep}_{G \times H}) &\equiv \dim(G) S_2(\text{rep}_G) + \dim(H) S_2(\text{rep}_H) \\ &= \dim(\text{rep}_{G \times H}) C_2(\text{rep}_{G \times H}) \equiv \dim(\text{rep}_G) C_2(\text{rep}_G) + \dim(\text{rep}_H) C_2(\text{rep}_H) . \end{aligned} \quad (39.60)$$

So defined, the relation (39.11) between Dynkin index and Casimir invariant continues to hold for a direct sum representation.

39.3 Self-dual compactification tori

A Lie group of rank n defines an integral lattice in an n -dimensional Euclidean geometry, as illustrated in Figure 39.1. The geometry can be made into a torus by treating each of the n simple roots α_i of the Lie algebra as a periodic vector. Such a torus is called the **maximal torus** of a Lie group, the term maximal referring to the condition that the dimension of the torus equals the rank of the group.

In string theory, such tori provide one way of generating a gauge group from a compact geometry, §???. In string theory, the condition that the theory be invariant under exchange $\tau \leftrightarrow \sigma$ of the string's two coordinates imposes that the lattice remain integral under exchange $\{\alpha_i\} \leftrightarrow \{\hat{\alpha}_i\}$ of the simple roots with their duals defined by equations (39.61). Such a torus is called **self-dual**. In string theory, a compactification torus must be self-dual.

A set of vectors $\hat{\alpha}_i$ dual to the vectors α_i may be defined by

$$\alpha_i = \sum_{j=1}^n (C^{-1})_{ij} \hat{\alpha}_j, \quad \hat{\alpha}_i = \sum_{j=1}^n C_{ij} \alpha_j, \quad (39.61)$$

where C_{ij} is the Cartan matrix (39.29). The vectors α_i and their duals $\hat{\alpha}_i$ are orthogonal,

$$\alpha_i \cdot \hat{\alpha}_j = \delta_{ij} \frac{c_j}{2}, \quad (39.62)$$

with c_j from equation (39.25). The metric defined by equation (39.62) is Euclidean if and only if $c_j = 2$ for all j , the same condition as for the Cartan matrix to be symmetric. The metric is Euclidean for $SU(n+1)$ and $Spin(2n)$, but not for $Spin(2n+1)$ and $Sp(n)$. Their scalar products with themselves are

$$\alpha_i \cdot \alpha_j = C_{ij} \frac{c_j}{2}, \quad \hat{\alpha}_i \cdot \hat{\alpha}_j = (C^{-1})_{ij} \frac{c_j}{2}. \quad (39.63)$$

The Cartan matrix C_{ij} defined by equation (39.29) is integral, and the squared lengths $\alpha_i \cdot \alpha_i$ of the simple roots are all integral, equation (39.25). The squared lengths $\hat{\alpha}_i \cdot \hat{\alpha}_i$ of the dual roots are integral if and only if the inverse Cartan matrix $(C^{-1})_{ij}$ is integral. The inverse Cartan matrix is integral if and only if the Cartan matrix has unit determinant.

In string theory, a further condition arises from the requirement that the compactification torus of the group yield a full complement of massless gauge bosons. When bosonic strings are quantized, zero-mass gauge bosons arise from modes of momentum squared equal to 2 (Gross et al., 1985). For a gauge group of rank n and dimension d , there are n Kaluza-Klein modes, and a full complement of massless gauge bosons requires that there be exactly $d-n$ nearest-neighbour pairs of separation squared 2. There can be no nearest-neighbour pairs with separations squared less than 2, which would yield problematic tachyonic gauge bosons; and if there were fewer than $d-n$ nearest-neighbour pairs of separation squared 2, then the full complement

of massless gauge bosons of the gauge group would not be realized. A compactification torus with exactly $d-n$ nearest-neighbour pairs of separation squared 2 is said to be **simply-laced**.

The remainder of this section 39.3.2 shows that $SU(n+1)$, $Spin(4k+1)$, $Sp(2k)$, and $Spin(8k)$ all admit self-dual compactification tori. Section 39.4 shows that only $SU(n+1)$ and $Spin(8k)$ admit compactification tori that are simply-laced; $Spin(4k+1)$ and $Sp(2k)$ do not admit simply-laced tori, and are not viable as gauge groups in string theory.

39.3.1 Cartan matrices

The Cartan matrix C_{ij} is defined by equation (39.29). The diagonal elements of the standard Cartan matrices are all equal to 2. For $SU(n+1)$, the standard Cartan matrix is

$$SU(n+1) : C_{ij} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 & 0 \\ & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 2 \end{pmatrix}. \quad (39.64)$$

For $Spin(2n+1)$, $Sp(n)$, and $Spin(2n)$, the standard Cartan matrix is the same as that (39.64) for $SU(n+1)$, but with the final rows replaced by:

$$Spin(2n+1) : C_{ij} = \begin{pmatrix} & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 & -2 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 2 \end{pmatrix}, \quad (39.65a)$$

$$Sp(n) : C_{ij} = \begin{pmatrix} & & & \vdots & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -2 & 2 \end{pmatrix}, \quad (39.65b)$$

$$Spin(2n) : C_{ij} = \begin{pmatrix} & & & \vdots & & & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & 2 \end{pmatrix}. \quad (39.65c)$$

The Cartan matrices of $SU(n+1)$ and $Spin(2n)$ are symmetric, while the Cartan matrices of $Spin(2n+1)$ and $Sp(n)$ are not symmetric, and are transposes of each other. The determinants of the standard Cartan matrices are

$$|C_{ij}| = \begin{cases} n+1 & SU(n+1), \\ 2 & Spin(2n+1) \text{ or } Sp(n), \\ 4 & Spin(2n). \end{cases} \quad (39.66)$$

39.3.2 Self-dual root systems

None of the standard Cartan matrices of the non-exceptional Lie groups is self-dual, since none of them have unit determinant. The only standard Cartan matrices that are self-dual, having unit determinant, are those of the exceptional groups G_2 , E_4 , and E_8 . The Cartan matrices of the exceptional groups are symmetric for E_n , $n = 6, 7, 8$, but not symmetric for G_2 or E_4 . The only Lie group whose standard Cartan matrix is self-dual and symmetric is E_8 .

The unimodular lattices listed by Conway and Sloane (1982) include only those with a symmetric Cartan matrix, hence a Euclidean inner product $\alpha_i \cdot \hat{\alpha}_j$ between the simple roots and their duals. Conway and Sloane (1982) were interested in the lattices of quadratic forms, whose Cartan matrix is indeed symmetric. However, the restriction to a symmetric Cartan matrix is unwarranted in string theory.

Self-dual Cartan matrices for the non-exceptional Lie groups can be obtained by replacing the first standard simple root α_1 with a spinor root, in much the same way that Gross et al. (1985) were able to construct a self-dual Spin(32) torus for heterotic theory by substituting one of its roots with a spinor root. Spinor roots σ_p constitute a subset of the roots with Dynkin coordinate $\{..0.., \overset{p}{1}, ..0..\}$ (1 in the p th place, 0 elsewhere). Equivalently, spinor roots constitute a subset $\sigma_p = \hat{\alpha}_p$ of the dual roots of the standard set of simple roots α_i .

For $SU(n+1)$, the n spinor roots σ_p , $p = 1, \dots, n$ are given by the $n \times (n+1)$ matrix

$$SU(n+1) : \quad \sigma_{pj} = \frac{1}{n+1} \begin{pmatrix} n & -1 & -1 & \dots & -1 & -1 & -1 \\ n-1 & n-1 & -2 & \dots & -2 & -2 & -2 \\ n-2 & n-2 & n-2 & \dots & -3 & -3 & -3 \\ & & & \ddots & & & \\ 2 & 2 & 2 & \dots & 2 & -(n-1) & -(n-1) \\ 1 & 1 & 1 & \dots & 1 & 1 & -n \end{pmatrix}. \quad (39.67)$$

For each of Spin(2n+1) and Sp(n) there is just one spinor root σ_n , given by the $1 \times n$ matrices

$$\text{Spin}(2n+1) : \quad \sigma_{nj} = \left(\frac{1}{2} \quad \dots \quad \frac{1}{2} \right), \quad (39.68a)$$

$$\text{Sp}(n) : \quad \sigma_{nj} = \left(1 \quad \dots \quad 1 \right). \quad (39.68b)$$

For Spin(2n), there are two spinor roots σ_p , $p = n-1, n$, given by the $2 \times n$ matrix

$$\text{Spin}(2n) : \quad \sigma_{pj} = \begin{pmatrix} \frac{1}{2} & \dots & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \dots & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \quad (39.69)$$

The lengths squared of the spinor roots σ_p are

$$\sigma_p \cdot \sigma_p = s_p, \quad (39.70)$$

with

$$s_p = \begin{cases} \frac{p(n+1-p)}{n+1} & \text{SU}(n+1) , \\ \frac{n}{4} & \text{Spin}(2n+1) \text{ or } \text{Spin}(2n) , \\ n & \text{Sp}(n) . \end{cases} \quad (39.71)$$

The Cartan matrix associated with the root system $\{\sigma_p, \alpha_2, \dots, \alpha_n\}$, in which the spinor root σ_p replaces the first simple root α_1 , is given by equation (39.29) with s_p replacing c_1 . The resulting Cartan matrices are the same as the standard Cartan matrices (39.64) and (39.65), but with the first row and column replaced by:

$$\begin{aligned} \text{SU}(n+1) : \quad C_{1j} &= \begin{pmatrix} \frac{p}{n+1-p} & \dots 0 \dots & \overset{p}{1} & \dots 0 \dots \end{pmatrix} , \\ C_{i1} &= \begin{pmatrix} \frac{p}{n+1-p} & \dots 0 \dots & \frac{\overset{p}{n+1}}{(n+1-p)^2} & \dots 0 \dots \end{pmatrix}^\top , \end{aligned} \quad (39.72a)$$

$$\begin{aligned} \text{Spin}(2n+1) \text{ or } \text{Sp}(n) : \quad C_{1j} &= \begin{pmatrix} \frac{n}{2} & \dots 0 \dots & 1 \end{pmatrix} , \\ C_{i1} &= \begin{pmatrix} \frac{n}{2} & \dots 0 \dots & 1 \end{pmatrix}^\top , \end{aligned} \quad (39.72b)$$

$$\begin{aligned} \text{Spin}(2n) : \quad C_{1j} &= \begin{pmatrix} \frac{n}{4} & \dots 0 \dots & \overset{p}{1} & \dots 0 \dots \end{pmatrix} , \\ C_{i1} &= \begin{pmatrix} \frac{n}{4} & \dots 0 \dots & \overset{p}{1} & \dots 0 \dots \end{pmatrix}^\top , \end{aligned} \quad (39.72c)$$

where $\dots 0 \dots$ denotes a (possibly empty) sequence of zeroes, and the p overscript signifies the p th column or row.

All of the Cartan matrices (39.72) have unit determinant. However, the self-dual condition also requires the Cartan matrix to be integral. From the expressions (39.72) it follows that the Cartan matrices are integral if and only if n satisfies the condition

$$\begin{cases} n+1 = km^2 \text{ where } m \equiv n+1-p & \text{SU}(n+1) , \\ n = 2k & \text{Spin}(2n+1) \text{ or } \text{Sp}(n) , \\ n = 4k & \text{Spin}(2n) , \end{cases} \quad (39.73)$$

for some integral k . For n satisfying condition (39.73), the lengths squared of the spinor roots σ_p are, from equation (39.71),

$$s_p = \begin{cases} (km-1)/k & \text{SU}(km^2) , \\ k/2 & \text{Spin}(4k+1) , \\ 2k & \text{Sp}(2k) , \\ k & \text{Spin}(8k) . \end{cases} \quad (39.74)$$

According to condition (39.73), an $\text{SU}(n+1)$ lattice with spinor root σ_p is self-dual if $n+1 = km^2$ with $m \equiv n+1-p$. It is convenient to call such a lattice an m lattice.

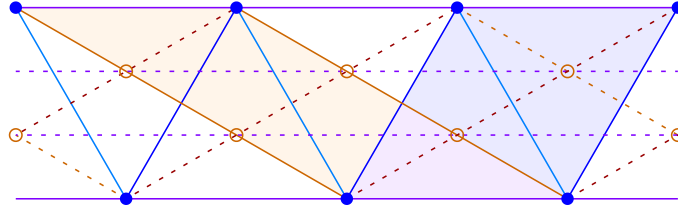


Figure 39.2 Standard (blue) and self-dual $m = 1$ (orange) $SU(3)$ lattices. Vertices with integral coordinates are coloured blue; those with nonintegral fraction coordinates are coloured orange. The blue-shaded region (two equilateral triangles) is a fundamental tile of the standard $SU(3)$ lattice, generated by the two simple roots α_1 and α_2 , signified by the purple and blue solid lines. Altogether there are $\frac{1}{2}n(n+1) = 3$ simple roots of length squared equal to two, signified by the purple/blue/cyan solid lines, which represent massless gauge bosons; the simple roots are raising operators in one direction, lowering operators in the opposite direction, for a total of $n(n+1) = 6$ degrees of freedom. The self-dual (orange) lattice contains spinor roots (dashed orange/maroon lines) of length squared $(km-1)/k = 2/3$, equation (39.74), which is less than two, which would represent tachyonic (negative mass-squared) gauge bosons if they were present. These tachyonic roots are eliminated by taking the $(km = 3)$ -fold cover of the self-dual lattice. The orange-shaded region is a fundamental tile of the 3-fold cover of the self-dual lattice, generated by the spinor root $km\sigma_p = 3\sigma_2$ together with the simple root α_2 , signified by the orange and purple solid lines. The 3-fold cover of the self-dual lattice is again self-dual, and its nearest-neighbour roots of length squared two coincide with those of the standard lattice.

Figure 39.2 illustrates the standard and self-dual $m = 1$ lattices for $SU(3)$. The $m = 1$ self-dual lattice has $n = 2$, $p = 2$, and $k = 3$. The Cartan matrix for the $SU(3)$ $m = 1$ self-dual lattice happens to coincide with the inverse of the Cartan matrix of the exceptional rank 2 Lie group G_2 (whose Cartan matrix has unit determinant); but this is just a low-dimensional coincidence.

The spinor-root-substituted lattices considered above have root systems in which the spinor root σ_p replaces the standard simple root α_1 . Spinor roots comprise a subset σ_p of the dual roots of the standard set of simple roots α_i . The dual roots are given in terms of the standard roots by equation (39.61) with C_{ij} the standard Cartan matrix (39.29). Thus the first root α_1 is given in terms of spinor root $\sigma_p = \hat{\alpha}_p$ and standard simple roots α_i with $i \geq 2$ by

$$\alpha_1 = \frac{1}{C_{p1}^{-1}} \left(\sigma_p - \sum_{i=2}^n C_{pi}^{-1} \alpha_i \right), \quad (39.75)$$

with C the standard Cartan matrix (39.29). A spinor root σ_q is then given in terms of spinor root σ_p and standard simple roots α_i with $a \geq 2$ by

$$\sigma_q = \frac{C_{q1}^{-1}}{C_{p1}^{-1}} \sigma_p + \sum_{i=2}^n \left(C_{qi}^{-1} - \frac{C_{q1}^{-1} C_{pi}^{-1}}{C_{p1}^{-1}} \right) \alpha_i. \quad (39.76)$$

Thus the spinor root σ_q belongs to the lattice with spinor root σ_p provided that C_{q1}^{-1}/C_{p1}^{-1} is integral. But

$$C_{p1}^{-1} = \begin{cases} (n+1-p)/(n+1) & \text{SU}(n+1) , \\ 1/2 & \text{Spin}(2n+1) \text{ or Spin}(2n) , \\ 1 & \text{Sp}(n) . \end{cases} \quad (39.77)$$

The ratio C_{q1}^{-1}/C_{p1}^{-1} is therefore

$$\frac{C_{q1}^{-1}}{C_{p1}^{-1}} = \begin{cases} \frac{n+1-q}{n+1-p} & \text{SU}(n+1) , \\ 1 & \text{otherwise} . \end{cases} \quad (39.78)$$

For $\text{SU}(n+1)$, the ratio C_{q1}^{-1}/C_{p1}^{-1} equals an integer j if $n+1-q = jm$ and $n+1-p = m$. Therefore a jm lattice is a factor of j coarser than an m lattice. The standard lattice is itself a factor of $1/C_{p1}^{-1} = (n+1)/(n+1-p) = km$ coarser than an m lattice.

For example, the subgroup $\text{SU}(16)$ of $G^2(10) = \text{U}(16)$ from Table ?? has three self-dual lattices, with $m = 1, 2, 4$. The $m = 1$ lattice is a factor of 2 finer than the $m = 2$ lattice, which is a factor of 2 finer than the $m = 4$ lattice, which is in turn a factor of 4 finer than the standard lattice. The subgroup $\text{SU}(8)$ of $\text{SU}(16)$ has two self-dual lattices, with $m = 1, 2$. The $m = 1$ lattice is a factor of 2 finer than the $m = 2$ lattice, which is in turn a factor of 4 finer than the standard lattice. The $\text{SU}(8) \times \text{SU}(8)$ $m = 1$ and $m = 2$ lattices are sublattices of the $\text{SU}(16)$ $m = 2$ and $m = 4$ lattices.

According to equation (39.77), the self-dual lattices of $\text{Spin}(2n+1)$ and $\text{Spin}(2n)$ are a factor of 2 finer than the standard lattice, while the self-dual lattices of $\text{Sp}(n)$ coincide with the standard lattice.

A direct product of self-dual Cartan matrices of groups yields a self-dual Cartan matrix for the direct product of the groups. The self-dual Cartan matrix of a group of rank 1 can be taken to be the 1×1 unit matrix. There exist self-dual Cartan matrices of direct products of groups that are not just direct products over self-dual Cartan matrices (Conway and Sloane, 1982).

39.4 Simply-laced root systems

When open strings are quantized on a compactification torus, zero-mass modes of the bosonic string must have momentum squared equal to 2, §??, as originally argued for heterotic strings by Gross et al. (1985). This translates into the requirement that the string modes be quantized on nearest-neighbour pairs with separation squared equal to 2.

The standard root system of $\text{SU}(n+1)$ has $n(n+1)$ roots (raising and lowering operators) of the form

$$\{..0.., 1, ..0.., -1, ..0..\} \text{ and } \{..0.., -1, ..0.., 1, ..0..\} , \quad (39.79)$$

which have lengths squared 2. The standard root systems of $\text{Spin}(2n+1)$, $\text{Sp}(n)$ and $\text{Spin}(2n)$ have, in addition to $(n-1)n$ roots of the form (39.79), $(n-1)n$ roots of the form

$$\{..0.., 1, ..0.., 1, ..0..\} \text{ and } \{..0.., -1, ..0.., -1, ..0..\} , \quad (39.80)$$

which likewise have lengths squared 2.

Consider first $SU(n+1)$ and $Spin(2n)$; the cases of $Spin(2n+1)$ and $Sp(n)$ will be addressed below. According to equation (39.77), self-dual lattices of $SU(n+1)$ are a factor $1/C_{p1}^{-1} = (n+1)/(n+1-p) = km$ finer than the standard lattice, while self-dual lattices of $Spin(2n)$ are a factor of $1/C_{p1}^{-1} = 2$ finer than the standard lattice. These self-dual lattices have nearest-neighbour pairs with separations squared less than 2, which would yield problematic negative mass-squared tachyonic modes if they were present. The solution to the tachyonic threat is to take the km -fold cover of the torus for $SU(n+1)$, and the 2-fold cover of the torus for $Spin(2n)$. This is accomplished by taking the lattice to be generated by km times its spinor root σ_p for $SU(n+1)$, or 2 times the spinor root σ_p for $Spin(2n)$, along with the standard roots α_i , $i = 2, \dots, n$. The resulting lattice coincides with the standard lattice. The self-dual Cartan matrix C is deemed unchanged from the original expressions (39.72) by the covering, and the lattice remains self-dual. This strategy was previously adopted by Gross et al. (1985) to construct a self-dual lattice for $Spin(32)$ in heterotic string theory, by taking the double cover of the self-dual lattice with spinor root $\sigma_{n-1} = \sigma_{15}$.

The covering procedure is illustrated for the $SU(3)$ $m = 1$ lattice in Figure 39.2. The 3-fold cover of the $SU(3)$ $m = 1$ torus coincides with the standard self-dual torus of the exceptional group G_2 .

The standard lattices of $Spin(2n+1)$ and $Sp(n)$ have, in addition to $(n-1)n$ roots of the form (39.79) and $(n-1)n$ roots of the form (39.80), $2n$ roots of the form

$$\begin{cases} \{..0.., 1, ..0..\} \text{ and } \{..0.., -1, ..0..\} & Spin(2n+1) , \\ \{..0.., 2, ..0..\} \text{ and } \{..0.., -2, ..0..\} & Sp(n) , \end{cases} \quad (39.81)$$

which have lengths squared respectively 1 and 4. The potentially tachyonic length-squared 1 roots of $Spin(2n+1)$ can be eliminated by taking the double cover of the torus with respect to the simple root α_n , yielding roots with lengths squared 4 like those of $Sp(n)$. However, the length-squared 4 roots yield massive rather than massless bosons. The spectrum of massless gauge bosons is not that of $Spin(2n+1)$ or $Sp(n)$, but rather that of the subgroup $Spin(2n)$. The groups $Spin(2n+1)$ and $Sp(n)$ are not viable as gauge groups for bosonic strings.

The supergeometric algebra

The **supergeometric algebra** generalizes the geometric algebra to include spinors, which are spin- $\frac{1}{2}$ objects.

Introduced originally by Cartan in 1913, a spinor is the fundamental representation of the group of rotations in N spacetime dimensions. Cartan's (1938) introduction to spinor algebras is an excellent resource.

For simplicity, this Chapter focuses on the supergeometric algebra in 3 spatial dimensions. The generalization to arbitrarily many spatial dimensions is given as Exercise 40.3 at the end of the Chapter. The generalization of the supergeometric algebra to Minkowski space, with a time dimension in addition to spatial dimensions, is presented in Chapter 41. The generalization to arbitrarily many space and time dimensions is given as Exercise 41.5.

40.1 Spin basis vectors in 3D

A systematic way to project tensors into spin components is to work in a spin basis. Start with an orthonormal triad $\{\gamma_1, \gamma_2, \gamma_3\}$ (or $\{\gamma_x, \gamma_y, \gamma_z\}$ if you prefer). Choose a pair of basis vectors, in three dimensions conventionally taken to be the pair $\{\gamma_1, \gamma_2\}$, and from them form the chiral basis vectors $\{\gamma_+, \gamma_-\}$, the complex combinations

$$\boxed{\gamma_+ \equiv \frac{1}{\sqrt{2}}(\gamma_1 + i\gamma_2)} , \quad (40.1a)$$

$$\boxed{\gamma_- \equiv \frac{1}{\sqrt{2}}(\gamma_1 - i\gamma_2)} . \quad (40.1b)$$

This is the same trick used to define the spin components $L_{\pm}$ of the angular momentum operator $\mathbf{L}$ in quantum mechanics. The metric of the spin triad $\{\gamma_+, \gamma_-, \gamma_3\}$ is

$$\gamma_{ab} \equiv \gamma_a \cdot \gamma_b = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \quad (40.2)$$

Notice that the chiral basis vectors $\{\gamma_+, \gamma_-\}$ are themselves null, $\gamma_+ \cdot \gamma_+ = \gamma_- \cdot \gamma_- = 0$, whereas their scalar product with each other is nonzero $\gamma_+ \cdot \gamma_- = 1$.

40.2 Spin weight

An object is defined to have **spin weight** s if it varies by

$$e^{-is\theta} \quad (40.3)$$

under a right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane. In 3D, a right-handed rotation in the $\gamma_1\gamma_2$ -plane is the same as a right-handed rotation about the 3-axis, and the spin weight is the projection of the spin along the 3-axis, the spin analogue of the projection L_3 (or L_z) of the angular momentum along the 3-axis (or z -axis). Sometimes the term spin weight is abbreviated to spin, when there is no ambiguity. An object of spin weight s is unchanged by a rotation of $2\pi/s$ in the $\gamma_1\gamma_2$ -plane. An object of spin weight 0 is rotationally symmetric, unchanged by a rotation by any angle in the $\gamma_1\gamma_2$ -plane.

Under a right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane, the basis vectors γ_a transform as (13.51)

$$\begin{aligned} \gamma_1 &\rightarrow \cos\theta \gamma_1 + \sin\theta \gamma_2, \\ \gamma_2 &\rightarrow \sin\theta \gamma_1 - \cos\theta \gamma_2, \\ \gamma_3 &\rightarrow \gamma_3. \end{aligned} \quad (40.4)$$

It follows that the chiral basis vectors γ_+ and γ_- transform under a right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane

$$\gamma_{\pm} \rightarrow e^{\mp i\theta} \gamma_{\pm}. \quad (40.5)$$

The transformation (40.5) identifies the spin vectors γ_+ and γ_- as having spin weight $+1$ and -1 respectively. The γ_3 vector has spin weight 0, since it is unchanged by a rotation in the $\gamma_1\gamma_2$ -plane.

The components of a tensor in a spin basis inherit their spin properties from that of the spin basis. The general rule is that the spin weight s of any tensor component is equal to the number of $+$ covariant indices minus the number of $-$ covariant indices:

$$\boxed{\text{spin weight } s = \text{number of } + \text{ minus } - \text{ covariant indices}}. \quad (40.6)$$

The spin properties of the components of a tensor are thus manifest when expressed in a spin basis.

40.3 Pauli representation of chiral basis vectors

In the Pauli representation (13.112), the chiral basis vectors $\gamma_{\pm}$ are represented by the real 2×2 Pauli matrices

$$\gamma_+ = \sigma_+ \equiv \frac{1}{\sqrt{2}}(\sigma_1 + i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \gamma_- = \sigma_- \equiv \frac{1}{\sqrt{2}}(\sigma_1 - i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \quad (40.7)$$

The basis vector γ_3 is represented as usual by the real Pauli matrix σ_3 ,

$$\gamma_3 = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (40.8)$$

40.4 Basis spinors

Introduce a dyad of basis spinors ϵ_a with the index a running over spin up $\uparrow$ and spin down $\downarrow$ (the braces in equation (40.9) signify a set of spinors, not anticommutation),

$$\boxed{\epsilon_a \equiv \{\epsilon_\uparrow, \epsilon_\downarrow\}} . \quad (40.9)$$

The basis spinors $\epsilon_\uparrow$ and $\epsilon_\downarrow$ physically signify spin up and spin down eigenstates. It will be seen in §40.11 that the basis spinors ϵ_a are related to the 3D basis vectors γ_a through a supergeometric algebra that is essentially the square root of the geometric algebra. Elements of the geometric algebra act by pre-multiplication on the basis spinors ϵ_a . Under a rotation by rotor R , the basis spinors ϵ_a are defined to transform in the same way as rotors,

$$R : \epsilon_a \rightarrow R\epsilon_a . \quad (40.10)$$

In the Pauli representation (13.112) the basis spinors ϵ_a are the column vectors

$$\epsilon_\uparrow = \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \quad \epsilon_\downarrow = \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \quad (40.11)$$

that are rotated by pre-multiplying by elements of the special unitary group SU(2). Rotations transform the basis spinors ϵ_a into linear combinations of each other.

The rotor R corresponding to a right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane is $e^{-\imath_3\theta/2}$, equation (13.106). In the Pauli representation (40.9), the action of $\imath_3 = I_3\sigma_3$ on the basis spinors is $\imath_3\epsilon_\uparrow = i\epsilon_\uparrow$ and $\imath_3\epsilon_\downarrow = -i\epsilon_\downarrow$. Under a right-handed rotation by angle θ in the $\gamma_1\gamma_2$ -plane, the basis spinors ϵ_a therefore transform as

$$\epsilon_\uparrow \rightarrow e^{-i\theta/2}\epsilon_\uparrow , \quad \epsilon_\downarrow \rightarrow e^{i\theta/2}\epsilon_\downarrow . \quad (40.12)$$

The behaviour (40.12), along with the definition (40.3) of spin, shows that the basis spinors $\epsilon_\uparrow$ and $\epsilon_\downarrow$ have respective spin weights $+\frac{1}{2}$ and $-\frac{1}{2}$. A rotation by $\theta = 2\pi$ changes the sign of the basis spinors ϵ_a . A rotation by 4π is required to rotate the basis spinors back to their original values.

Spinor tensors inherit their spin properties from those of the basis spinors. The rule (40.6) generalizes to the statement that the spin weight of a spinor tensor is

$$\boxed{\text{spin weight } s = \frac{1}{2} (\text{number of } \uparrow \text{ minus } \downarrow \text{ covariant indices})} . \quad (40.13)$$

In any equality between vector and spinor tensors, the spin weights of the left and right hand sides must be equal. The rule (40.13) hold not only for column spinors ϵ_a , but also for row spinors ϵ_a , §40.7, and for inner and outer products of spinors, §§40.8 and 40.10.

40.5 Pauli spinor

A **Pauli spinor** φ is a complex (with respect to i) linear combination of the basis spinors ϵ_a ,

$$\boxed{\varphi = \varphi^a \epsilon_a} . \quad (40.14)$$

Just as a multivector $a^a \gamma_a$ is a vector in the geometric algebra, so also $\varphi^a \epsilon_a$ is a spinor in the supergeometric algebra.

By construction, a Pauli spinor transforms under a spatial rotation by rotor R like the basis spinors, equation (40.10),

$$R : \varphi \rightarrow R\varphi . \quad (40.15)$$

A Pauli spinor φ is a spin- $\frac{1}{2}$ object, in the sense that a rotation by 2π changes the sign of the spinor, and a rotation by 4π is required to return the spinor to its original value.

40.6 Spinor metric

In a matrix representation, the tensor product of basis spinors ϵ_a and ϵ_b can be represented as the 2×2 matrix $\epsilon_a \epsilon_b^\top$, a matrix product of the column spinor ϵ_a with the row spinor $\epsilon_b^\top$. In accordance with the transformation rule (40.10), the tensor product of basis spinors rotates as

$$R : \epsilon_a \epsilon_b^\top \rightarrow R \epsilon_a \epsilon_b^\top R^\top . \quad (40.16)$$

Consider the spinor tensor ε with the defining property that for any rotor R

$$\varepsilon R^\top = \tilde{R} \varepsilon . \quad (40.17)$$

The condition (40.17) implies that the spinor tensor ε is invariant under rotations,

$$R : \varepsilon \rightarrow R \varepsilon R^\top = R \tilde{R} \varepsilon = \varepsilon . \quad (40.18)$$

The spinor tensor ε is the **spinor metric**. Like the Euclidean metric, it is that tensor which remains invariant under rotations.

Since a rotor is a linear combination of even elements 1 and $I_3 \gamma_a$ of the geometric algebra, and bivectors $I_3 \gamma_a$ change sign under reversal, a necessary and sufficient condition for (40.17) is

$$\varepsilon (I_3 \gamma_a)^\top = -I_3 \gamma_a \varepsilon \quad \text{for } a = 1, 2, 3 . \quad (40.19)$$

In the Pauli representation (13.112), where $\gamma_a = \sigma_a$ and I_3 equals i times the unit matrix, the condition (40.19) requires that ε commutes with γ_2 , and anticommutes with γ_1 and γ_3 . The only basis element of the spacetime algebra with the required (anti)commutation properties is γ_2 , so the spinor metric ε must equal γ_2 up to a possible scalar normalization,

$$\varepsilon \equiv i \gamma_2 . \quad (40.20)$$

In the Pauli representation (13.112), the spinor metric (40.20) is the antisymmetric matrix

$$\varepsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \quad (40.21)$$

The chosen normalization is such that ε is real (with respect to i). The spinor metric ε is then orthogonal, and its square is minus the unit matrix,

$$\varepsilon^{-1} = \varepsilon^\top , \quad \varepsilon^2 = -1 . \quad (40.22)$$

Despite the equality of ε and $i\gamma_2$ in the Pauli representation, ε is defined to transform as a spinor tensor under spatial rotations, not as an element of the geometric algebra. The components of the spinor metric matrix ε constitute the spinor metric ε_{ab} ,

$$\epsilon_a^\top \varepsilon \epsilon_b = \varepsilon_{ab} . \quad (40.23)$$

Commuting the spinor metric ε through the orthonormal basis vectors γ_a converts them to minus their transposes,

$$\gamma_a^\top \varepsilon = -\varepsilon \gamma_a . \quad (40.24)$$

40.7 Row basis spinors

It is convenient to use the symbol $\epsilon_a \cdot$ with a trailing dot, symbolic of the trailing ε , to denote the row spinor $\epsilon_a^\top \varepsilon$,

$$\boxed{\epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon} . \quad (40.25)$$

The motivation for the trailing dot notation is equation (40.29) below. The two row spinors (the braces in equation (40.26) signify a set of spinors, not anticommutation)

$$\epsilon_a \cdot = \{\epsilon_\uparrow \cdot, \epsilon_\downarrow \cdot\} \quad (40.26)$$

provide a basis for row spinors. The spin weights of the row basis spinors are in accord with their covariant indices: $\epsilon_\uparrow \cdot$ has spin weight $+\frac{1}{2}$, while $\epsilon_\downarrow \cdot$ has spin weight $-\frac{1}{2}$. The row spinors $\epsilon_a \cdot$ rotate as

$$R : \epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon \rightarrow \epsilon_a^\top R^\top \varepsilon = \epsilon_a^\top \varepsilon \tilde{R} = \epsilon_a \cdot \tilde{R} . \quad (40.27)$$

Thus row spinors $\epsilon_a \cdot$ transform like reverse rotors, just as column spinors ϵ_a transform like rotors. In the Pauli representation (13.112) the row basis spinors $\epsilon_a \cdot$ are the row spinors

$$\epsilon_\uparrow \cdot = \begin{pmatrix} 0 & 1 \end{pmatrix} , \quad \epsilon_\downarrow \cdot = \begin{pmatrix} -1 & 0 \end{pmatrix} . \quad (40.28)$$

40.8 Inner products of basis spinors

The product of the row spinor $\epsilon_a \cdot$ with the column spinor ϵ_b defines their inner product, or scalar product, which equals the spinor metric ε_{ab} in accordance with equation (40.23),

$$\boxed{\epsilon_a \cdot \epsilon_b = \varepsilon_{ab}} . \quad (40.29)$$

Equation (40.29) motivates the trailing dot notation for the row spinor. The scalar product is antisymmetric,

$$\boxed{\epsilon_a \cdot \epsilon_b = -\epsilon_b \cdot \epsilon_a} . \quad (40.30)$$

In the Pauli representation, the nonzero components of the scalar product are explicitly, equation (40.21),

$$\epsilon_{\uparrow} \cdot \epsilon_{\downarrow} = -\epsilon_{\downarrow} \cdot \epsilon_{\uparrow} = 1 . \quad (40.31)$$

The antisymmetry of the spinor scalar product contrasts with the symmetry of the usual vector scalar product. The scalar product (40.29) is a scalar,

$$R : \epsilon_a \cdot \epsilon_b \rightarrow \epsilon_a \cdot \widetilde{R} R \epsilon_b = \epsilon_a \cdot \epsilon_b . \quad (40.32)$$

Thus the spinor metric ε_{ab} is invariant under rotations, just like the Euclidean metric δ_{ab} .

40.9 Lowering and raising spinor indices

The antisymmetric spinor metric ε_{ab} is given in the Pauli representation by equation (40.23). The inverse metric ε^{ab} is defined by $\varepsilon^{ab}\varepsilon_{bc} = \delta_c^a$. The spinor metric and its inverse satisfy

$$\varepsilon_{ab} = -\varepsilon_{ba} = -\varepsilon^{ab} = \varepsilon^{ba} . \quad (40.33)$$

Indices on a spinor tensor are lowered and raised by pre-multiplying by the metric ε_{ab} and its inverse ε^{ab} . The contravariant components ϵ^a of the column basis spinors, satisfying $\epsilon^a = \varepsilon^{ab}\epsilon_b$, are

$$\epsilon^{\uparrow} = -\epsilon_{\downarrow} , \quad \epsilon^{\downarrow} = \epsilon_{\uparrow} . \quad (40.34)$$

For example, $\epsilon^{\uparrow} = \varepsilon^{\uparrow\downarrow}\epsilon_{\downarrow} = -\epsilon_{\downarrow}$. A spinor index is lowered or raised by pre-multiplying by the metric or its inverse: post-multiplying by the metric or its inverse yields a result of opposite sign, $\epsilon^a = \varepsilon^{ab}\epsilon_b = -\epsilon_b\varepsilon^{ba}$. The contravariant components $\epsilon^a \cdot$ of the row basis spinors satisfy the same relations (40.34) with a trailing dot appended on left and right hand sides. The scalar products of contravariant row with covariant column basis spinors form the unit matrix,

$$\epsilon^a \cdot \epsilon_b = -\epsilon_b \cdot \epsilon^a = \delta_b^a . \quad (40.35)$$

40.9.1 Scalar products of Pauli spinors

A general row spinor $\varphi \cdot$ is a complex (with respect to i) linear combination of the row basis spinors

$$\boxed{\varphi \cdot \equiv \varphi^\top \varepsilon = \varphi^a \epsilon_a \cdot} . \quad (40.36)$$

It rotates as

$$R: \varphi \cdot \rightarrow \varphi \cdot \tilde{R} . \quad (40.37)$$

A row spinor $\varphi \cdot$ transforms like a reverse rotor.

The product of a row Pauli spinor $\varphi \cdot = \varphi^a \epsilon_a \cdot$ with a column Pauli spinor $\chi = \chi^a \epsilon_a$ forms a scalar, which may be written variously

$$\varphi \cdot \chi = \varphi^\top \varepsilon \chi = \varphi^a \epsilon_a \cdot \chi^b \epsilon_b = \varepsilon_{ab} \varphi^a \chi^b = \varphi^a \chi_a = -\varphi_a \chi^a = -\varepsilon^{ab} \varphi_a \chi_b . \quad (40.38)$$

Notice that when the scalar product $\varphi \cdot \chi$ is written in the contracted form $\varphi^a \chi_a$, the first index is raised and the second is lowered. An additional minus sign appears if the first index is lowered and the second is raised.

The components φ^a of a column spinor φ can be projected out by pre-multiplying by the row basis spinor $\epsilon^a \cdot$,

$$\epsilon^a \cdot \varphi = \epsilon^a \cdot \varphi^b \epsilon_b = \delta_b^a \varphi^b = \varphi^a . \quad (40.39)$$

The components φ^a of a row spinor $\varphi \cdot$ can be projected out by post-multiplying by minus the column basis spinor ϵ^a ,

$$-\varphi \cdot \epsilon^a = -\varphi^b \epsilon_b \cdot \epsilon^a = \delta_b^a \varphi^b = \varphi^a . \quad (40.40)$$

If the coefficients φ^a and χ^b of Pauli spinors $\varphi = \varphi^a \epsilon_a$ and $\chi = \chi^b \epsilon_b$ are taken to be ordinary commuting complex numbers, then the Pauli scalar product is anticommuting

$$\boxed{\varphi \cdot \chi = -\chi \cdot \varphi} . \quad (40.41)$$

40.10 Outer products of basis spinors

A row spinor $\epsilon_a \cdot$ multiplied by a column spinor ϵ_b yields their scalar product. In the opposite order, a column spinor ϵ_a multiplied by a row spinor $\epsilon_b \cdot$ yields their outer product. The outer product $\epsilon_a \epsilon_b \cdot$ rotates like a multivector in the geometric algebra,

$$R: \epsilon_a \epsilon_b \cdot \equiv \epsilon_a \epsilon_b^\top \varepsilon \rightarrow R \epsilon_a \epsilon_b^\top R^\top \varepsilon = R \epsilon_a \epsilon_b^\top \varepsilon \tilde{R} = R \epsilon_a \epsilon_b \cdot \tilde{R} . \quad (40.42)$$

The trailing dot on the outer product $\epsilon_a \epsilon_b \cdot$ is symbolic of the trailing ε , necessary to convert the spinor tensor $\epsilon_a \epsilon_b^\top$ into an object that transforms like a multivector.

The products of the 2 column basis spinors ϵ_a with the 2 row basis spinors $\epsilon_b \cdot$ form 4 outer products. The 3D geometric algebra has 8 basis elements, but the pseudoscalar I_3 is a commuting imaginary which in the

Pauli representation is just i times the unit matrix, so the 3D geometric algebra is equivalent to a complex algebra with 4 basis elements. The 4 outer products of basis spinors thus suffice to generate the complete complex 3D geometric algebra. In the Pauli representation (13.112), the 4 outer products of basis spinors map to elements of the 3D geometric algebra as follows.

The antisymmetric outer products of spinors form a scalar singlet,

$$[\epsilon_{\downarrow}, \epsilon_{\uparrow}] \cdot = 1 , \quad (40.43)$$

where the 1 on the right hand side denotes the unit element of the 3D geometric algebra, the 2×2 identity matrix. The trailing dot on the commutator indicates that the right partner of each product is a row spinor, $[\epsilon_{\uparrow}, \epsilon_{\downarrow}] \cdot = \epsilon_{\uparrow} \epsilon_{\downarrow}^{\top} \varepsilon - \epsilon_{\downarrow} \epsilon_{\uparrow}^{\top} \varepsilon$. The spin weight of the singlet (40.43) is zero according to the rule (40.13), as it should be for a scalar.

The symmetric outer products of spinors form a triplet,

$$\{\epsilon_{\uparrow}, \epsilon_{\uparrow}\} \cdot = \sqrt{2} \gamma_+ , \quad \{\epsilon_{\uparrow}, \epsilon_{\downarrow}\} \cdot = -\gamma_3 , \quad \{\epsilon_{\downarrow}, \epsilon_{\downarrow}\} \cdot = -\sqrt{2} \gamma_- . \quad (40.44)$$

The spin weights of the triplet (40.44) are respectively +1, 0, -1 according to the rules (40.6) and (40.13). The spin weights of left and right hand sides match, as they should.

The trace of the outer product of a pair of basis spinors gives their scalar product (note that the 1 on the right hand side of equation (40.43) is the unit matrix, whose trace is 2),

$$\text{Tr } \epsilon_a \epsilon_b \cdot = \epsilon_b \cdot \epsilon_a = \varepsilon_{ba} . \quad (40.45)$$

The expansion of the 4 outer products $\epsilon_a \epsilon_b \cdot$ of spinors in terms of the basis elements γ_A of the geometric algebra, and vice versa, define the matrix of coefficients γ_{ab}^A and its inverse γ_A^{ab} ,

$$\epsilon_a \epsilon_b \cdot = \gamma_{ab}^A \gamma_A , \quad \gamma_A = \gamma_A^{ab} \epsilon_a \epsilon_b \cdot . \quad (40.46)$$

The coefficients γ_{ab}^A and γ_A^{ab} in the chiral representation are

$$\gamma_{ab}^A = \frac{1}{2} \epsilon_b \cdot \gamma^A \epsilon_a , \quad \gamma_A^{ab} = -\epsilon^a \cdot \gamma_A \epsilon^b . \quad (40.47)$$

Exercise 40.1 Consistency of spinor and multivector scalar products. Confirm that the spinor and multivector scalar products are consistent.

Solution. Multivector vectors are equivalent to outer products of Pauli spinors in accordance with equations (40.44). For example, the scalar product of the multivectors γ_+ and γ_- is

$$\begin{aligned} \gamma_+ \cdot \gamma_- &= \frac{1}{2} (\gamma_+ \gamma_- + \gamma_- \gamma_+) \\ &= -\frac{1}{4} (\{\epsilon_{\uparrow}, \epsilon_{\uparrow}\} \cdot \{\epsilon_{\downarrow}, \epsilon_{\downarrow}\} \cdot + \{\epsilon_{\downarrow}, \epsilon_{\downarrow}\} \cdot \{\epsilon_{\uparrow}, \epsilon_{\uparrow}\} \cdot) \\ &= -(\epsilon_{\uparrow}(\epsilon_{\uparrow} \cdot \epsilon_{\downarrow})\epsilon_{\downarrow} \cdot + \epsilon_{\downarrow}(\epsilon_{\downarrow} \cdot \epsilon_{\uparrow})\epsilon_{\uparrow} \cdot) \\ &= -\epsilon_{\uparrow} \epsilon_{\downarrow} \cdot + \epsilon_{\downarrow} \epsilon_{\uparrow} \cdot \\ &= [\epsilon_{\downarrow}, \epsilon_{\uparrow}] \cdot \\ &= 1 , \end{aligned} \quad (40.48)$$

the fourth step of which invokes the spinor scalar product (40.31), and the last step of which is from the equivalence (40.43). The result agrees with the multivector scalar product (40.2).

40.11 The 3D supergeometric algebra

The 3D supergeometric algebra comprises 4 distinct species of objects: true scalars, column spinors, row spinors, and multivectors. In a matrix representation, they are complex (with respect to i) matrices with dimensions 1×1 , 1×2 , 2×1 , and 2×2 . A true scalar, the value of a scalar product of spinors, is just a complex number. A column spinor φ is a complex linear combination of column basis spinors ϵ_a ,

$$\varphi = \varphi^a \epsilon_a , \quad (40.49)$$

while a row spinor $\varphi \cdot$ is a complex linear combination of row basis spinors $\epsilon_a \cdot$,

$$\varphi \cdot = \varphi^a \epsilon_a \cdot . \quad (40.50)$$

A multivector $\mathbf{a}$ is a complex linear combination of outer products of the column and row basis spinors,

$$\mathbf{a} = a^{ab} \epsilon_a \epsilon_b \cdot . \quad (40.51)$$

Linearity and the transformation law (40.42) imply that the algebra of sums and products of outer products of spinors is isomorphic to the geometric algebra.

A unit multivector of grade 0, proportional to the unit 2×2 matrix, the value of a scalar product of vectors, is conventionally referred to as a scalar (and a multivector of grade 3 as a pseudoscalar), but it must be distinguished from a true scalar, a complex number. This is the reason for prepending the adjective “true” to a true scalar.

As seen in §40.8 and §40.10, a column spinor φ and a row spinor $\chi \cdot$ can be multiplied in either order, yielding an inner product which is a scalar, and an outer product which is a multivector. However, a column spinor cannot be multiplied by a column spinor, and likewise a row spinor cannot be multiplied by a row spinor, as is manifestly true in a matrix representation. Rather than prohibiting certain kinds of multiplication, it is advantageous instead to assert that prohibited multiplications yield zero,

$$\varphi \chi = 0 , \quad \varphi \cdot \chi \cdot = 0 , \quad \varphi \mathbf{a} = 0 , \quad \mathbf{a} \varphi \cdot = 0 , \quad (40.52)$$

which allows all objects in the supergeometric algebra to be added and multiplied, regardless of their species. In quantum field theory, row and column spinors are promoted to destruction and creation operators, §44.4.2, and the result of acting on the vacuum state $|0\rangle$ with a destruction operator, equations (44.50), is asserted to be 0. This is the 0 in equations (40.52); the 0 is not an element of the supergeometric algebra.

In general, a sequence of products of spinors yields a nonzero result provided that they alternate between column spinor and row spinor,

$$\varphi \chi \cdot \psi \quad \text{or} \quad \varphi \cdot \chi \psi \cdot . \quad (40.53)$$

Both product sequences are associative,

$$\varphi \chi \cdot \psi = (\varphi \chi \cdot) \psi = \varphi (\chi \cdot \psi) , \quad (40.54a)$$

$$\varphi \cdot \chi \psi \cdot = (\varphi \cdot \chi) \psi \cdot = \varphi \cdot (\chi \psi \cdot) . \quad (40.54b)$$

A product of an even number of spinors yields a scalar or a multivector depending on whether the first spinor is a row or a column spinor. A product of an odd number of spinors yields a row spinor or a column spinor depending on whether the first spinor is a row or a column spinor.

The scalar product and the associative law make it straightforward to simplify long sequences of products. Let $\mathbf{a} = a^{ab} \epsilon_a \epsilon_b \cdot$ and $\mathbf{b} = b^{ab} \epsilon_a \epsilon_b \cdot$ be two multivectors expressed as a sum of outer products of spinors. Their product is the multivector

$$\mathbf{a} \mathbf{b} = a^{ab} \epsilon_a \epsilon_b \cdot b^{cd} \epsilon_c \epsilon_d \cdot = \epsilon_a a^{ab} \epsilon_{bc} b^{cd} \epsilon_d \cdot = \epsilon_a a^{ab} b_b{}^d \epsilon_d \cdot . \quad (40.55)$$

A sequence such as $\varphi \cdot \mathbf{a} \chi$ simplifies as

$$\varphi \cdot \mathbf{a} \chi = \varphi^a \epsilon_a \cdot a^{bc} \epsilon_b \epsilon_c \cdot \chi^d \epsilon_d = \varphi^a \epsilon_{ab} a^{bc} \epsilon_{cd} \chi^d = \varphi^a a_a{}^c \chi_c . \quad (40.56)$$

The trace, equation (40.45), of an outer product of spinors is a true scalar

$$\text{Tr } \chi \varphi \cdot = \chi^a \varphi^b \epsilon_{ba} = -\chi \cdot \varphi = \varphi \cdot \chi , \quad (40.57)$$

the last step of which assumes that the coefficients χ^a and φ^b are ordinary commuting complex numbers, equation (40.41).

40.12 Conjugate Pauli spinor

The 3D supergeometric algebra possesses a discrete transformation called conjugation. The conjugate Pauli spinor $\bar{\varphi}$ is defined by equation (40.60). It has the defining properties that (a) its components are complex conjugates (with respect to i) of those of the parent spinor φ , and (b) the conjugate spinor $\bar{\varphi}$ rotates in the same way as the spinor φ .

The complex conjugate φ^* of a Pauli spinor $\varphi = \varphi^a \epsilon_a$ is defined to be the spinor with complex conjugate (with respect to i) coefficients,

$$\varphi^* \equiv \varphi^{a*} \epsilon_a . \quad (40.58)$$

In effect, the basis spinors ϵ_a are taken to be real, just as the basis vectors $\gamma_{\pm}$ and γ_3 in the spin basis are real, equations (40.7). Since the spinor φ rotates under a rotor R as $\varphi \rightarrow R\varphi$, its complex conjugate φ^* rotates according to the complex conjugate representation of the Pauli matrices,

$$R : \varphi^* \rightarrow (R\varphi)^* = R^* \varphi^* . \quad (40.59)$$

The conjugate Pauli spinor $\bar{\varphi}$ is defined by

$$\boxed{\bar{\varphi} \equiv \varepsilon \varphi^*} . \quad (40.60)$$

The 3D spinor metric tensor ε was constructed earlier to have the property (40.24) that commutation with ε converts orthonormal basis vectors γ_a of the geometric algebra to minus their transposes. The spinor metric tensor ε has the additional property that commutation with it converts even (but not odd) orthonormal basis elements 1 and $I_3\gamma_a$ of the geometric algebra to their complex conjugates (with respect to i) in the Pauli representation (13.112). Consequently commutation with ε converts rotors R , which are real linear combinations of the even orthonormal basis elements, to their complex conjugates,

$$\varepsilon R^* = R\varepsilon, \quad (40.61)$$

which also implies that $\varepsilon R = R^*\varepsilon$, since a rotor R is a real linear combination of even orthonormal basis multivectors, so the complex conjugate R^* of a rotor R is a rotor. It follows that the conjugate Pauli spinor $\bar{\varphi}$ rotates in the same way as the spinor φ ,

$$R: \bar{\varphi} \equiv \varepsilon\varphi^* \rightarrow \varepsilon R^*\varphi^* = R\varepsilon\varphi^* = R\bar{\varphi}. \quad (40.62)$$

In components,

$$\bar{\varphi} = \varphi^{a*}\bar{\epsilon}_a, \quad \bar{\epsilon}_a \equiv \varepsilon\epsilon_a = \mp\epsilon_{\bar{a}}, \quad (40.63)$$

where the index $\bar{a}$ is the bit-flip of the index a , and the $\mp$ sign is $-$ for $\uparrow$ and $+$ for $\downarrow$, that is, $\varepsilon\epsilon_{\uparrow} = -\epsilon_{\downarrow}$ and $\varepsilon\epsilon_{\downarrow} = \epsilon_{\uparrow}$.

Conjugating a Pauli spinor φ twice changes its sign,

$$\bar{\bar{\varphi}} = -\varphi. \quad (40.64)$$

40.13 Scalar products of spinors and conjugate spinors

The row conjugate Pauli spinor $\bar{\varphi}\cdot$ corresponding to the column conjugate spinor $\bar{\varphi}$ coincides with the Hermitian conjugate spinor $\varphi^\dagger$, which in turn coincides with the reverse spinor $\tilde{\varphi}$, equation (13.129),

$$\bar{\varphi}\cdot \equiv \bar{\varphi}^\top\varepsilon = \varphi^\dagger\varepsilon^\top\varepsilon = \varphi^\dagger = \tilde{\varphi}. \quad (40.65)$$

Note that the reverse spinor $\tilde{\varphi}$ equals the row conjugate spinor $\bar{\varphi}\cdot$; the reverse spinor $\tilde{\varphi}$ does *not* equal the column conjugate spinor $\bar{\varphi}$ defined by equation (40.60), and the two should not be confused.

The scalar product of a row conjugate Pauli spinor $\bar{\varphi}\cdot$ with a column Pauli spinor χ coincides with the product of the Hermitian conjugate spinor $\varphi^\dagger$ with the spinor χ ,

$$\bar{\varphi}\cdot\chi = \varphi^\dagger\chi = \begin{pmatrix} \varphi^{\uparrow*} & \varphi^{\downarrow*} \end{pmatrix} \begin{pmatrix} \chi^\uparrow \\ \chi^\downarrow \end{pmatrix} = \varphi^{\uparrow*}\chi^\uparrow + \varphi^{\downarrow*}\chi^\downarrow. \quad (40.66)$$

In particular, the scalar product $\bar{\varphi}\cdot\varphi$ of a spinor with its own conjugate is real and positive,

$$\bar{\varphi}\cdot\varphi = \varphi^\dagger\varphi. \quad (40.67)$$

The complex conjugate of the scalar product satisfies

$$(\bar{\varphi}\cdot\chi)^* \equiv ((\varepsilon\varphi^*)^\top\varepsilon\chi)^* = \varphi^\top\varepsilon^\top\bar{\chi} = -\varphi^\top\varepsilon\bar{\chi} = -\varphi\cdot\bar{\chi}. \quad (40.68)$$

The sign flip in the fourth expression occurs because the spinor metric tensor ε is antisymmetric, $\varepsilon^\top = -\varepsilon$. In particular, the complex conjugate of the product $\bar{\varphi} \cdot \varphi$ of a spinor with its own conjugate is

$$(\bar{\varphi} \cdot \varphi)^* = -\varphi \cdot \bar{\varphi} . \quad (40.69)$$

Equation (40.69), along with the condition that the scalar product be real, $(\bar{\varphi} \cdot \varphi)^* = \bar{\varphi} \cdot \varphi$, equation (40.67), requires that the scalar product $\bar{\varphi} \cdot \varphi$ be anticommuting,

$$\bar{\varphi} \cdot \varphi = -\varphi \cdot \bar{\varphi} . \quad (40.70)$$

Equation (40.70) proves that the scalar product of Pauli spinors must be anticommuting, as asserted earlier, equation (40.41).

In nonrelativistic quantum mechanics, the real positive scalar (40.67) is interpreted as the probability of the Pauli spinor φ . Since conjugating a Pauli spinor twice flips its sign, equation (40.64), the scalar product (40.67) is the same regardless of whether the spinor φ or its conjugate $\bar{\varphi}$ is taken:

$$\bar{\bar{\varphi}} \cdot \bar{\varphi} = -\varphi \cdot \bar{\varphi} = \bar{\varphi} \cdot \varphi . \quad (40.71)$$

Concept question 40.2 Imaginary spinor metric? Would making the spinor metric ε imaginary allow the spinor scalar product to be commuting instead of anticommuting? **Answer.** No. If the spinor metric ε were multiplied by i , or more generally by some arbitrary complex phase (which is possible since the spinor metric is defined only up to a scalar normalization factor), then the conjugate spinor must be defined by $\bar{\varphi} = \varepsilon^* \varphi^*$ in place of the definition (40.60) in order that the scalar product of the spinor and its conjugate remain real and positive, equation (40.67). A manipulation similar to equation (40.68) carries through, with the result that equation (40.69) continues to hold regardless of any complex phase in the spinor metric ε . The minus sign comes from $\varepsilon^\top = -\varepsilon$ regardless of any complex phase. The scalar product of Pauli scalars is necessarily anticommuting.

40.14 Conjugate multivectors

Conjugate multivectors $\bar{\mathbf{a}}$ in the supergeometric algebra are defined, similarly to conjugate Pauli spinors, such that their components are complex conjugates of the parent multivector $\mathbf{a}$, and they rotate in the same way as multivectors (the conjugate multivector $\bar{\mathbf{a}}$ is *not* the same as the reverse multivector $\tilde{\mathbf{a}}$).

The complex conjugate multivector $\mathbf{a}^*$ of a multivector $\mathbf{a} \equiv a^A \boldsymbol{\gamma}_A$ is defined to be

$$\mathbf{a}^* \equiv a^{A*} \boldsymbol{\gamma}_A^* , \quad (40.72)$$

where $\boldsymbol{\gamma}_A^*$ is the complex conjugate of the basis multivector $\boldsymbol{\gamma}_A$ in the Pauli representation. The chiral basis vectors $\boldsymbol{\gamma}_\pm$ and $\boldsymbol{\gamma}_3$ are real in the Pauli representation, which is consistent with the basis spinors $\boldsymbol{\epsilon}_a$ being taken to be real, equation (40.58). Since $\mathbf{a}$ rotates as $\mathbf{a} \rightarrow R \mathbf{a} \tilde{R}$, the complex conjugate $\mathbf{a}^*$ rotates as

$$R : \mathbf{a}^* \rightarrow (R \mathbf{a} \tilde{R})^* = R^* \mathbf{a}^* \tilde{R}^* . \quad (40.73)$$

Complex conjugation commutes with the isomorphism between multivectors and outer products of spinors in the supergeometric algebra. That is, if the multivector is an outer product of spinors, $\mathbf{a} = \varphi\chi\cdot$, then the complex conjugate multivector is the outer product of the complex conjugate spinors, $\mathbf{a}^* = \varphi^*\chi^*\cdot$.

Similarly, consistent with the definition (40.60) of the conjugate spinor φ^* , the conjugate multivector $\bar{\mathbf{a}}$ is defined by

$$\bar{\mathbf{a}} \equiv \varepsilon \mathbf{a}^* \varepsilon^{-1} . \quad (40.74)$$

If the multivector is an outer product of spinors, $\mathbf{a} = \varphi\chi\cdot$, then the conjugate multivector is the outer product of the conjugate spinors, $\bar{\mathbf{a}} = \bar{\varphi}\bar{\chi}\cdot$. Like the conjugate spinor, equation (40.62), the conjugate multivector $\bar{\mathbf{a}}$ rotates in the same way as a multivector,

$$R : \bar{\mathbf{a}} \equiv \varepsilon \mathbf{a}^* \varepsilon^{-1} \rightarrow \varepsilon R^* \mathbf{a}^* \tilde{R}^* \varepsilon^{-1} = R \varepsilon \mathbf{a}^* \varepsilon^{-1} \tilde{R} = R \bar{\mathbf{a}} \tilde{R} . \quad (40.75)$$

In the Pauli representation, the conjugates of the orthonormal basis vectors γ_a are minus themselves,

$$\bar{\gamma}_a \equiv \varepsilon \gamma_a^* \varepsilon^{-1} = -\gamma_a . \quad (40.76)$$

The conjugate of a grade- p multivector $\mathbf{a}$ is, in components,

$$\bar{\mathbf{a}} = a^{A*} \bar{\gamma}_A , \quad \bar{\gamma}_A = (-)^p \gamma_A . \quad (40.77)$$

40.14.1 Real subalgebra

In the Pauli representation, the basis vectors $\gamma_{\pm}$ and γ_3 in a spin basis are real, equations (40.7) and (40.8), and the basis spinors $\epsilon_{\uparrow}$ are similarly real, equations (40.11). One might therefore contemplate forming a real subalgebra of the supergeometric algebra from real linear combinations of these basis spinors and their products. This does not work however, because spatial rotations transform the basis spinors into complex combinations of each other, equation (13.120). Any viable real subalgebra must be closed under rotations.

Orthonormal basis multivectors on the other hand do transform into real linear combinations of each other under rotations. A real subalgebra of the geometric algebra may be obtained by restricting to multivectors satisfying the reality condition that they are their own conjugates,

$$\bar{\mathbf{a}} = \mathbf{a} . \quad (40.78)$$

Since conjugates of even and odd orthonormal basis vectors γ_A are respectively plus and are minus themselves, equation (40.77), in the Pauli representation there is a real subalgebra consisting of linear combinations $a^A \gamma_A$ of odd orthonormal multivectors with pure imaginary coefficients, and even orthonormal multivectors with pure real coefficients. But in the Pauli algebra the (odd) pseudoscalar I_3 is identified with i times the unit matrix, so the real Pauli subalgebra reduces to real linear combinations of even orthonormal multivectors.

40.15 The supergeometric algebra in arbitrarily many spatial dimensions

Exercise 40.3 Generalize the supergeometric algebra to an arbitrary number of dimensions. Generalize the supergeometric algebra to an arbitrary number of spatial dimensions N . Exercise 41.5 generalizes this exercise to an arbitrary number of space and time dimensions.

Solution.

1. **Basis of spin vectors.** Let γ_a , $a = 1, \dots, N$ be an orthonormal ($\gamma_a \cdot \gamma_b = \delta_{ab}$) basis of vectors in the N -dimensional geometric algebra. The geometric algebra, Chapter 13, generated by inner and outer products of the N basis vectors γ_a is a vector space of dimension 2^N .

In the presence of spinors, it is convenient to group the orthonormal basis elements into pairs $\gamma_i^\pm$,

$$\gamma_i^+ \equiv \gamma_{2i-1}, \quad \gamma_i^- \equiv \gamma_{2i}, \quad i = 1, \dots, [N/2]. \quad (40.79)$$

If the dimension N is odd, then one basis vector, γ_N , will remain unpaired. Define chiral basis vectors γ_i and $\gamma_{\bar{i}}$ by

$$\gamma_i \equiv \frac{1}{\sqrt{2}}(\gamma_i^+ + i\gamma_i^-), \quad \gamma_{\bar{i}} \equiv \frac{1}{\sqrt{2}}(\gamma_i^+ - i\gamma_i^-), \quad i = 1, \dots, [N/2], \quad (40.80)$$

generalizing equations (40.1). The normalization is $\gamma_i \gamma_{\bar{i}} = 1$. The orthonormal vectors γ_i^+ and γ_i^- can be thought of as, modulo a normalization, the real and imaginary parts of a complex vector γ_i whose complex conjugate is $\gamma_{\bar{i}}$. Under a right-handed rotation by angle θ in the $\gamma_i^+ \gamma_i^-$ -plane, the i th pair of chiral basis vectors transform as

$$\gamma_i \rightarrow e^{-i\theta} \gamma_i, \quad \gamma_{\bar{i}} \rightarrow e^{+i\theta} \gamma_{\bar{i}}. \quad (40.81)$$

The transformation (40.81) identifies the chiral basis vectors γ_i and $\gamma_{\bar{i}}$ as having i th spin weight respectively equal to $+1$ and -1 . All other chiral basis vectors, γ_j and $\gamma_{\bar{j}}$ with $j \neq i$, together with the unpaired basis vector γ_N if N is odd, have zero i th spin weight. There are $[N/2]$ different spin weights i . The components of a tensor in a spin basis inherit their spin properties from those of the spin basis. The i th spin-weight s_i of any tensor component is

$$\text{spin weight } s_i = \text{number of } i \text{ minus } \bar{i} \text{ covariant indices}, \quad (40.82)$$

generalizing equation (40.6).

2. **Basis of spinors ϵ_a .** Spinor axes are defined by $2^{[N/2]}$ basis spinors ϵ_a ,

$$\epsilon_a \equiv \epsilon_{a_1 \dots a_{[N/2]}} \quad (40.83)$$

where $a_1 \dots a_{[N/2]}$ denotes not a set of indices, but rather a bitcode specifying the single index a . Each bit a_i is either up $\uparrow$ or down $\downarrow$. For example, one of the basis spinors is the all-up basis spinor $\epsilon_{\uparrow \dots \uparrow}$.

Under a right-handed rotation by angle θ in the $\gamma_i^+ \gamma_i^-$ -plane, a basis spinor ϵ_a transforms as

$$\epsilon_{\dots i \dots} \rightarrow e^{-i\theta/2} \epsilon_{\dots i \dots}, \quad \epsilon_{\dots \bar{i} \dots} \rightarrow e^{i\theta/2} \epsilon_{\dots \bar{i} \dots}. \quad (40.84)$$

The transformation (40.84) shows that each basis spinor ϵ_a has i th spin weight either $+\frac{1}{2}$ or $-\frac{1}{2}$ in each

of its $[N/2]$ bits. The components of a spinor tensor in a spin basis inherit their spin properties from those of the spin basis. The i th spin-weight s_i of any spinor tensor component is

$$\text{spin weight } s_i = \frac{1}{2}(\text{number of } i \text{ minus } \bar{i} \text{ covariant indices}) , \quad (40.85)$$

generalizing equation (40.13).

A spinor φ ,

$$\varphi = \varphi^a \epsilon_a , \quad (40.86)$$

is a linear combination of the $2^{[N/2]}$ basis spinors ϵ_a . The spinor can be represented as a column vector φ^a of dimension $2^{[N/2]}$, the index a running over bitcodes $a_1 \dots a_{[N/2]}$.

3. **Spinor metric tensor.** A spinor metric ε can be defined as that spinor tensor that is invariant under rotations, suitably normalized, §40.6. Invariance of the spinor metric ε under rotations requires that for any rotor R ,

$$\varepsilon R^\top = \tilde{R} \varepsilon , \quad (40.87)$$

the same as condition (40.17). A rotor R is a real linear combination of even elements of the geometric algebra in an orthonormal basis. Thus the condition (40.87) is determined by the commutation properties of ε with the orthonormal bivectors of the geometric algebra (an orthonormal bivector is defined here to be a wedge product of orthonormal vectors; the square of an orthonormal bivector is thus -1). In the canonical chiral representation defined by the construction (40.108), orthonormal basis bivectors $\gamma_a \wedge \gamma_b$ are represented by traceless, unitary ($A^{-1} = A^\dagger$), skew-Hermitian ($A^\dagger = -A$) matrices. Then condition (40.87) holds if ε commutes with orthonormal basis bivectors whose representation is real, and anticommutes with orthonormal basis bivectors whose representation is imaginary. In the construction (40.108), all chiral basis vectors γ_i and $\gamma_{\bar{i}}$ are real, so orthonormal basis vectors γ_i^+ are real while $\gamma_{\bar{i}}^-$ are imaginary. The only matrix ε with the required commutation properties with basis bivectors is, up to a scalar or pseudoscalar normalization factor, the product of all the odd basis vectors γ_i^+ ,

$$\varepsilon = \prod_{i=1}^{[(N+1)/2]} \gamma_i^+ . \quad (40.88)$$

An alternative version ε_{alt} of the spinor metric may be obtained by multiplying the spinor metric (40.88) by the chiral factor $\varkappa_N$, which is the pseudoscalar I_N , equation (40.121), normalized by a power of i so that $\varkappa_N^2$ equals one, equation (40.124),

$$\varepsilon_{\text{alt}} \equiv \varkappa_N \varepsilon = \prod_{i=1}^{[N/2]} i \gamma_i^- . \quad (40.89)$$

The factors of the imaginary i are introduced so that the spinor metric ε is real.

If N is odd, and if the odd algebra is constructed, as described in part 10 of this Exercise, by embedding the odd algebra in one extra dimension and treating either the final (odd) dimension γ_N or the extra (even) dimension γ_{N+1} as a scalar, then there are further options for the spinor metric. The

Table 40.1: Symmetry of spinor metric

$N \pmod{8}$	$\varepsilon^2 = (-)^{[(N+1)/4]}$	$\varepsilon_{\text{alt}}^2 = (-)^{[(N+2)/4]}$	$\tilde{\varepsilon}^2 = (-)^{[N/4]}$	$\tilde{\varepsilon}_{\text{alt}}^2 = (-)^{[(N+3)/4]}$
1	+	+	+	−
2	+	−		
3	−	−	+	−
4	−	−		
5	−	−	−	+
6	−	+		
7	+	+	−	+
8	+	+		

invariance condition (40.87) need hold only for rotors not involving the scalar dimension γ_N or γ_{N+1} . If the scalar dimension is the odd dimension γ_N , then γ_N can be dropped from the standard spinor metric ε , leaving ε in $N-1$ dimensions. If the scalar dimension is the even dimension γ_{N+1} , then $i\gamma_{N+1}$ can be adjoined to the alternative spinor metric ε_{alt} , giving ε_{alt} in $N+1$ dimensions. The resulting spinor metrics, distinguished with a tilde, are

$$\tilde{\varepsilon}_N = \varepsilon_N \gamma_N = \varepsilon_{N-1} , \quad \tilde{\varepsilon}_{\text{alt},N} = \varepsilon_{\text{alt},N} i\gamma_{N+1} = \varepsilon_{\text{alt},N+1} \quad (N \text{ odd}) . \quad (40.90)$$

The spinor metric ε , in any of the forms (40.88)–(40.90), is real and orthogonal, and its square is plus or minus the unit matrix,

$$\varepsilon^{-1} = \varepsilon^\top , \quad \varepsilon^2 = \pm 1 , \quad (40.91)$$

where the $\pm$ sign is as tabulated in Table 40.1. The square of the spinor metric coincides with the symmetry of the spinor metric under exchange of its indices, equation (40.96) below. The spinor metric matrix ε is always Hermitian,

$$\varepsilon^{-1} = \varepsilon^\dagger . \quad (40.92)$$

Despite the equality of ε and $\prod_i \gamma_i^+$ (or of ε_{alt} and $\prod_i i\gamma_i^-$) in the representation (40.108), ε (or ε_{alt}) is defined to transform as a spinor tensor under rotations, not as an element of the geometric algebra. In the representation (40.108), the ordering of rows or columns indexed by spinor index $a = a_1 \dots a_{[N/2]}$ is that of binary numbers $a_{[N/2]} \dots a_1$ with 0 for up $\uparrow$ and 1 for down $\downarrow$. The components ε_{ba} of the spinor metric ε ,

$$\varepsilon_{ba} \equiv \epsilon_b \cdot \epsilon_a \equiv \epsilon_b^\top \varepsilon \epsilon_a , \quad (40.93)$$

are non-vanishing only between basis spinors ϵ_b and ϵ_a that are bit flips of each other. The sign of $\varepsilon_{\bar{a}a}$, where $\bar{a}$ denotes the bit flip of a , follows inductively from equations (40.106), and is

$$\varepsilon \epsilon_a = \text{sign}(\varepsilon_{\bar{a}a}) \epsilon_{\bar{a}} , \quad \text{sign}(\varepsilon_{\bar{a}a}) \equiv \text{sign}(\varepsilon_{\bar{a}_1 \dots \bar{a}_{[N/2]} a_1 \dots a_{[N/2]}}) = \prod_{a_i = \uparrow} (-)^{i-1} . \quad (40.94)$$

Table 40.2: Sign of $\gamma_a^\top \varepsilon = \pm \varepsilon \gamma_a$

$N \pmod{8}$	$\varepsilon : (-)^{[(N+3)/2]}$	$\varepsilon_{\text{alt}} : (-)^{[N/2]}$	$\tilde{\varepsilon} : (-)^{[(N+2)/2]}$	$\tilde{\varepsilon}_{\text{alt}} : (-)^{[(N+1)/2]}$
1	+	+	−	−
2	+	−		
3	−	−	+	+
4	−	+		
5	+	+	−	−
6	+	−		
7	−	−	+	+
8	−	+		

For the alternative spinor metric (40.89), the sign is

$$\varepsilon_{\text{alt}} \epsilon_a = \text{sign}(\varepsilon_{\bar{a}a}^{\text{alt}}) \epsilon_{\bar{a}} , \quad \text{sign}(\varepsilon_{\bar{a}a}^{\text{alt}}) \equiv \text{sign}(\varepsilon_{\bar{a}_1 \dots \bar{a}_{[N/2]} a_1 \dots a_{[N/2]}}^{\text{alt}}) = \prod_{a_i = \uparrow} (-)^i . \quad (40.95)$$

The spinor metric is symmetric or antisymmetric as its square is positive or negative,

$$\varepsilon_{ab} = \pm \varepsilon_{ba} , \quad (40.96)$$

where the $\pm$ sign is as tabulated in Table 40.1.

Commuting the spinor metric ε through the orthonormal basis vectors γ_a converts them to plus or minus their transposes,

$$\gamma_a^\top \varepsilon = \pm \varepsilon \gamma_a . \quad (40.97)$$

Table 40.2 tabulates the sign in equation (40.97) for the spinor metric ε and the alternative spinor metric ε_{alt} , along with the tilde'd versions (40.90) for odd N . For tilde'd spinor metrics, equation (40.97) holds for all orthonormal basis vectors γ_a *excepting* the scalar vector γ_N or γ_{N+1} , for which there is an extra minus sign, that is, $\gamma_N^\top \tilde{\varepsilon} = -\pm \tilde{\varepsilon} \gamma_N$ if the scalar dimension is γ_N , or $\gamma_{N+1}^\top \tilde{\varepsilon}_{\text{alt}} = -\pm \tilde{\varepsilon}_{\text{alt}} \gamma_{N+1}$ if the scalar dimension is γ_{N+1} . Equation (40.97) is proved by induction: equations (40.109) and (40.115) imply that if (40.97) has a certain sign in $N-2$ dimensions, then it has the same sign in N dimensions; the sign is then determined at the smallest dimension for which the spinor metric ε is defined, $N = 1$ or 2 .

Equation (40.97) implies that the commutation rule of an orthonormal multivector γ_A of grade p with the spinor metric ε is

$$\gamma_A^\top \varepsilon = (\pm)^p \varepsilon \tilde{\gamma}_A = (\pm)^p (-)^{[p/2]} \varepsilon \gamma_A , \quad (40.98)$$

where $\tilde{\gamma}_A$ is the reverse (not conjugate) of γ_A , and the $\pm$ sign in $(\pm)^p$ is that in equation (40.97), which depends on dimension N as tabulated in Table 40.2.

4. **Scalar product of spinors.** Corresponding to any column basis spinor ϵ_a is a row basis spinor ϵ_a .

defined by

$$\epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon . \quad (40.99)$$

(or by $\epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon_{\text{alt}}$ if the alternative spinor metric is used). The row spinor $\varphi \cdot$ corresponding to a column spinor $\varphi = \varphi^a \epsilon_a$ is

$$\varphi \cdot \equiv \varphi^\top \varepsilon = \varphi^a \epsilon_a \cdot . \quad (40.100)$$

The scalar product of row and column spinors is

$$\varphi \cdot \chi = \varepsilon_{ab} \varphi^a \chi^b . \quad (40.101)$$

The scalar product is symmetric or antisymmetric as the spinor metric is symmetric or antisymmetric,

$$\varphi \cdot \chi = \varepsilon^2 \chi \cdot \varphi , \quad (40.102)$$

the sign of ε^2 being as given in Table 40.1.

Linear combinations of outer products $\epsilon_a \epsilon_b \cdot$ of basis spinors,

$$\varphi \chi \cdot = \varphi^a \chi^b \epsilon_a \epsilon_b \cdot , \quad (40.103)$$

form a vector space of dimension $2^{2[N/2]}$. Multiplication of outer products satisfies the associative rule

$$(\varphi \chi \cdot)(\psi \xi \cdot) = \varphi(\chi \cdot \psi) \xi \cdot , \quad (40.104)$$

which since $\chi \cdot \psi$ is a scalar is proportional to the outer product $\varphi \xi \cdot$.

5. **Chiral representation of the supergeometric algebra.** There is an isomorphism between the algebra of outer products of spinors and the geometric algebra (Brauer and Weyl, 1935). The isomorphism may be established by an explicit representation in terms of column and row vectors for spinors, and matrices for multivectors in the geometric algebra. This part 5 of this Exercise takes the spinor metric to be the standard spinor metric ε , equation (40.88). The next part 6 of this Exercise describes the modifications that must be made if the spinor metric is taken to be the alternative spinor metric ε_{alt} , equation (40.89). Since $[N/2]$, the integer part of half the number of dimensions N , appears so ubiquitously, it is convenient to define

$$n \equiv [N/2] . \quad (40.105)$$

The construction below yields the chiral representation, generated inductively starting from $N = 0$. Given a representation of column and row basis spinors ϵ_A and $\epsilon_A \cdot$ in $N-2$ dimensions, a representation of column and row basis spinors ϵ_{Aa} and $\epsilon_{Aa} \cdot$ (with one extra index $a = \uparrow$ or $\downarrow$) in N dimensions are column and row matrices of length 2^n ,

$$\epsilon_{A\uparrow} = \begin{pmatrix} \epsilon_A \\ 0 \end{pmatrix} , \quad \epsilon_{A\uparrow} \cdot = \begin{pmatrix} 0 & \epsilon_A \cdot \end{pmatrix} , \quad (40.106a)$$

$$\epsilon_{A\downarrow} = \begin{pmatrix} 0 \\ \epsilon_A \end{pmatrix} , \quad \epsilon_{A\downarrow} \cdot = \begin{pmatrix} (-)^{n-1} \epsilon_A \cdot & 0 \end{pmatrix} , \quad (40.106b)$$

where 0 represents respectively a zero column or row vector of length 2^{n-1} . The induction starts at $N = 2$ where A is empty and $\epsilon_A = \epsilon_A \cdot = 1$. The trailing dot signifies the spinor metric tensor ε . The construction (40.106) assumes that the spinor metric ε is a product (40.88) of factors, the last factor γ_n^+ taking the form (40.113), so that the relation between the spinor metric in N and $N-2$ dimensions is given by equation (40.115).

The outer products of the column basis spinors ϵ_{Aa} and row basis spinors ϵ_{Bb} given by the inductive relations (40.106) are $2^n \times 2^n$ matrices

$$\epsilon_{A\uparrow}\epsilon_{B\uparrow}\cdot = \begin{pmatrix} 0 & \epsilon_A\epsilon_B\cdot \\ 0 & 0 \end{pmatrix}, \quad (40.107a)$$

$$\epsilon_{A\uparrow}\epsilon_{B\downarrow}\cdot = \begin{pmatrix} (-)^{n-1}\epsilon_A\epsilon_B\cdot & 0 \\ 0 & 0 \end{pmatrix}, \quad (40.107b)$$

$$\epsilon_{A\downarrow}\epsilon_{B\uparrow}\cdot = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon_A\epsilon_B\cdot \end{pmatrix}, \quad (40.107c)$$

$$\epsilon_{A\downarrow}\epsilon_{B\downarrow}\cdot = \begin{pmatrix} 0 & 0 \\ (-)^{n-1}\epsilon_A\epsilon_B\cdot & 0 \end{pmatrix}, \quad (40.107d)$$

where the 0's in equations (40.107) represent zero $2^{n-1} \times 2^{n-1}$ matrices. Again, the induction (40.107) starts at $N = 2$ where A and B are empty, and $\epsilon_A\epsilon_B\cdot = 1$. The outer products (40.107) can be rewritten

$$\epsilon_{A\uparrow}\epsilon_{B\uparrow}\cdot = \frac{1}{\sqrt{2}} \begin{pmatrix} \epsilon_A\epsilon_B\cdot & 0 \\ 0 & \pm\epsilon_A\epsilon_B\cdot \end{pmatrix} \begin{pmatrix} 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad (40.108a)$$

$$\epsilon_{A\uparrow}\epsilon_{B\downarrow}\cdot = (-)^{n-1} \frac{1}{2} \begin{pmatrix} \epsilon_A\epsilon_B\cdot & 0 \\ 0 & \pm\epsilon_A\epsilon_B\cdot \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \quad (40.108b)$$

$$\epsilon_{A\downarrow}\epsilon_{B\uparrow}\cdot = \pm \frac{1}{2} \begin{pmatrix} \epsilon_A\epsilon_B\cdot & 0 \\ 0 & \pm\epsilon_A\epsilon_B\cdot \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}, \quad (40.108c)$$

$$\epsilon_{A\downarrow}\epsilon_{B\downarrow}\cdot = \pm(-)^{n-1} \frac{1}{\sqrt{2}} \begin{pmatrix} \epsilon_A\epsilon_B\cdot & 0 \\ 0 & \pm\epsilon_A\epsilon_B\cdot \end{pmatrix} \begin{pmatrix} 0 & 0 \\ \sqrt{2} & 0 \end{pmatrix}, \quad (40.108d)$$

where the upper/lower sign is for even/odd $\epsilon_A\epsilon_B\cdot$ (that is, the total spin weight $\sum_i s_i$ of $\epsilon_A\epsilon_B\cdot$ is even/odd). The first matrix on the right hand sides of equations (40.108) is the matrix representation of the multivector $\epsilon_A\epsilon_B\cdot$ in N dimensions in terms of its representation in $N-2$ dimensions,

$$\epsilon_A\epsilon_B\cdot = (-)^{n-1}\epsilon_{A\uparrow}\epsilon_{B\downarrow}\cdot \pm \epsilon_{A\downarrow}\epsilon_{B\uparrow}\cdot = \begin{pmatrix} \epsilon_A\epsilon_B\cdot & 0 \\ 0 & \pm\epsilon_A\epsilon_B\cdot \end{pmatrix}. \quad (40.109)$$

The rightmost factors in equations (40.108) constitute the matrix representations of γ_n , $\gamma_n\gamma_{\bar{n}}$, $\gamma_{\bar{n}}\gamma_n$, and $\gamma_{\bar{n}}$ in N dimensions,

$$\gamma_n = \begin{pmatrix} 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad \gamma_n\gamma_{\bar{n}} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \quad \gamma_{\bar{n}}\gamma_n = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}, \quad \gamma_{\bar{n}} = \begin{pmatrix} 0 & 0 \\ \sqrt{2} & 0 \end{pmatrix}, \quad (40.110)$$

which have the correct normalization and commutation rules with respect to each other. The signs in equations (40.108) are arranged so that the correct commutation rules of the geometric algebra are respected: γ_n and $\gamma_{\bar{n}}$, which are odd, commute/anticommute with $\epsilon_A \epsilon_B \cdot$ according as the latter is even/odd; and $\gamma_n \gamma_{\bar{n}}$ and $\gamma_{\bar{n}} \gamma_n$, which are even, always commute with $\epsilon_A \epsilon_B \cdot$. In terms of scalar and wedge products, the multivectors $\gamma_n \gamma_{\bar{n}}$ and $\gamma_{\bar{n}} \gamma_n$ in equations (40.110) are

$$\gamma_n \gamma_{\bar{n}} = \gamma_n \cdot \gamma_{\bar{n}} + \gamma_n \wedge \gamma_{\bar{n}}, \quad \gamma_{\bar{n}} \gamma_n = \gamma_n \cdot \gamma_{\bar{n}} - \gamma_n \wedge \gamma_{\bar{n}}, \quad \gamma_n \cdot \gamma_{\bar{n}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \gamma_n \wedge \gamma_{\bar{n}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (40.111)$$

Note that the chiral and orthonormal bivectors are related by

$$\gamma_n \wedge \gamma_{\bar{n}} = -i \gamma_n^+ \wedge \gamma_{\bar{n}}^-, \quad (40.112)$$

so that $(\gamma_n \wedge \gamma_{\bar{n}})^2 = 1$. The orthonormal basis vectors γ_n^+ and $\gamma_{\bar{n}}^-$ at the n th step are

$$\gamma_n^+ = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma_{\bar{n}}^- = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad (40.113)$$

which are traceless, unitary, and Hermitian. The orthonormal basis bivector $\gamma_n^+ \wedge \gamma_{\bar{n}}^-$ is

$$\gamma_n^+ \wedge \gamma_{\bar{n}}^- = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad (40.114)$$

which is traceless, unitary, and skew-Hermitian. An iterative expression for the spinor metric ε_N follows from its expression (40.88) as a product of basis vectors, and is the antidiagonal matrix

$$\varepsilon_N = \varepsilon_{N-2} \gamma_n^+ = \begin{pmatrix} \varepsilon_{N-2} & 0 \\ 0 & (-)^{n-1} \varepsilon_{N-2} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon_{N-2} \\ (-)^{n-1} \varepsilon_{N-2} & 0 \end{pmatrix}. \quad (40.115)$$

The left factor in the third expression of equations (40.115) is the matrix representation of ε_{N-2} in N dimensions in terms of its representation in $N-2$ dimensions, in accordance with equation (40.109). The factor of $(-)^{n-1}$ comes from the fact that the spinor metric ε_{N-2} is a product of $n-1$ basis vectors, equation (40.88), so is even/odd (total spin weight even/odd) as $n-1$ is even or odd. Equation (40.115), which was assumed in the initial step (40.106) of the construction of the chiral representation of the supergeometric algebra, proves the consistency of the construction.

The matrix representation of the column and row basis spinors (40.106) and of their outer products (40.108) is entirely real (with respect to i). The expansion of the 2^N outer products $\epsilon_a \epsilon_b \cdot$ of spinors in terms of the 2^N basis multivectors γ_A of the spacetime algebra, and vice versa, define the matrix of coefficients γ_{ab}^A and its inverse γ_A^{ab} ,

$$\epsilon_a \epsilon_b \cdot = \gamma_{ab}^A \gamma_A, \quad \gamma_A = \gamma_A^{ab} \epsilon_a \epsilon_b \cdot. \quad (40.116)$$

The coefficients γ_{ab}^A and γ_A^{ab} in the chiral representation are

$$\gamma_{ab}^A = \frac{1}{2^n} \epsilon_b \cdot \gamma^A \epsilon_a, \quad \gamma_A^{ab} = \text{sign}(\varepsilon^2) \epsilon^a \cdot \gamma_A \epsilon^b, \quad (40.117)$$

where $\text{sign}(\varepsilon^2)$ is the symmetry of the spinor metric, Table 40.1.

For even N , the above construction establishes an isomorphism between outer products of spinors and the geometric algebra,

$$\text{outer products of spinors} \cong \text{geometric algebra} \quad (N \text{ even}) . \quad (40.118)$$

Both spaces are complex 2^N -dimensional vector spaces. Their representation as $2^n \times 2^n$ dimensional matrices is minimal: there is no representation of the geometric algebra with matrices of smaller dimension.

6. Chiral representation of the supergeometric algebra using the alternative spinor metric.

The chiral representation of the supergeometric algebra with the alternative spinor metric (40.89) is the same as the construction in part 5, but with the replacement

$$(-)^{n-1} \rightarrow (-)^n \quad (40.119)$$

in equations (40.106) to (40.109). Analogously to equation (40.115), an iterative equation for the alternative spinor metric follows from its expression (40.89) as a product of basis vectors, and is the antidiagonal matrix

$$\varepsilon_N^{\text{alt}} = \varepsilon_{N-2}^{\text{alt}} i \gamma_n^- = \begin{pmatrix} \varepsilon_{N-2}^{\text{alt}} & 0 \\ 0 & (-)^{n-1} \varepsilon_{N-2}^{\text{alt}} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon_{N-2}^{\text{alt}} \\ (-)^n \varepsilon_{N-2}^{\text{alt}} & 0 \end{pmatrix} . \quad (40.120)$$

7. Supergeometric algebra in odd dimensions, version 1. The construction of the supergeometric algebra in part 5 works in even dimensions N . What about N odd? One approach, dealt with in this part, is to project the odd-dimensional algebra into one lower dimension, which requires identifying the chiral operator $\varkappa_N$ with 1, equation (40.125). The resulting algebra of outer products of spinors, besides not yielding the full odd- N geometric algebra, does not include a parity operator. A richer approach, put forward in part 10, is to embed the odd-dimensional algebra in the algebra with one higher dimension, and to treat the extra dimension as a scalar, which proves to be a parity operator.

Consider that the pseudoscalar I_N of the geometric algebra can be written

$$I_N \equiv \gamma_1^+ \wedge \gamma_1^- \wedge \dots \wedge \gamma_n^+ \wedge \gamma_n^- \{ \wedge \gamma_N \text{ if } N \text{ is odd} \} = i^n \varkappa_N , \quad (40.121)$$

where the chiral operator $\varkappa_N$ (the generalization of the 4D Dirac chiral operator γ_5) is defined by

$$\varkappa_N \equiv \gamma_1 \wedge \gamma_{\bar{1}} \wedge \dots \wedge \gamma_n \wedge \gamma_{\bar{n}} \{ \wedge \gamma_N \text{ if } N \text{ is odd} \} . \quad (40.122)$$

In the chiral representation (40.108), the representation of the chiral operator $\varkappa_N$ in N even dimensions in terms of its representation $\varkappa_{N-2}$ in $N-2$ dimensions is the diagonal matrix

$$\varkappa_N = \begin{pmatrix} \varkappa_{N-2} & 0 \\ 0 & -\varkappa_{N-2} \end{pmatrix} \quad (N \text{ even}) . \quad (40.123)$$

The chiral operator is diagonal in the chiral representation by construction. The square of the pseudoscalar is $I_N^2 = (-)^n$, equation (13.21), so the square of the chiral operator is the unit matrix 1,

$$\varkappa_N^2 = 1 . \quad (40.124)$$

Like the pseudoscalar I_N , the chiral operator $\varkappa_N$ is invariant under rotations. For even N , the chiral operator $\varkappa_N$ is defined through equation (40.122) as a prescribed member of both algebras, the algebra of spinor outer products and the geometric algebra. But for odd N , since the definition (40.122) involves γ_N which (as yet) has no expression in the algebra of outer products of spinors, there is the possibility that $\varkappa_N$ could be a distinct element not belonging to the algebra of spinor outer products. The element $\varkappa_N$ is a rotationally invariant scalar that squares to 1, and that (for odd N) commutes with all basis vectors γ_a . The other element of the odd- N algebra of spinor outer products that possesses those properties is (up to a possible sign) the unit element. Thus if the chiral operator $\varkappa_N$ is identified with 1,

$$\varkappa_N = 1 \quad (N \text{ odd}) , \quad (40.125)$$

then there is an isomorphism between the algebra of outer products of spinors in $N-1$ dimensions and the geometric algebra in N dimensions modulo the chiral operator $\varkappa_N$,

$$\text{outer products of spinors} \cong \text{geometric algebra (mod } \varkappa_N) \quad (N \text{ odd}) . \quad (40.126)$$

Given the identification (40.125) of the chiral operator with 1, it then follows from the definition equation (40.122) of $\varkappa_N$ that the final element γ_N of the geometric algebra is

$$\gamma_N = \varkappa_{N-1} = \gamma_1 \wedge \gamma_{\bar{1}} \wedge \dots \wedge \gamma_n \wedge \gamma_{\bar{n}} \quad (N \text{ odd}) . \quad (40.127)$$

In the case $N = 3$, this gives

$$\gamma_3 = \gamma_1 \wedge \gamma_{\bar{1}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} , \quad (40.128)$$

in agreement with the Pauli matrix equation (40.8). With the identification (40.125), the pseudoscalar I_N itself is, equation (40.121),

$$I_N = i^n \quad (N \text{ odd}) . \quad (40.129)$$

For odd N , the chiral operator $\varkappa_N$ defined by equation (40.122) is (before $\varkappa_N$ is identified with 1) an odd element of the geometric algebra. Thus for odd N , the odd part of the geometric algebra is isomorphic to $\varkappa_N$ times the even geometric algebra. Only the odd geometric algebra is affected by the identification (40.125) of the chiral operator with unity; the even geometric algebra is unaffected. The square of the chiral operator is always 1, equation (40.124), so the product of two odd multivectors yields the correct even multivector regardless of the identification (40.125).

In summary, the algebra of spinor outer products in $2n$ dimensions is isomorphic to the geometric algebra for both even and odd N , modulo $\varkappa_N$ in the case of odd N . The algebra is a complex (with respect to i) vector space of dimension 2^{2n} , represented in the chiral construction (40.108) by $2^n \times 2^n$ matrices. For example, the $N = 2$ geometric algebra is the algebra generated by $1, \gamma_1, \gamma_{\bar{1}}, \gamma_1 \wedge \gamma_{\bar{1}}$, while the $N = 3$ geometric algebra (the Pauli algebra) is the algebra generated by $1, \gamma_1, \gamma_{\bar{1}}, \gamma_3$, the pseudoscalar I_3 being identified with the imaginary i .

8. **Extra symmetry of the supergeometric algebra in odd dimensions.** Given that, if $\varkappa_N$ is identified with 1, the geometric algebra for odd N is isomorphic to the geometric algebra for even $N-1$,

what is the difference between the two algebras? Since the algebras are isomorphic, there is of course no difference. However, bivectors are special in that they are the only generators that generate transformations that preserve grade, and therefore correspond to what one usually thinks of as spatial rotations. If one restricts only to rotations generated by bivectors, then the odd algebra has a higher degree of symmetry. The equivalence (40.127) means that the pseudoscalar $\varkappa_{N-1}$ in the even algebra is promoted to a vector γ_N in the odd algebra, and pseudovectors $\gamma_i^\pm \varkappa_{N-1}$ in the even algebra become bivectors $\gamma_i^\pm \wedge \gamma_N$ in the odd algebra. Thus the odd algebra has $N-1$ more rotations than the even algebra.

The final basis vector $\gamma_N = \varkappa_{N-1}$ of the odd algebra has the same properties as the other orthonormal basis vectors $\gamma_i^\pm$: its square is 1, it anticommutes with the other orthonormal basis vectors, it is represented by a traceless, unitary, Hermitian matrix, and its reverse is (by definition) itself, $\tilde{\gamma}_N = \gamma_N$. And, like the other orthonormal basis vectors γ_i^+ of odd index, the representation of γ_N is real.

The Pauli algebra (13.115) in $N = 3$ dimensions offers a familiar example. In both 2 and 3 dimensions there are just 2 basis spinors, $\epsilon_\uparrow$ and $\epsilon_\downarrow$, which one commonly conceptualizes as being up and down along a “3-axis”. But whereas in 2 dimensions there is just one rotation, generated by the bivector $\gamma_1^+ \wedge \gamma_1^-$ (rotation about the “3-axis”), in 3 dimensions there are 2 more rotations, generated by the bivectors $\gamma_1^+ \wedge \gamma_3$ and $\gamma_3 \wedge \gamma_1^-$ (rotations about the “2-axis” and “1-axis”).

9. **Parity reversal.** A second approach to the odd- N algebra is put forward in the next part 10, but first it is necessary to consider the issue of parity reversal. Parity reversal is the operation of reflecting an odd number of spatial axes γ_a , corresponding to an improper rotation with determinant -1 . By contrast, reflecting an even number of axes can be accomplished by a continuous rotation with determinant 1.

If the number N of dimensions is even, then parity reversal may be realised by picking one particular axis, say $P = \gamma_a$, and transforming spinors ψ and multivectors $\mathbf{a}$ by

$$P: \quad \psi \rightarrow P\psi, \quad \mathbf{a} \rightarrow P\mathbf{a}P^{-1}. \quad (40.130)$$

The transformation (40.130) reflects all axes *except* the axis $P = \gamma_a$, so reflects an odd number of axes provided that N is even.

If the number N of dimensions is odd, and if the geometric algebra is projected into one dimension lower as proposed in part 7, equation (40.125), then there is no element of the geometric algebra that accomplishes parity reversal P by the operation (40.130). The difficulty is that any anticommutation of P with a basis vector γ_a is cancelled by a corresponding anticommutation with the final basis vector γ_N , equation (40.127), for no net anticommutation. The absence of a parity operator in the geometric algebra holds true even if the odd-dimensional chiral operator $\varkappa_N$ is *not* identified with unity, since all vectors commute with the odd-dimensional chiral operator. The problem of constructing an odd- N supergeometric algebra that incorporates a parity operator is solved in the next part 10.

10. **Supergeometric algebra in odd dimensions, version 2.** The previous part 9 brought up the fact that the geometric algebra in odd N dimensions does not contain a parity operator P , at least if the path proposed in part 7 is followed, that is, if the odd- N algebra is projected into one lower dimension.

The problem is not that the operation of parity reversal does not exist, but rather, how to construct such a parity operator out of products of spinors.

The solution is to embed the odd N -dimensional algebra in the even $(N+1)$ -dimensional algebra, and

to treat either the final (odd) dimension γ_N or the extra (even) orthonormal dimension γ_{N+1} as the scalar parity operator P ,

$$P = \gamma_N \text{ or } \gamma_{N+1} . \quad (40.131)$$

The vectors γ_N or γ_{N+1} have the usual property that they anticommute with all orthonormal vectors γ_a other than themselves, so the parity operator P defined by equation (40.131) has the property that it reflects all axes except itself,

$$P : \gamma_a \rightarrow P\gamma_a P^{-1} = -\gamma_a . \quad (40.132)$$

Since N is odd, this choice of P reflects an odd number of axes, so indeed reverses parity. The operation P of reflecting all axes (other than the scalar axis γ_N or γ_{N+1}) is rotationally invariant with respect to rotations in N dimensions (with the scalar axis γ_N or γ_{N+1} fixed).

As usual, there is a spin bit (the $[(N+1)/2]$ th bit) associated with the pair γ_N and γ_{N+1} of axes. Normally a rotation in the $\gamma_N \wedge \gamma_{N+1}$ plane would rotate spinors by a phase $e^{\mp i\theta/2}$ with sign $\mp$ depending on whether the spin bit is up $\uparrow$ or down $\downarrow$. But since P is a scalar, there is no such rotation. Notwithstanding the absence of a rotation by a phase, the spin bit is still there, part of the bitcode index $a = a_1 \dots a_{[(N+1)/2]}$ of a basis spinor ϵ_a .

11. **Properties of orthonormal basis multivectors in the chiral representation.** In the chiral representation constructed in part 5, all orthonormal basis vectors γ_a , and all orthonormal basis p -vectors $\gamma_{a_1 \dots a_p} \equiv \gamma_{a_1} \wedge \dots \wedge \gamma_{a_p}$, are traceless (except for the unit basis element 1), unitary, and either Hermitian (if $[p/2]$ is even, i.e. $p = 0, 1, 4, 5, \dots$) or skew-Hermitian (if $[p/2]$ is odd, i.e. $p = 2, 3, 6, 7, \dots$) $2^n \times 2^n$ matrices. All matrices have determinant 1, except that for $N = 2$ the vectors (grade $p = 1$) have determinant -1 . The unit element is represented by the unit matrix. Most of these assertions can be proved by induction using the expression (40.109), which gives the representation of a multivector in N dimensions in terms of its representation in $N-2$ dimensions.
12. **Right- and left-handed chiral subalgebras in even dimensions.** In even N dimensions, a spinor is said to be right- or left-handed depending on whether its chirality is even or odd. A basis spinor ϵ_a is right- or left-handed depending on whether the number of spin flips of the index $a = a_1 \dots a_n$, relative to the all-up index $\uparrow \uparrow \dots \uparrow$, is even or odd,

$$\varkappa_N \epsilon_a = \left(\prod_{a_i = \downarrow} (-) \right) \epsilon_a . \quad (40.133)$$

In other words, a basis spinor ϵ_a is right- or left-handed as the number of down $\downarrow$ indices is even or odd. In even N dimensions, the chirality of a spinor is invariant under rotations.

In odd N dimensions, if the path proposed in part 7 is followed, where the algebra is projected into one lower dimension, which requires identifying the chiral operator $\varkappa_N$ with unity, equation (40.125), then rotations mix right- and left-handed spinors, and chirality is not a rotationally invariant property of spinors.

If on the other hand N is odd and the path proposed in part 10 is followed, where the algebra is embedded in one higher dimension, then a basis spinor ϵ_a has $[(N+1)/2]$ bits, and its chirality is that

of the algebra in one higher dimension. The chirality operator is $\varkappa_{N+1}$. In the rest of this part of this Exercise, replace N by $N+1$ if N is odd and part 10 is followed.

Right- and left-handed chiral multivectors are eigenvalues of the chiral operator $\varkappa_N$ (or $\varkappa_{N+1}$ if N is odd and part 10 is followed), with eigenvalues ± 1 ,

$$\varkappa_N \mathbf{a}_R = \pm \mathbf{a}_R . \quad (40.134)$$

Right- and left-handed chirality projection operators P_R may be defined by

$$P_R \equiv \frac{1}{2}(1 \pm \varkappa_N) = \frac{1}{2}(1 \pm i^{-n} I_N) , \quad (40.135)$$

which are projection operators because their squares are one, $(P_N^\pm)^2 = 1$, and their product is zero, $P_N^+ P_N^- = 0$. A multivector $\mathbf{a}$ splits into right- and left-handed chiral parts,

$$\mathbf{a} = \mathbf{a}_R + \mathbf{a}_L, \quad \mathbf{a}_R \equiv P_R \mathbf{a} . \quad (40.136)$$

Since the chiral operator $\varkappa_N$ is proportional to the pseudoscalar I_N , a purely right- or left-handed multivector is necessarily a linear combination of a multivector and its Hodge dual.

An outer product of a right-handed column spinor with any row spinor (right- or left-handed) is a right-handed multivector. An outer product of a left-handed column spinor with any row spinor is a left-handed multivector.

Equations (40.108) provide a matrix representation of the isomorphism between spinor outer products and multivectors. To make the split into right- and left-handed algebras more transparent, it can be convenient to permute the rows and columns of the matrices so that the chiral operator $\varkappa_N$ is represented by the matrix with all positive diagonal entries $+1$ coming first, and all negative diagonal entries -1 coming last (for example, this is the ordering adopted for Dirac spinors in $N = 4$ dimensions, equation (41.20)),

$$\varkappa_N = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \quad (40.137)$$

The 0's and 1's represent zero and unit $2^{n-1} \times 2^{n-1}$ matrices. There are many ways to accomplish the permutation. Since the chirality of a basis spinor ϵ_a is right- or left-handed as the number of down bits in the index a is even or odd, equation (40.133), one possibility is to reorder the rows and columns on a single bit, say the first bit a_1 of the index a , leaving the ordering with respect to all other bits unchanged. The ordering on the chosen single bit is such that the index with total number of down bits even (right-handed) joins the first 2^{n-1} indices, while the index with total number of down bits odd (left-handed) joins the last 2^{n-1} indices.

The result of the permutation is that the matrix representation of a multivector is block diagonal with all right-handed chiral multivectors in the top half, all left-handed chiral multivectors in the bottom half, and with all even multivectors on-diagonal and all odd multivectors off-diagonal,

$$\text{multivector} = \begin{pmatrix} R \text{ even} & R \text{ odd} \\ L \text{ odd} & L \text{ even} \end{pmatrix} . \quad (40.138)$$

The splitting into even and odd multivectors follows because the chiral operator $\varkappa_N$ commutes with all even multivectors but anticommutes with all odd multivectors.

13. **Pure grade components of spinor outer products.** An outer product $\chi\varphi\cdot$ of spinors is a multivector, and its grade p component may be denoted in the usual way, equation (13.27),

$$\langle\chi\varphi\cdot\rangle_p . \quad (40.139)$$

The trace of the outer product is the scalar product,

$$\text{Tr}(\chi\varphi\cdot) = \varphi\cdot\chi . \quad (40.140)$$

The grade 0 component of the outer product $\chi\varphi\cdot$ is the scalar product $\varphi\cdot\chi$ multiplied by the $2^n \times 2^n$ unit matrix 1 normalized by the reciprocal of its trace, $\text{Tr} 1 = 2^n$ (the 1 in equations (40.141)–(40.143) denotes the unit matrix),

$$\langle\chi\varphi\cdot\rangle_0 = (\varphi\cdot\chi) \frac{1}{2^n} . \quad (40.141)$$

If $\mathbf{a}$ is a multivector of grade p , then the scalar sequence $\varphi\cdot\mathbf{a}\chi$, multiplied by the normalized unit matrix, may be re-expressed as the scalar product of $\mathbf{a}$ with the grade p part of $\chi\varphi\cdot$,

$$(\varphi\cdot\mathbf{a}\chi) \frac{1}{2^n} = \langle\mathbf{a}\chi\varphi\cdot\rangle_0 = \mathbf{a}\cdot\langle\chi\varphi\cdot\rangle_p . \quad (40.142)$$

The Hodge dual of the grade p multivector $\mathbf{a}$ is $I_N\mathbf{a}$, and the scalar sequence $\varphi\cdot I_N\mathbf{a}\chi$, multiplied by the normalized unit matrix, may be re-expressed as the Hodge dual of the wedge product of $\mathbf{a}$ with the grade $N-p$ part of $\chi\varphi\cdot$,

$$(\varphi\cdot I_N\mathbf{a}\chi) \frac{1}{2^n} = (I_N\mathbf{a})\cdot\langle\chi\varphi\cdot\rangle_{N-p} = I_N(\mathbf{a}\wedge\langle\chi\varphi\cdot\rangle_{N-p}) . \quad (40.143)$$

14. **Conjugation.** The rotationally-invariant conjugation operator C is defined such that commutation with it converts rotors R in the chiral representation (40.108) to their complex conjugates (with respect to i) (compare equation (40.61)),

$$CR^* = RC . \quad (40.144)$$

Note that since a rotor R is a real linear combination of even orthonormal basis multivectors, the complex conjugate R^* of a rotor R is a rotor. The complex conjugate φ^* of a spinor φ is defined to be its complex conjugate (with respect to i) in the representation (40.108), where the basis spinors ϵ_a are real column vectors,

$$\varphi^* = \varphi^{a*} \epsilon_a . \quad (40.145)$$

The conjugate spinor $\bar{\varphi}$ of a spinor $\varphi = \varphi^a \epsilon_a$ is defined by, equation (40.60),

$$\bar{\varphi} \equiv C\varphi^* = C\varphi^{a*} \epsilon_a . \quad (40.146)$$

The condition (40.144) on the conjugation operator C is imposed precisely so that the conjugate spinor $\bar{\varphi}$ rotates under a rotor R in the same way as the spinor φ ,

$$R : \bar{\varphi} \equiv C\varphi^* \rightarrow C(R\varphi)^* = CR^*\varphi^* = RC\varphi^* = R\bar{\varphi} . \quad (40.147)$$

A necessary and sufficient condition for condition (40.144) to hold is that C commute with all real (with respect to i) orthonormal bivectors, and anticommute with all imaginary orthonormal bivectors. This is the same condition that previously the spinor metric tensor ε was required to satisfy, so C must equal ε (up to a possible normalization factor),

$$C = \varepsilon = \prod_{i=1}^{[(N+1)/2]} \gamma_i^+ . \quad (40.148)$$

If the alternative spinor metric (40.89) is used, then the conjugation operator is

$$C_{\text{alt}} = \varepsilon_{\text{alt}} = \prod_{i=1}^n i\gamma_i^- . \quad (40.149)$$

Choosing $C = \varepsilon$ (or $C_{\text{alt}} = \varepsilon_{\text{alt}}$) without any additional normalization factor ensures that the scalar product $\bar{\varphi} \cdot \varphi$ of a spinor with its own conjugate is real and positive, equation (40.154). There is no loss of generality in imposing that ε , hence C , be real. If ε were multiplied by an arbitrary complex phase, then the conjugation operator would have to be defined by $C = \varepsilon^*$ in place of the definition (40.148), in order that the scalar product of a spinor with its conjugate remain real and positive, equation (40.154). The modification by a phase leaves various key results unchanged; for example the double conjugate of a spinor, equation (40.151), becomes $\bar{\bar{\varphi}} = CC^*\varphi$, which is unaffected by a complex phase in C .

The conjugate of a basis spinor ϵ_a is

$$\bar{\epsilon}_a \equiv C\epsilon_a = \pm\epsilon_{\bar{a}} , \quad (40.150)$$

where the conjugate index $\bar{a}$ is the index a with all bits flipped. Conjugation flips the chirality of a spinor if n is odd, and leaves the chirality unchanged if n is even. The $\pm$ sign in equation (40.150) is as given by equation (40.94), or by equation (40.95) if the alternative spinor metric is used. Conjugation flips all the bits of a spinor; for example, the conjugate of the all-up basis spinor is the all-down basis spinor, $\bar{\epsilon}_{\uparrow\uparrow\dots\uparrow} = \pm\epsilon_{\downarrow\downarrow\dots\downarrow}$. The conjugate spinor $\bar{\varphi}$ of a spinor φ is, equation (40.146), $\bar{\varphi} = \varphi^{a*}\bar{\epsilon}_a$. The double conjugate of a spinor is

$$\bar{\bar{\varphi}} = C^2\varphi = \varepsilon^2\varphi , \quad (40.151)$$

where the sign ε^2 is as given in Table 40.1.

The scalar product of a conjugate spinor $\bar{\varphi}$ with a spinor χ is

$$\bar{\varphi} \cdot \chi = (C\varphi^*)^\top \varepsilon \chi = \varphi^\dagger C^\top \varepsilon \chi = \varphi^\dagger \chi , \quad (40.152)$$

which is a complex number. In particular, the scalar product of a conjugate basis spinor $\bar{\epsilon}_a$ with a basis spinor ϵ_b is a Kronecker delta,

$$\bar{\epsilon}_a \cdot \epsilon_b = \delta_{ab} . \quad (40.153)$$

The scalar product of a spinor φ with its own conjugate is real and positive,

$$\bar{\varphi} \cdot \varphi = \varphi^\dagger \varphi . \quad (40.154)$$

The scalar product of $\bar{\varphi}$ with its conjugate is the same as the scalar product (40.154) of φ with its conjugate,

$$\bar{\bar{\varphi}} \cdot \bar{\varphi} = C^2 \varphi \cdot \bar{\varphi} = \bar{\varphi} \cdot \varphi . \quad (40.155)$$

The complex conjugate (with respect to i) $\mathbf{a}^*$ of a multivector $\mathbf{a} = a^A \boldsymbol{\gamma}_A$ is defined to be its complex conjugate in the chiral representation (40.108) of multivectors,

$$\mathbf{a}^* = a^{A*} \boldsymbol{\gamma}_A^* . \quad (40.156)$$

In the representation (40.108), the chiral basis vectors $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_{\bar{i}}$ (and the final vector $\boldsymbol{\gamma}_N$ if N is odd) are real, so the orthonormal basis vectors $\boldsymbol{\gamma}_i^+$ and $\boldsymbol{\gamma}_i^-$ are respectively real and imaginary. The conjugate $\bar{\mathbf{a}}$ of a multivector $\mathbf{a} = a^A \boldsymbol{\gamma}_A$ is defined to be, consistent with the definition (40.146) of the conjugate of a spinor,

$$\bar{\mathbf{a}} \equiv C \mathbf{a}^* C^{-1} . \quad (40.157)$$

The conjugate multivector $\bar{\mathbf{a}}$ rotates under a rotor R in the same way as the multivector $\mathbf{a}$,

$$R : \bar{\mathbf{a}} \equiv C \mathbf{a}^* C^{-1} \rightarrow C(R \mathbf{a} \tilde{R})^* C^{-1} = C R^* \mathbf{a}^* \tilde{R}^* C^{-1} = R C \mathbf{a}^* C^{-1} \tilde{R} = R \bar{\mathbf{a}} \tilde{R} . \quad (40.158)$$

Conjugation is multiplicative over multivectors, and over multivectors with spinors,

$$\overline{\mathbf{a} \mathbf{b}} = \bar{\mathbf{a}} \bar{\mathbf{b}} , \quad \overline{\varphi \psi} = \bar{\varphi} \bar{\psi} . \quad (40.159)$$

If the outer product of two spinors φ and χ equals the multivector $\mathbf{a}$, then the outer product of conjugate spinors $\bar{\varphi}$ and $\bar{\chi}$ equals the conjugate multivector $\bar{\mathbf{a}}$,

$$\varphi \chi \cdot = \mathbf{a} , \quad \bar{\varphi} \bar{\chi} \cdot = \bar{\mathbf{a}} . \quad (40.160)$$

Equation (41.152) holds because (with $C = \varepsilon$)

$$\bar{\varphi} \bar{\chi} \cdot \equiv \varepsilon \varphi^* (\varepsilon \chi^*)^\top \varepsilon = \varepsilon \varphi^* (\chi^*)^\top = \varepsilon (\varphi \chi^\top \varepsilon)^* \varepsilon^{-1} = \varepsilon \mathbf{a}^* \varepsilon^{-1} = \bar{\mathbf{a}} . \quad (40.161)$$

The conjugate of a basis multivector $\boldsymbol{\gamma}_A$ is defined to be

$$\bar{\boldsymbol{\gamma}}_A \equiv C \boldsymbol{\gamma}_A^* C^{-1} , \quad (40.162)$$

so that a conjugate multivector $\bar{\mathbf{a}}$ is

$$\bar{\mathbf{a}} = a^{A*} \bar{\boldsymbol{\gamma}}_A . \quad (40.163)$$

The conjugate of an orthonormal basis vector $\boldsymbol{\gamma}_a$ is

$$\bar{\boldsymbol{\gamma}}_a = \pm \boldsymbol{\gamma}_a , \quad (40.164)$$

where the $\pm$ factor depends on the choice of spinor metric, and is as given in Table 40.2. The conjugates of chiral basis vectors γ_i and $\gamma_{\bar{i}}$ defined by equations (40.80) have their index flipped,

$$\bar{\gamma}_i = \pm \gamma_{\bar{i}} , \quad \bar{\gamma}_{\bar{i}} = \pm \gamma_i , \quad (40.165)$$

where the $\pm$ factor is again as given in Table 40.2.

15. **Real subalgebra.** The chiral matrix representations (40.106) of the column and row basis spinors ϵ_a and $\epsilon_a \cdot$, and (40.108) of their outer products (which yield the full set of basis multivectors in the chiral representation), are all real. One might therefore contemplate forming a real subalgebra consisting of spinors $\varphi^a \epsilon_a$ and multivectors $a^A \gamma_A$ with real coefficients φ^a and a^A in the chiral representation. This does not work however, because spatial rotations transform the basis spinors (and their outer products) into complex combinations of each other, equations (40.84). Any viable subalgebra must be closed under rotations.

Orthonormal basis multivectors on the other hand do transform into real linear combinations of each other under rotations. A real subalgebra of the complex geometric algebra may be obtained by restricting to multivectors satisfying the reality condition that they are their own conjugates,

$$\bar{a} = a . \quad (40.166)$$

Conjugates of orthonormal basis vectors γ_a are equal to either plus themselves, or minus themselves, depending on the choice of spinor metric, equation (40.164). If the conjugates of the orthonormal basis vectors are themselves, $\bar{\gamma}_a = \gamma_a$ (+ in Table 40.2), then the real subalgebra consists of real linear combinations of orthonormal basis multivectors. If the conjugates of the orthonormal basis vectors are minus themselves, $\bar{\gamma}_a = -\gamma_a$ (− in Table 40.2), then the real subalgebra consists of linear combinations of odd-grade orthonormal multivectors with pure imaginary coefficients and even-grade orthonormal multivectors with pure real coefficients.

A real supergeometric subalgebra may similarly be obtained by restricting to spinors satisfying the reality condition that they are their own conjugates,

$$\bar{\varphi} = \varphi . \quad (40.167)$$

The spinor reality condition (40.167) is more restrictive than the multivector reality condition (40.166). Whereas the multivector reality condition (40.166) can always be imposed, the spinor reality condition (40.167) can be imposed only if the double conjugate spinor is itself, equation (40.151), which is to say, only if the spinor metric is symmetric, Table 40.1.

If the self-conjugate condition (40.167) holds, then the relation (41.152) implies that outer products of self-conjugate spinors φ and χ are self-conjugate multivectors,

$$\bar{a} = \bar{\varphi} \bar{\chi} \cdot = \varphi \chi \cdot = a . \quad (40.168)$$

Thus the multivector part of the real supergeometric subalgebra is the real geometric subalgebra corresponding to the reality condition (40.166) for a symmetric choice of spinor metric.

16. **Transformations that leave the spinor scalar product unchanged.** The spinor metric ε , hence the spinor scalar product, is by definition invariant under rotations, that is, under the rotor group generated by bivectors of the geometric algebra. However, the geometric algebra contains multivectors of other grades, that generate other Lie groups of transformations of the algebra, Exercise 13.6. An element R of a Lie group generated by a set of orthonormal multivectors γ_A takes the form $R = \exp(-\frac{1}{2} \sum_A \theta_A \gamma_A)$, where, depending on the choice of group, the coefficients θ_A could be real or imaginary or complex. The element R transforms multivectors $\mathbf{a}$ by $R : \mathbf{a} \rightarrow R\mathbf{a}R^{-1}$, the inverse of R being $R^{-1} = \exp(\frac{1}{2} \sum_A \theta_A \gamma_A)$. All such transformations preserve the scalar product of multivectors. However, not all such transformations preserve the scalar product of spinors.

Let $R = e^{-\theta \gamma_A/2}$ be a transformation generated by an orthonormal basis multivector γ_A , with θ real, imaginary, or complex. The condition for the spinor metric ε to be invariant under the transformation R is that commuting the generator γ_A through ε should convert it to minus its transpose,

$$\gamma_A^\top \varepsilon = -\varepsilon \gamma_A . \quad (40.169)$$

For then $R^\top \varepsilon = e^{-\theta \gamma_A^\top/2} \varepsilon = \varepsilon e^{\theta \gamma_A/2} = \varepsilon R^{-1}$, which implies that a scalar product $\varphi \cdot \chi$ of spinors is invariant under R ,

$$(R\varphi) \cdot (R\chi) = \varphi^\top R^\top \varepsilon R \chi = \varphi^\top \varepsilon R^{-1} R \chi = \varphi \cdot \chi . \quad (40.170)$$

Comparing the condition (40.169) to the actual commutation rule (40.98) shows that the grades of orthonormal multivectors that generate transformations that leave the spinor scalar product unchanged are, with the $+$ or $-$ from Table 40.2,

$$+ : \text{grades } (2 \text{ or } 3) \bmod 4 \quad (\text{thus } 2, 3, 6, 7, \dots) , \quad (40.171a)$$

$$- : \text{grades } (1 \text{ or } 2) \bmod 4 \quad (\text{thus } 1, 2, 5, 6, \dots) . \quad (40.171b)$$

For tilde'd spinor metrics, equation (40.98) holds, hence the list (40.171) holds, for multivectors γ_A that do *not* include a factor of whichever is the scalar dimension, γ_N or γ_{N+1} . If the multivector γ_A includes a factor of the scalar dimension γ_N or γ_{N+1} , then equation (40.98) holds with an extra minus sign (the grade p being that of γ_A including the scalar dimension), and the grades of generators that leave the spinor scalar product unchanged are the complement of those in the list (40.171).

If the scalar product is between spinors and conjugate spinors, then whether a transformation $R = e^{-\theta \gamma_A/2}$ generated by a grade- p multivector γ_A preserves the spinor scalar product depends on whether the coefficient θ is real or imaginary. A conjugate spinor $\bar{\varphi}$ transforms under R as

$$R : \bar{\varphi} \equiv C\varphi^* \rightarrow C(R\varphi)^* . \quad (40.172)$$

The commutation rule (40.144) for rotors is replaced by

$$CR^* = Ce^{-\theta^* \gamma_A^*/2} = e^{-\theta^* \gamma_A/2} C = e^{-(\pm)^p \theta^* \gamma_A/2} C , \quad (40.173)$$

where the $\pm$ sign in $(\pm)^p$, from equation (40.164), is as given in Table 40.2. For rotors, which are generated by real linear combinations of bivectors, the grade p is 2, and θ is real, and equation (40.173)

recovers the commutation rule (40.144). The scalar product $\bar{\varphi} \cdot \chi$ of a conjugate spinor with a spinor transforms under R to

$$(C(R\varphi)^*) \cdot (R\chi) = \bar{\varphi} e^{-(\pm)^p \theta^* \gamma_A^\top / 2} \varepsilon R\chi = \bar{\varphi} \varepsilon e^{-(\pm)^{[p/2]} \theta^* \gamma_A / 2} R\chi, \quad (40.174)$$

where the sign $(-)^{[p/2]}$ in the final expression is the product of $(\pm)^p$ and the sign $(\pm)^p (-)^{[p/2]}$ in the commutation rule (40.98) of a multivector γ_A through the spinor metric ε . The spinor product is preserved provided that $e^{-(\pm)^{[p/2]} \theta^* \gamma_A / 2} = R^{-1}$, which is to say provided that

$$-(-)^{[p/2]} \theta^* = \theta. \quad (40.175)$$

Therefore the scalar product of spinors and conjugate spinors is preserved under transformations generated by multivectors of grade p provided that the coefficient θ satisfies

$$\theta \text{ real} \quad \text{grades } (2 \text{ or } 3) \bmod 4 \quad (\text{thus } 2, 3, 6, 7, \dots), \quad (40.176a)$$

$$\theta \text{ imaginary} \quad \text{grades } (0 \text{ or } 1) \bmod 4 \quad (\text{thus } 0, 1, 4, 5, \dots). \quad (40.176b)$$

17. **Rotor group.** Unimodular elements of the (even or odd) geometric algebra generated by the $N(N-1)/2$ orthonormal bivectors $\gamma_a \wedge \gamma_b$ form a group, the rotor group, also called the spin group, or $\text{Spin}(N)$. The rotor group $\text{Spin}(N)$ comprises all distinct rotations of spinors (spin- $\frac{1}{2}$ objects) in N dimensions, and is the double cover of the special orthogonal group $\text{SO}(N)$, which comprises all distinct rotations of vectors (spin-1 objects) in N dimensions.

As noted in part 11 of this Exercise, the chiral representation represents orthonormal basis bivectors $\gamma_a \wedge \gamma_b$ in even N dimensions by traceless, skew-Hermitian, unitary $2^n \times 2^n$ matrices. The rotor group generated by the basis bivectors is then represented by unitary $2^n \times 2^n$ matrices. Thus the rotor group in even N dimensions is a subgroup of $\text{SU}(2^n)$, the special unitary group in 2^n dimensions,

$$\text{Spin}(N) \subset \text{SU}(2^n). \quad (40.177)$$

The embedding (40.177) holds also if N is odd, since as described in part 7, in odd N dimensions the N -dimensional chiral operator $\varkappa_N$ can be identified with unity, equation (40.125), in which case the final odd vector γ_N is equivalent to the $(N-1)$ -dimensional chiral operator $\varkappa_{N-1}$, and bivectors $\gamma_a \wedge \gamma_N$ are again represented by traceless, skew-Hermitian, unitary $2^n \times 2^n$ matrices.

The generators of the unitary group are skew-Hermitian. The orthonormal basis multivectors of the geometric algebra are either skew-Hermitian (grades $p = (2 \text{ or } 3) \bmod 4$) or Hermitian (grades $p = (0 \text{ or } 1) \bmod 4$). Multiplying a Hermitian generator by the imaginary i makes it skew-Hermitian. The set of 2^N orthonormal basis multivectors in even N dimensions, with Hermitian multivectors multiplied by i , generates the full unitary group $\text{U}(2^n)$. This is the group denoted $G^{23i01}(N)$ by Shirokov (2018),

$$G^{23i01}(N) \cong \text{U}(2^n). \quad (40.178)$$

If the generator consisting of i times the unit matrix is excised, the result is the special unitary group $\text{SU}(2^n)$.

18. **Grade-preserving subgroup of $\text{Spin}(2N)$.** The rotor group $\text{Spin}(2N)$ contains a subgroup that preserves the spinor grade, the number of up bits, of a spinor (Atiyah, Bott, and Shapiro, 1964). The subgroup is isomorphic to $U(N)$, so that

$$\text{SU}(N) \subset U(N) \subset \text{Spin}(2N) . \quad (40.179)$$

The generators of $\text{Spin}(2N)$ that preserve spinor grade are bivectors with zero total spin. These generators must be real linear combinations of orthonormal $\text{Spin}(2N)$ bivectors that, when expressed in terms of spin vectors $\gamma_i^\pm$, are (complex) linear combinations of bivectors of the form $\gamma_i \wedge \gamma_j$. Such a bivector flips the i th bit of a spinor from down to up, and the j th bit from up to down, preserving the total number of up bits of the spinor. Linearly independent generators satisfying these conditions are

$$\gamma_i^+ \wedge \gamma_j^+ + \gamma_i^- \wedge \gamma_j^- = \gamma_i \wedge \gamma_j - \gamma_j \wedge \gamma_i \quad (N(N-1)/2 \text{ generators}) , \quad (40.180a)$$

$$\gamma_i^+ \wedge \gamma_j^- - \gamma_i^- \wedge \gamma_j^+ = i(\gamma_i \wedge \gamma_j + \gamma_j \wedge \gamma_i) \quad (N(N-1)/2 \text{ generators}) , \quad (40.180b)$$

$$\gamma_i^+ \wedge \gamma_i^- = i \gamma_i \wedge \gamma_i \quad (N \text{ generators}) , \quad (40.180c)$$

a total of N^2 generators. The Lie algebra of commutators of the generators (40.180) coincides with the Lie algebra of commutators in which $\frac{1}{2}\gamma_i \wedge \gamma_j$ is represented by the $N \times N$ matrix 1_{ij} with 1 in the ij th entry and 0 elsewhere,

$$\frac{1}{2}\gamma_i \wedge \gamma_j \rightarrow 1_{ij} . \quad (40.181)$$

But that algebra is just that of the group $U(N)$ of unitary $N \times N$ matrices. The generator $\frac{i}{2} \sum_i \gamma_i \wedge \gamma_i$ is represented by i times the unit matrix, which generates a rotation by an overall phase. Eliminating that generator yields the algebra of the group $\text{SU}(N)$ of special unitary $N \times N$ matrices. Thus $U(N)$ and $\text{SU}(N)$ are subgroups of $\text{Spin}(2N)$ as claimed. The chain (40.179) of subgroups extends (trivially) to

$$\text{SU}(N) \subset U(N) \subset \text{Spin}(2N) \subset \text{Spin}(2N+1) . \quad (40.182)$$

Super spacetime algebra

This Chapter presents the **super spacetime algebra**, the generalization of the spacetime algebra in 3+1 spacetime dimensions to include spinors. The generalization to arbitrarily many space and time dimensions is given as Exercise 41.5.

41.1 Newman-Penrose formalism

The extension of the spacetime algebra to spinors is most direct when the basis vectors of the spacetime algebra are expressed in a Newman-Penrose basis (Newman and Penrose, 1962). Newman-Penrose adopts a tetrad in which two of the tetrad axes are lightlike, γ_v (outgoing) and γ_u (ingoing), while the remaining two axes γ_+ and γ_- are spin axes.

41.1.1 Newman-Penrose tetrad

A Newman-Penrose tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$ is defined in terms of an orthonormal tetrad $\{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$, (or $\{\gamma_t, \gamma_x, \gamma_y, \gamma_z\}$ if you prefer), by

$$\boxed{\gamma_v \equiv \frac{1}{\sqrt{2}}(\gamma_0 + \gamma_3)} \quad , \quad (41.1a)$$

$$\boxed{\gamma_u \equiv \frac{1}{\sqrt{2}}(\gamma_0 - \gamma_3)} \quad , \quad (41.1b)$$

$$\boxed{\gamma_+ \equiv \frac{1}{\sqrt{2}}(\gamma_1 + i\gamma_2)} \quad , \quad (41.1c)$$

$$\boxed{\gamma_- \equiv \frac{1}{\sqrt{2}}(\gamma_1 - i\gamma_2)} \quad , \quad (41.1d)$$

or in matrix form

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & i & 0 \\ 0 & 1 & -i & 0 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}. \quad (41.2)$$

All four tetrad axes are null

$$\gamma_v \cdot \gamma_v = \gamma_u \cdot \gamma_u = \gamma_+ \cdot \gamma_+ = \gamma_- \cdot \gamma_- = 0. \quad (41.3)$$

The tetrad metric of the Newman-Penrose tetrad $\{\gamma_v, \gamma_u, \gamma_+, \gamma_-\}$ is

$$\gamma_{mn} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \quad (41.4)$$

41.1.2 Boost and spin weight

An object is defined to have **boost weight** n if it varies by

$$e^{n\theta} \quad (41.5)$$

under a boost by rapidity θ along the positive 3-direction.

Under a boost by rapidity θ in the 3-direction, the basis vectors γ_m transform as (14.44)

$$\gamma_0 \rightarrow \gamma_0 \cosh \theta + \gamma_3 \sinh \theta, \quad (41.6a)$$

$$\gamma_3 \rightarrow \gamma_3 \cosh \theta + \gamma_0 \sinh \theta, \quad (41.6b)$$

$$\gamma_a \rightarrow \gamma_a \quad (a = 1, 2). \quad (41.6c)$$

It follows that a boost by rapidity θ in the 3-direction multiplies the outgoing and ingoing axes γ_v and γ_u by a blueshift factor e^θ and its reciprocal,

$$\gamma_v \rightarrow e^\theta \gamma_v, \quad \gamma_u \rightarrow e^{-\theta} \gamma_u. \quad (41.7)$$

In terms of the boost velocity $v = \tanh \theta$ (not to be confused with the Newman-Penrose index v), the blueshift factor is the special relativistic Doppler shift factor

$$e^\theta = \left(\frac{1+v}{1-v} \right)^{1/2}. \quad (41.8)$$

Thus γ_v has boost weight $+1$, and γ_u has boost weight -1 . The spin axes $\gamma_\pm$ both have boost weight 0. The Newman-Penrose components of a tensor inherit their boost weight properties from those of the Newman-Penrose basis. The general rule is that the boost weight n of any tensor component is equal to the number of v covariant indices minus the number of u covariant indices:

$$\boxed{\text{boost weight } n = \text{number of } v \text{ minus } u \text{ covariant indices}}. \quad (41.9)$$

The operation of boosting along the 3-axis, which is the same as a rotation in the $\gamma_0\gamma_3$ -plane, commutes with the operation of rotating in the $\gamma_1\gamma_2$ -plane. The concept of spin weight presented in §40.2 holds unchanged. The outgoing and ingoing basis vectors γ_v and γ_u have spin weight zero, while γ_+ and γ_- have spin weight $+1$ and -1 . The general rule is that the spin weight s of any tensor component equals the number of $+$ covariant indices minus the number of $-$ covariant indices (this repeats rule (40.13)):

$$\boxed{\text{spin weight } s = \text{number of } + \text{ minus } - \text{ covariant indices}} . \quad (41.10)$$

The boost and spin properties of the components of a tensor are thus manifest in a Newman-Penrose tetrad.

41.2 Chiral representation of γ -matrices

The **chiral representation** of the Dirac γ -matrices provides the natural extension of the Newman-Penrose tetrad to spin- $\frac{1}{2}$ particles. The chiral representation may be obtained from the Dirac representation (14.102) by the transformation (Dirac $\rightarrow$ chiral)

$$X : \gamma_m \rightarrow X\gamma_m X^{-1} , \quad (41.11)$$

where X is the symmetric ($X = X^\top$), unitary ($X^{-1} = X^\dagger$) matrix

$$X \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & i & 0 \\ 0 & 1 & 0 & i \\ i & 0 & 1 & 0 \\ 0 & i & 0 & 1 \end{pmatrix} . \quad (41.12)$$

As in the Dirac representation, all the γ -matrices in the chiral representation are traceless; the only basis matrix of the algebra with finite trace is the unit matrix. The γ -matrices in the chiral representation are the unitary matrices

$$\gamma_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} , \quad \gamma_a = \begin{pmatrix} 0 & \sigma_a \\ \sigma_a & 0 \end{pmatrix} . \quad (41.13)$$

The bivectors σ_a and $I\sigma_a$ and the pseudoscalar I are

$$\gamma_0\gamma_a = \sigma_a = \begin{pmatrix} \sigma_a & 0 \\ 0 & -\sigma_a \end{pmatrix} , \quad \frac{1}{2}\varepsilon_{abc}\gamma_b\gamma_c = I\sigma_a = i \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_a \end{pmatrix} , \quad I = i \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \quad (41.14)$$

The Newman-Penrose basis vectors in the chiral representation are the real matrices

$$\gamma_v = \begin{pmatrix} 0 & \sigma_v \\ -\sigma_u & 0 \end{pmatrix} , \quad \gamma_u = \begin{pmatrix} 0 & \sigma_u \\ -\sigma_v & 0 \end{pmatrix} , \quad \gamma_+ = \begin{pmatrix} 0 & \sigma_+ \\ \sigma_+ & 0 \end{pmatrix} , \quad \gamma_- = \begin{pmatrix} 0 & \sigma_- \\ \sigma_- & 0 \end{pmatrix} , \quad (41.15)$$

where σ_m are the Newman-Penrose Pauli matrices

$$\sigma_v \equiv \frac{1}{\sqrt{2}}(1 + \sigma_3) = \sqrt{2} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \sigma_u \equiv \frac{1}{\sqrt{2}}(1 - \sigma_3) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad (41.16a)$$

$$\sigma_+ \equiv \frac{1}{\sqrt{2}}(\sigma_1 + i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_- \equiv \frac{1}{\sqrt{2}}(\sigma_1 - i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \quad (41.16b)$$

The Newman-Penrose bivectors form 6 real matrices that group into three right-handed bivectors (notation $\gamma_{mn} \equiv \gamma_m \wedge \gamma_n$),

$$\gamma_{v+} = \sqrt{2} \begin{pmatrix} \sigma_+ & 0 \\ 0 & 0 \end{pmatrix}, \quad \frac{1}{2}(\gamma_{vu} - \gamma_{+-}) = \begin{pmatrix} -\sigma_3 & 0 \\ 0 & 0 \end{pmatrix}, \quad \gamma_{u-} = \sqrt{2} \begin{pmatrix} \sigma_- & 0 \\ 0 & 0 \end{pmatrix}, \quad (41.17)$$

and three left-handed bivectors,

$$\gamma_{u+} = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & -\sigma_+ \end{pmatrix}, \quad \frac{1}{2}(\gamma_{vu} + \gamma_{+-}) = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_3 \end{pmatrix}, \quad \gamma_{v-} = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & -\sigma_- \end{pmatrix}. \quad (41.18)$$

The chiral matrix γ_5 is

$$\gamma_5 \equiv -iI = -\gamma_v \wedge \gamma_u \wedge \gamma_+ \wedge \gamma_- = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (41.19)$$

By construction, the chiral matrix γ_5 is diagonal in the chiral representation.

41.3 Basis spinors

Introduce a tetrad of basis spinors ϵ_a ,

$$\epsilon_a \equiv \{\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}, \epsilon_{\uparrow\downarrow}\}. \quad (41.20)$$

The indices $\{\uparrow\uparrow, \downarrow\downarrow, \downarrow\uparrow, \uparrow\downarrow\}$ signify the transformation properties of the basis spinors: $\uparrow$ and $\downarrow$ signify boost weight $+\frac{1}{2}$ and $-\frac{1}{2}$, while $\uparrow$ and $\downarrow$ signify spin weight $+\frac{1}{2}$ and $-\frac{1}{2}$. Under a Lorentz transformation, the basis spinors ϵ_a are defined to transform in the same way as rotors,

$$R : \epsilon_a \rightarrow R\epsilon_a. \quad (41.21)$$

In the chiral representation (41.13) the basis spinors ϵ_a are the column spinors

$$\epsilon_{\uparrow\uparrow} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon_{\downarrow\downarrow} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon_{\downarrow\uparrow} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \epsilon_{\uparrow\downarrow} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad (41.22)$$

which are Lorentz-transformed by pre-multiplying by rotors expressed in the chiral representation. The

basis spinors ϵ_a in the chiral representation may be obtained from those in the Dirac representation by the transformation (Dirac $\rightarrow$ chiral)

$$X : \epsilon_a \rightarrow X \epsilon_a , \quad (41.23)$$

where the matrix X is defined by equation (41.12).

The basis spinors ϵ_a are eigenvectors of the chirality operator γ_5 , equation (41.19), with eigenvalues ± 1 . Positive chirality spinors are called right-handed, while negative chirality spinors are called left-handed. The first two basis spinors are right-handed, while the last two are left-handed,

$$\gamma_5 \epsilon_{\uparrow\uparrow} = \epsilon_{\uparrow\uparrow} , \quad \gamma_5 \epsilon_{\downarrow\downarrow} = \epsilon_{\downarrow\downarrow} , \quad \gamma_5 \epsilon_{\downarrow\uparrow} = -\epsilon_{\downarrow\uparrow} , \quad \gamma_5 \epsilon_{\uparrow\downarrow} = -\epsilon_{\uparrow\downarrow} . \quad (41.24)$$

Lorentz transformation preserves chirality, as is evident from the block-diagonal form of the even elements of the spacetime algebra in the chiral representation, equations (41.14). The right-handed basis spinors $\epsilon_{\uparrow\uparrow}$ and $\epsilon_{\downarrow\downarrow}$ are called right-handed because the boost axis and the spin axis point in the same direction (along the 3-direction for $\epsilon_{\uparrow\uparrow}$, and along the negative 3-direction for $\epsilon_{\downarrow\downarrow}$). Conversely, the left-handed basis spinors $\epsilon_{\downarrow\uparrow}$ and $\epsilon_{\uparrow\downarrow}$ are called left-handed because the boost axis and the spin axis point in opposite directions.

A Lorentz boost $R = e^{\sigma_3 \theta/2} = \cosh(\theta/2) + \sigma_3 \sinh(\theta/2)$ by rapidity θ along the spin axis (3-axis) multiplies the basis spinors ϵ_a by $e^{\pm\theta/2}$ according to

$$\epsilon_{\uparrow\uparrow} \rightarrow e^{\theta/2} \epsilon_{\uparrow\uparrow} , \quad \epsilon_{\downarrow\downarrow} \rightarrow e^{-\theta/2} \epsilon_{\downarrow\downarrow} , \quad \epsilon_{\downarrow\uparrow} \rightarrow e^{-\theta/2} \epsilon_{\downarrow\uparrow} , \quad \epsilon_{\uparrow\downarrow} \rightarrow e^{\theta/2} \epsilon_{\uparrow\downarrow} . \quad (41.25)$$

The transformations (41.25) confirm that the basis spinors with a $\uparrow$ index have boost weight $+\frac{1}{2}$, while the basis spinors with a $\downarrow$ index have boost weight $-\frac{1}{2}$. A right-handed spatial rotation $R = e^{-I\sigma_3 \theta/2} = \cos(\theta/2) - I\sigma_3 \sin(\theta/2)$ by rotation angle θ about the spin axis (3-axis) multiplies the basis spinors ϵ_a by $e^{\pm i\theta/2}$ according to

$$\epsilon_{\uparrow\uparrow} \rightarrow e^{-i\theta/2} \epsilon_{\uparrow\uparrow} , \quad \epsilon_{\downarrow\downarrow} \rightarrow e^{i\theta/2} \epsilon_{\downarrow\downarrow} , \quad \epsilon_{\downarrow\uparrow} \rightarrow e^{-i\theta/2} \epsilon_{\downarrow\uparrow} , \quad \epsilon_{\uparrow\downarrow} \rightarrow e^{i\theta/2} \epsilon_{\uparrow\downarrow} . \quad (41.26)$$

The transformations (41.26) confirm that the basis spinors with a $\uparrow$ index have spin weight $+\frac{1}{2}$, while the basis spinors with a $\downarrow$ index have spin weight $-\frac{1}{2}$. This justifies the choice of indices on the basis spinors. Spinor tensors inherit their boost and spin weights from those of the basis spinors. The rules are

$$\boxed{\text{boost weight } n = \frac{1}{2} (\text{number of } \uparrow \text{ minus } \downarrow \text{ covariant indices})} , \quad (41.27a)$$

$$\boxed{\text{spin weight } s = \frac{1}{2} (\text{number of } \uparrow \text{ minus } \downarrow \text{ covariant indices})} , \quad (41.27b)$$

which generalize the rules (41.9) and (41.10). The rules (41.27) hold not only for column spinors ϵ_a , but also for row spinors $\epsilon_a \cdot$, §41.6.2, and for inner and outer products of spinors, §41.6.3 and §41.7.1.

41.4 Weyl spinors

Pure right- or left-handed spinors are called Weyl spinors.

$$\psi_R = \psi^{\uparrow\uparrow} \epsilon_{\uparrow\uparrow} + \psi^{\downarrow\downarrow} \epsilon_{\downarrow\downarrow} . \quad (41.28)$$

41.5 Dirac and Weyl spinors

Dirac basis spinors $\epsilon_a \equiv \{\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\downarrow}, \epsilon_{\downarrow\uparrow}, \epsilon_{\downarrow\downarrow}\}$ are massive linear combinations of chiral spinors,

$$\epsilon_a^{\text{Dirac}} = X_{ab} \epsilon_b^{\text{chiral}} , \quad (41.29)$$

The Dirac basis spinors $\epsilon_{\uparrow\uparrow}$ and $\epsilon_{\uparrow\downarrow}$ represent massive spinors in their rest frames, while $\epsilon_{\downarrow\uparrow}$ and $\epsilon_{\downarrow\downarrow}$ represent massive antispinors in their rest frames.

A **Dirac spinor** ψ is a complex (with respect to i) linear combination of the 4 basis spinors ϵ_a ,

$$\psi = \psi^a \epsilon_a . \quad (41.30)$$

A Dirac spinor has 4 complex components, making 8 degrees of freedom in all. Just as a multivector $a^m \gamma_m$ is a vector in the spacetime algebra, so also $\psi^a \epsilon_a$ is a spinor in the super spacetime algebra.

A Dirac spinor ψ Lorentz transforms as

$$R : \psi \rightarrow R\psi . \quad (41.31)$$

A Dirac spinor ψ is a spin- $\frac{1}{2}$ object, in the sense that a rotation by 2π changes the sign of the spinor, and a rotation by 4π is required to return the spinor to its original value.

Concept question 41.1 Lorentz transformation of the phase of a spinor. Should not a Lorentz transformation also change the phase of a spinor ψ as a function of position? For example, if the phase is $\psi \sim e^{-imt}$ in the spinor rest frame, would not the phase be $\psi \sim e^{-i\omega t + i\mathbf{k} \cdot \mathbf{x}}$ in the Lorentz-transformed frame?

Answer. No. A Lorentz transformation is a tetrad transformation, not a coordinate transformation. That being said, in flat (Minkowski) space it is possible to choose inertial coordinates $\{t, \mathbf{x}\}$ aligned everywhere with the tetrad frame. It is true that $\psi \sim e^{-i\omega t + i\mathbf{k} \cdot \mathbf{x}}$ with respect to Lorentz-transformed inertial coordinates.

41.5.1 Weyl decomposition of a Dirac spinor

A Dirac spinor ψ can be decomposed into a sum of right- and left-handed chiral **Weyl spinors** ψ_R and ψ_L

$$\psi = \psi_R + \psi_L , \quad (41.32)$$

that are right- and left-handed eigenvectors of the chiral operator γ_5 ,

$$\gamma_5 \psi_{\begin{smallmatrix} R \\ L \end{smallmatrix}} = \pm \psi_{\begin{smallmatrix} R \\ L \end{smallmatrix}} . \quad (41.33)$$

The right- and left-handed chiral spinors can be projected out by applying the chiral projection operators $\frac{1}{2}(1 \pm \gamma_5)$ (which are projection operators because their squares are themselves) to the Dirac spinor ψ ,

$$\psi_{\begin{smallmatrix} R \\ L \end{smallmatrix}} = \frac{1}{2}(1 \pm \gamma_5) \psi . \quad (41.34)$$

Since chirality is Lorentz invariant, the chiral decomposition of a Dirac spinor is unique. The right- and

left-handed components of a Dirac spinor each contain 2 complex components. The right- and left-handed components of a Dirac spinor cannot be rotated into each other by any Lorentz transformation.

41.6 Spinor scalar product

41.6.1 Spinor metric tensor

In a matrix representation, the tensor product of Dirac basis spinors ϵ_a and ϵ_b can be represented as the matrix $\epsilon_a \epsilon_b^\top$, a matrix product of the column spinor ϵ_a with the row spinor $\epsilon_b^\top$. In accordance with the transformation rule (41.21), the tensor product of basis spinors Lorentz transforms as

$$R : \epsilon_a \epsilon_b^\top \rightarrow R \epsilon_a \epsilon_b^\top R^\top . \quad (41.35)$$

Consider the spinor metric tensor ε with the defining property that for any Lorentz rotor R

$$R^\top \varepsilon = \varepsilon \tilde{R} . \quad (41.36)$$

That ε defines a Lorentz-invariant spinor metric will be seen in §41.6.3. A Lorentz rotor is a real (with respect to i) linear combination of even elements 1 , I , σ_a , and $I\sigma_a$ of the spacetime algebra. Consequently, in the Dirac representation (14.103) a necessary and sufficient condition for (41.36) to hold is that ε anti-commutes with I and σ_2 , and commutes with σ_1 and σ_3 . This requires that ε be proportional to γ_2 in the Dirac representation, with a proportionality factor that could be some arbitrary complex (with respect to i and/or I) number. To be consistent with standard Dirac theory, the spinor metric tensor ε in the Dirac representation (14.103) is taken to be the real unitary matrix

$$\varepsilon_{\text{Dirac}} = i\gamma_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} . \quad (41.37)$$

Despite the equality of ε and $i\gamma_2$ in the Dirac representation, ε is defined to transform as a spinor tensor under Lorentz transformations, not as an element of the spacetime algebra. The spinor metric (41.37) in the Dirac representation translates into the chiral representation (41.13) as $\varepsilon_{\text{chiral}} = X^{-\top} \varepsilon_{\text{Dirac}} X^{-1} = -iI\sigma_2$. However, the resulting chiral spinor metric $\varepsilon_{\text{chiral}}$ is imaginary. The chiral spinor metric can be made real by scaling it by a factor of i ,

$$\varepsilon_{\text{chiral}} = iX^{-\top} \varepsilon_{\text{Dirac}} X^{-1} = I\sigma_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} . \quad (41.38)$$

The normalization is chosen such that ε in either the Dirac or chiral representations is real and orthogonal. Its square is minus the unit matrix,

$$\varepsilon^{-1} = \varepsilon^\top , \quad \varepsilon^2 = -1 . \quad (41.39)$$

In both Dirac and chiral representations, commuting the spinor metric ε through the orthonormal basis vectors γ_m converts them to minus their transposes,

$$\gamma_m^\top \varepsilon = -\varepsilon \gamma_m . \quad (41.40)$$

The condition (41.36) implies that the spinor tensor ε is invariant under Lorentz transformations,

$$R : \varepsilon \rightarrow R^\top \varepsilon R = \varepsilon \tilde{R} R = \varepsilon . \quad (41.41)$$

The components of the spinor tensor ε define the antisymmetric spinor metric ε_{ab} ,

$$\epsilon_a^\top \varepsilon \epsilon_b = \varepsilon_{ab} . \quad (41.42)$$

Notice that the spinor metric tensor ε_{ab} is non-vanishing only between like-chiral indices ab .

41.6.2 Row basis spinors

It is convenient to use the symbol $\epsilon_a \cdot$ with a trailing dot, symbolic of the trailing ε , to denote the row spinor $\epsilon_a^\top \varepsilon$,

$$\epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon . \quad (41.43)$$

The motivation for the trailing dot notation is equation (41.47) below. The four row spinors

$$\epsilon_a \cdot = \{ \epsilon_{\uparrow\uparrow} \cdot, \epsilon_{\downarrow\downarrow} \cdot, \epsilon_{\downarrow\uparrow} \cdot, \epsilon_{\uparrow\downarrow} \cdot \} \quad (41.44)$$

provide a basis for row spinors. The boost and spin weights of the row basis spinors are in accord with their covariant indices: basis spinors with a $\uparrow$ index have boost weight $+\frac{1}{2}$, while basis spinors with a $\downarrow$ index have boost weight $-\frac{1}{2}$. Likewise basis spinors with a $\uparrow$ index have spin weight $+\frac{1}{2}$, while basis spinors with a $\downarrow$ index have spin weight $-\frac{1}{2}$. The row spinors $\epsilon_a \cdot$ Lorentz transform as

$$R : \epsilon_a \cdot \equiv \epsilon_a^\top \varepsilon \rightarrow \epsilon_a^\top R^\top \varepsilon = \epsilon_a^\top \varepsilon \tilde{R} = \epsilon_a \cdot \tilde{R} . \quad (41.45)$$

In the chiral representation (41.13) the row basis spinors $\epsilon_a \cdot$ are the row spinors

$$\epsilon_{\uparrow\uparrow} \cdot = (0 \ 1 \ 0 \ 0) , \quad \epsilon_{\downarrow\downarrow} \cdot = (-1 \ 0 \ 0 \ 0) , \quad \epsilon_{\downarrow\uparrow} \cdot = (0 \ 0 \ 0 \ 1) , \quad \epsilon_{\uparrow\downarrow} \cdot = (0 \ 0 \ -1 \ 0) . \quad (41.46)$$

41.6.3 Inner products of basis spinors

The product of the row spinor $\epsilon_a \cdot$ with the column spinor ϵ_b defines their inner product, or scalar product, which equals the **spinor metric** ε_{ab} in accordance with equation (41.42),

$$\boxed{\epsilon_a \cdot \epsilon_b = \varepsilon_{ab}} . \quad (41.47)$$

Equation (41.47) motivates the trailing dot notation for the row spinor. The scalar product is antisymmetric,

$$\epsilon_a \cdot \epsilon_b = -\epsilon_b \cdot \epsilon_a . \quad (41.48)$$

In the chiral representation, the nonzero components of the scalar product are explicitly, equation (41.38),

$$\epsilon_{\uparrow\uparrow} \cdot \epsilon_{\downarrow\downarrow} = -\epsilon_{\downarrow\downarrow} \cdot \epsilon_{\uparrow\uparrow} = 1, \quad \epsilon_{\downarrow\uparrow} \cdot \epsilon_{\uparrow\downarrow} = -\epsilon_{\uparrow\downarrow} \cdot \epsilon_{\downarrow\uparrow} = 1. \quad (41.49)$$

The scalar product (41.47) is a Lorentz scalar,

$$R: \epsilon_a \cdot \epsilon_b \rightarrow \epsilon_a \cdot \tilde{R} R \epsilon_b = \epsilon_a \cdot \epsilon_b. \quad (41.50)$$

Thus the spinor metric ε_{ab} is Lorentz invariant, just like the Minkowski metric η_{mn} .

41.6.4 Lowering and raising spinor indices

The antisymmetric spinor metric ε_{ab} is given in the chiral representation by equation (41.38). The inverse metric ε^{ab} is defined by $\varepsilon^{ab}\varepsilon_{bc} = \delta_c^a$. The spinor metric and its inverse satisfy

$$\varepsilon_{ab} = -\varepsilon_{ba} = -\varepsilon^{ab} = \varepsilon^{ba}. \quad (41.51)$$

Indices on a spinor tensor are lowered and raised by pre-multiplying by the metric ε_{ab} and its inverse ε^{ab} . The contravariant components ϵ^a of the column basis spinors, satisfying $\epsilon^a = \varepsilon^{ab}\epsilon_b$, are

$$\epsilon^{\uparrow\uparrow} = -\epsilon_{\downarrow\downarrow}, \quad \epsilon^{\downarrow\downarrow} = \epsilon_{\uparrow\uparrow}, \quad \epsilon^{\downarrow\uparrow} = \epsilon_{\uparrow\downarrow}, \quad \epsilon^{\uparrow\downarrow} = -\epsilon_{\downarrow\uparrow}. \quad (41.52)$$

For example, $\epsilon^{\uparrow\uparrow} = \varepsilon^{\uparrow\uparrow\downarrow\downarrow}\epsilon_{\downarrow\downarrow} = -\epsilon_{\downarrow\downarrow}$. Post-multiplying by the metric or its inverse yields a result of opposite sign, $\epsilon^a = \varepsilon^{ab}\epsilon_b = -\epsilon_b\varepsilon^{ba}$. The contravariant components ϵ^a of the row basis spinors satisfy the same relations (41.52) with a trailing dot appended on left and right hand sides. The scalar products of contravariant with covariant basis spinors form the unit matrix,

$$\epsilon^a \cdot \epsilon_b = -\epsilon_b \cdot \epsilon^a = \delta_b^a. \quad (41.53)$$

The scalar products of contravariant basis spinors are

$$\epsilon^a \cdot \epsilon^b = -\epsilon^b \cdot \epsilon^a = -\varepsilon^{ab}. \quad (41.54)$$

41.6.5 Scalar products of Dirac spinors

A general row Dirac spinor $\psi \cdot$ is a complex (with respect to i) linear combination of the 4 row basis spinors

$$\boxed{\psi \cdot \equiv \psi^\top \varepsilon = \psi^a \epsilon_a \cdot}. \quad (41.55)$$

It Lorentz transforms as

$$R: \psi \cdot \rightarrow \psi \cdot \tilde{R}. \quad (41.56)$$

A row spinor $\psi \cdot$ transforms like a reverse rotor.

The scalar product of a row spinor $\psi \cdot = \psi^a \epsilon_a \cdot$ with a column spinor $\chi = \chi^a \epsilon_a$ may be written variously

$$\psi \cdot \chi = \psi^\top \varepsilon \chi = \psi^a \epsilon_a \cdot \chi^b \epsilon_b = \varepsilon_{ab} \psi^a \chi^b = \psi^a \chi_a = -\psi_a \chi^a = -\varepsilon^{ab} \psi_a \chi_b. \quad (41.57)$$

Notice that when the scalar product $\psi \cdot \chi$ is written in the contracted form $\psi^a \chi_a$, the first index is raised and the second is lowered. An additional minus sign appears if the first index is lowered and the second is raised. Flipping the indices on the expansion $\psi^a \epsilon_a$ of a spinor in components similarly changes the sign,

$$\psi = \psi^a \epsilon_a = \psi^a \epsilon_{ab} \epsilon^b = -\psi_a \epsilon^a . \quad (41.58)$$

The components ψ^a of a column spinor ψ can be projected out by pre-multiplying by the row basis spinor ϵ^a ,

$$\epsilon^a \cdot \psi = \epsilon^a \cdot \psi^b \epsilon_b = \delta_b^a \psi^b = \psi^a . \quad (41.59)$$

The components ψ^a of a row spinor $\psi \cdot$ can be projected out by post-multiplying by minus the column basis spinor ϵ^a ,

$$-\psi \cdot \epsilon^a = -\psi^b \epsilon_b \cdot \epsilon^a = \delta_b^a \psi^b = \psi^a . \quad (41.60)$$

41.7 Super spacetime algebra

41.7.1 Outer products of basis spinors

A row spinor $\epsilon_a \cdot$ multiplied by a column spinor ϵ_b yields their scalar product. In the opposite order, a column spinor ϵ_a multiplied by a row spinor $\epsilon_b \cdot$ yields their outer product. The outer product $\epsilon_a \epsilon_b \cdot$ Lorentz transforms like a multivector in the spacetime algebra,

$$R : \epsilon_a \epsilon_b \cdot \equiv \epsilon_a \epsilon_b^\top \varepsilon \rightarrow R \epsilon_a \epsilon_b^\top R^\top \varepsilon = R \epsilon_a \epsilon_b^\top \varepsilon \tilde{R} = R \epsilon_a \epsilon_b \cdot \tilde{R} . \quad (41.61)$$

The trailing dot on the outer product $\epsilon_a \epsilon_b \cdot$ is symbolic of the trailing ε , necessary to convert the spinor tensor $\epsilon_a \epsilon_b^\top$ into an object that transforms like a multivector.

The products of the 4 column basis spinors ϵ_a with the 4 row basis spinors $\epsilon_b \cdot$ form 16 outer products. All 16 outer products are non-vanishing, and their algebra is isomorphic to the 4D spacetime algebra of multivectors. Unlike the spacetime algebra, the outer product contains both antisymmetric and symmetric products.

The 16 outer products divide into 8 outer products of spinors of like chirality (two right, or two left), and 8 outer products of spinors of opposite chirality (one right, one left). The outer products of spinors of like chirality yield the 8 even-grade elements of the spacetime algebra, while outer products of spinors of opposite chirality yield the 8 odd-grade elements of the spacetime algebra. The 8 even elements preserve chirality (they transform a spinor of given chirality to another of like chirality), while the 8 odd elements flip chirality (they transform a spinor of given chirality to another of opposite chirality).

In the chiral representation (41.13), the 8 outer products of basis spinors of like chirality map to even multivectors of the spacetime algebra as follows. The boost and spin weights of the left and right hand sides of each of equations (41.62)–(41.65) below match, as they should. The antisymmetric outer products of right-handed spinors form a right-handed scalar singlet,

$$[\epsilon_{\Downarrow\Downarrow}, \epsilon_{\Uparrow\Uparrow}] \cdot = \frac{1}{2}(1 + \gamma_5) . \quad (41.62)$$

The trailing dot on the commutator indicates that the right partner of each product is a row spinor, $[\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\uparrow}] \cdot = \epsilon_{\downarrow\downarrow} \epsilon_{\uparrow\uparrow} \cdot - \epsilon_{\uparrow\uparrow} \epsilon_{\downarrow\downarrow} \cdot$. Similarly the antisymmetric outer products of left-handed spinors form a left-handed scalar singlet,

$$[\epsilon_{\uparrow\downarrow}, \epsilon_{\downarrow\uparrow}] \cdot = \frac{1}{2}(1 - \gamma_5) . \quad (41.63)$$

The symmetric outer products of right-handed spinors form a triplet of right-handed bivectors,

$$\{\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\uparrow}\} \cdot = \gamma_{v+} , \quad \{\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\downarrow}\} \cdot = \frac{1}{2}(\gamma_{vu} - \gamma_{+-}) , \quad \{\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\downarrow}\} \cdot = -\gamma_{u-} . \quad (41.64)$$

The symmetric outer products of left-handed spinors form a triplet of left-handed bivectors,

$$\{\epsilon_{\downarrow\uparrow}, \epsilon_{\downarrow\uparrow}\} \cdot = -\gamma_{u+} , \quad \{\epsilon_{\downarrow\uparrow}, \epsilon_{\uparrow\downarrow}\} \cdot = -\frac{1}{2}(\gamma_{vu} + \gamma_{+-}) , \quad \{\epsilon_{\uparrow\downarrow}, \epsilon_{\uparrow\downarrow}\} \cdot = \gamma_{v-} . \quad (41.65)$$

In the chiral representation (41.13), the 8 outer products of basis spinors of opposite chirality map to odd multivectors of the spacetime algebra as follows. Again, the boost and spin weights of the left and right hand sides of each of equations (41.66)–(41.67) below match, as they should. The 4 symmetric outer products of right- with left-handed spinors yield the 4 Newman-Penrose basis vectors,

$$\{\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\downarrow}\} \cdot = -\frac{1}{\sqrt{2}}\gamma_v , \quad \{\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}\} \cdot = \frac{1}{\sqrt{2}}\gamma_u , \quad \{\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\uparrow}\} \cdot = \frac{1}{\sqrt{2}}\gamma_+ , \quad \{\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\downarrow}\} \cdot = -\frac{1}{\sqrt{2}}\gamma_- . \quad (41.66)$$

The 4 antisymmetric outer products of right- with left-handed spinors yield the 4 Newman-Penrose basis pseudovectors,

$$[\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\downarrow}] \cdot = -\frac{1}{\sqrt{2}}\gamma_5\gamma_v , \quad [\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}] \cdot = \frac{1}{\sqrt{2}}\gamma_5\gamma_u , \quad [\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\uparrow}] \cdot = \frac{1}{\sqrt{2}}\gamma_5\gamma_+ , \quad [\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\downarrow}] \cdot = -\frac{1}{\sqrt{2}}\gamma_5\gamma_- . \quad (41.67)$$

The trace of the outer product of a pair of basis spinors gives their scalar product (note that the 1 on the right hand sides of equations (41.62) and (41.63) is the unit matrix, whose trace is 4),

$$\text{Tr } \epsilon_a \epsilon_b \cdot = \epsilon_b \cdot \epsilon_a = \varepsilon_{ba} . \quad (41.68)$$

The expansion of the 16 outer products $\epsilon_a \epsilon_b \cdot$ of spinors in terms of the 16 basis elements γ_M of the spacetime algebra, and vice versa, define the matrix of coefficients γ_{ab}^M and its inverse γ_M^{ab} ,

$$\epsilon_a \epsilon_b \cdot = \gamma_{ab}^M \gamma_M , \quad \gamma_M = \gamma_M^{ab} \epsilon_a \epsilon_b \cdot . \quad (41.69)$$

The coefficients γ_{ab}^M and γ_M^{ab} are

$$\gamma_{ab}^M = \frac{1}{4} \epsilon_b \cdot \gamma^M \epsilon_a , \quad \gamma_M^{ab} = -\epsilon^a \cdot \gamma_M \epsilon^b . \quad (41.70)$$

The coefficients in the chiral representation in terms of Newman-Penrose basis multivectors can be read off from equations (41.62)–(41.67), and are all real.

Exercise 41.2 Consistency of spinor and multivector scalar products. Confirm that the spinor and multivector scalar products are consistent. This exercise is similar to Exercise 40.1.

Solution. Multivector vectors γ_m are equivalent to outer products of Dirac spinors in accordance with equations (41.66) and (41.66). For example, the scalar product of the multivectors γ_v and γ_u is

$$\begin{aligned}
 -\gamma_v \cdot \gamma_u &= -\frac{1}{2} (\gamma_v \gamma_u + \gamma_u \gamma_v) \\
 &= \{\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\downarrow}\} \cdot \{\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}\} \cdot + \{\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}\} \cdot \{\epsilon_{\uparrow\uparrow}, \epsilon_{\uparrow\downarrow}\} \cdot \\
 &= \epsilon_{\uparrow\uparrow}(\epsilon_{\uparrow\downarrow} \cdot \epsilon_{\downarrow\uparrow})\epsilon_{\downarrow\downarrow} \cdot + \epsilon_{\uparrow\downarrow}(\epsilon_{\uparrow\uparrow} \cdot \epsilon_{\downarrow\downarrow})\epsilon_{\downarrow\uparrow} \cdot + \epsilon_{\downarrow\downarrow}(\epsilon_{\downarrow\uparrow} \cdot \epsilon_{\uparrow\downarrow})\epsilon_{\uparrow\uparrow} \cdot + \epsilon_{\downarrow\uparrow}(\epsilon_{\downarrow\downarrow} \cdot \epsilon_{\uparrow\uparrow})\epsilon_{\uparrow\downarrow} \cdot \\
 &= -\epsilon_{\uparrow\uparrow}\epsilon_{\downarrow\downarrow} \cdot + \epsilon_{\uparrow\downarrow}\epsilon_{\downarrow\uparrow} \cdot + \epsilon_{\downarrow\downarrow}\epsilon_{\uparrow\uparrow} \cdot - \epsilon_{\downarrow\uparrow}\epsilon_{\uparrow\downarrow} \cdot \\
 &= [\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\uparrow}] \cdot + [\epsilon_{\uparrow\downarrow}, \epsilon_{\downarrow\uparrow}] \cdot \\
 &= \frac{1}{2}(1 + \gamma_5) + \frac{1}{2}(1 - \gamma_5) \\
 &= 1 ,
 \end{aligned} \tag{41.71}$$

the fourth step of which invokes the spinor scalar product (41.49), and the penultimate step is from the equivalences (41.62) and (41.63). Similarly,

$$\begin{aligned}
 \gamma_+ \cdot \gamma_- &= \frac{1}{2} (\gamma_+ \gamma_- + \gamma_- \gamma_+) \\
 &= -(\{\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\uparrow}\} \cdot \{\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\downarrow}\} \cdot + \{\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\downarrow}\} \cdot \{\epsilon_{\uparrow\uparrow}, \epsilon_{\downarrow\uparrow}\} \cdot) \\
 &= -(\epsilon_{\uparrow\uparrow}(\epsilon_{\downarrow\uparrow} \cdot \epsilon_{\uparrow\downarrow})\epsilon_{\downarrow\downarrow} \cdot + \epsilon_{\downarrow\uparrow}(\epsilon_{\uparrow\uparrow} \cdot \epsilon_{\downarrow\downarrow})\epsilon_{\uparrow\downarrow} \cdot + \epsilon_{\downarrow\downarrow}(\epsilon_{\uparrow\downarrow} \cdot \epsilon_{\downarrow\uparrow})\epsilon_{\uparrow\uparrow} \cdot + \epsilon_{\uparrow\downarrow}(\epsilon_{\downarrow\downarrow} \cdot \epsilon_{\uparrow\uparrow})\epsilon_{\downarrow\uparrow} \cdot) \\
 &= -\epsilon_{\uparrow\uparrow}\epsilon_{\downarrow\downarrow} \cdot - \epsilon_{\downarrow\uparrow}\epsilon_{\uparrow\downarrow} \cdot + \epsilon_{\downarrow\downarrow}\epsilon_{\uparrow\uparrow} \cdot + \epsilon_{\uparrow\downarrow}\epsilon_{\downarrow\uparrow} \cdot \\
 &= [\epsilon_{\downarrow\downarrow}, \epsilon_{\uparrow\uparrow}] \cdot + [\epsilon_{\downarrow\uparrow}, \epsilon_{\uparrow\downarrow}] \cdot \\
 &= \frac{1}{2}(1 + \gamma_5) + \frac{1}{2}(1 - \gamma_5) \\
 &= 1 .
 \end{aligned} \tag{41.72}$$

The results agrees with the multivector scalar product (41.4).

Concept question 41.3 Chiral scalar. A scalar field has no spin. How then can a scalar field have chirality? **Answer.** A chiral scalar is a sum of a scalar and a pseudoscalar. For example, a right-handed chiral scalar is

$$\varphi_R = \frac{1}{2}(1 + \gamma_5)\varphi , \tag{41.73}$$

where φ is a complex scalar. The chiral operator γ_5 is not a scalar, but rather a totally antisymmetric tensor of rank 4. The Newman-Penrose expression (41.117) for γ_5 shows that it has zero boost and spin weight.

41.7.2 The 4D super spacetime algebra

The super spacetime algebra comprises 4 distinct species of objects: true scalars, column spinors, row spinors, and multivectors. In a matrix representation, they are complex (with respect to i) matrices with dimensions

1×1 , 1×4 , 4×1 , and 4×4 . The super spacetime algebra consists of arbitrary sums and products of all 4 species.

A true scalar, the value of a scalar product of spinors, is a complex number. A column spinor ψ is a complex linear combination of column basis spinors ϵ_a ,

$$\psi = \psi^a \epsilon_a , \quad (41.74)$$

while a row spinor $\psi \cdot$ is a complex linear combination of row basis spinors $\epsilon_a \cdot$,

$$\psi \cdot = \psi^a \epsilon_a \cdot . \quad (41.75)$$

A multivector $\mathbf{a}$ is a complex linear combination of outer products of the column and row basis spinors,

$$\mathbf{a} = a^{ab} \epsilon_a \epsilon_b \cdot . \quad (41.76)$$

Linearity and the transformation law (41.61) imply that the algebra of sums and products of outer products of spinors is isomorphic to the spacetime algebra.

As seen in §41.6.3 and §41.7.1, a column spinor ψ and a row spinor $\chi \cdot$ can be multiplied in either order, yielding an inner product which is a true scalar, and an outer product which is a multivector. However, a column spinor cannot be multiplied by a column spinor, and likewise a row spinor cannot be multiplied by a row spinor, as is manifestly true in a matrix representation. Rather than prohibit multiplication, it is advantageous, because it permits interpretation of the column and row spinors as destruction and creation operators in quantum field theory, §44.4.2, to assert that the product of a column spinor with a column spinor is zero, and the product of a row spinor with a row spinor is zero,

$$\psi \chi = 0 , \quad \psi \cdot \chi \cdot = 0 . \quad (41.77)$$

Similarly, a multivector $\mathbf{a}$ can only pre-multiply a column spinor ψ , and can only post-multiply a row spinor $\psi \cdot$, as is again manifestly true in a matrix representation. Thus a multivector $\mathbf{a}$ post-multiplying a column spinor or pre-multiplying a row spinor yields zero,

$$\psi \mathbf{a} = 0 , \quad \mathbf{a}(\psi \cdot) = 0 . \quad (41.78)$$

In general, a sequence of products of spinors yields a nonzero result provided that they alternate between column spinor and row spinor,

$$\psi \chi \cdot \varphi \quad \text{or} \quad \psi \cdot \chi \varphi \cdot . \quad (41.79)$$

Both product sequences are associative,

$$\psi \chi \cdot \varphi = (\psi \chi \cdot) \varphi = \psi (\chi \cdot \varphi) , \quad (41.80a)$$

$$\psi \cdot \chi \varphi \cdot = (\psi \cdot \chi) \varphi \cdot = \psi \cdot (\chi \varphi \cdot) . \quad (41.80b)$$

One of the advantages of the trailing dot notation is that it makes the directionality of spinor multiplication, and the corresponding associative law, transparent. A multivector $\mathbf{a}$ is equivalent to an outer product of spinors, so products such as

$$\psi \cdot \mathbf{a} \chi \quad (41.81)$$

are admissible, and in general non-vanishing.

The trace of an outer product of spinors is a true scalar

$$\text{Tr } \psi \chi \cdot = \psi^a \chi^b \varepsilon_{ba} = -\psi \cdot \chi = \chi \cdot \psi . \quad (41.82)$$

41.7.3 Fierz identities

The associative law and the scalar product make it straightforward to simplify long strings of products of spinors and multivectors, a process known in quantum field theory as Fierz rearrangement.

Let $\mathbf{a} = a^{ab} \epsilon_a \epsilon_b \cdot$ and $\mathbf{b} = b^{ab} \epsilon_a \epsilon_b \cdot$ be two multivectors expressed as a sum of outer products of spinors. Their product is the multivector

$$\mathbf{ab} = a^{ab} \epsilon_a \epsilon_b \cdot b^{cd} \epsilon_c \epsilon_d \cdot = \epsilon_a a^{ab} \varepsilon_{bc} b^{cd} \epsilon_d \cdot = \epsilon_a a^{ab} b_b^{\ d} \epsilon_d \cdot . \quad (41.83)$$

A sequence of multivectors sandwiched by spinors simplifies as

$$\psi \cdot \mathbf{ab} \chi = \psi^a \epsilon_a \cdot a^{bc} \epsilon_b \epsilon_c \cdot b^{de} \epsilon_d \epsilon_e \cdot \chi^f \epsilon_f = \psi^a \varepsilon_{ab} a^{bc} \varepsilon_{cd} b^{de} \varepsilon_{ef} \chi^f = \psi^a a_a^{\ c} b_c^{\ e} \chi_e . \quad (41.84)$$

41.8 Charge conjugation C

The super spacetime algebra possesses a discrete transformation, called charge conjugation (or simply conjugation, when there is no ambiguity), denoted C , that transforms a particle into its antiparticle (Bjorken and Drell, 1964, §5.2). The charge-conjugate Dirac spinor $\bar{\psi}$ is defined by equation (41.94) below (Bjorken and Drell (1964) denote the charge conjugate by ψ_c). The conjugate spinor $\bar{\psi}$ has the defining properties that (a) its components are complex conjugates of those of the parent spinor ψ , and (b) it Lorentz transforms in the same way as ψ .

41.8.1 Conjugation operator C

Consider the conjugation operator C with the defining property that commutation with it converts any Lorentz rotor R in the chiral or Dirac representations to its complex conjugate (with respect to i),

$$CR^* = RC . \quad (41.85)$$

Note that the complex conjugate R^* of a Lorentz rotor R is also a Lorentz rotor, since a Lorentz rotor R is a real (with respect to i) linear combination of even orthonormal basis multivectors of the spacetime algebra. In the Dirac representation (14.103), a necessary and sufficient condition for (41.85) to hold is that C commutes with I and σ_2 , and anticommutes with σ_1 and σ_3 . This requires that in the Dirac representation C_{Dirac} is proportional to σ_2 with a proportionality factor that could be some arbitrary complex (with respect to i and/or I) number. To be consistent with standard Dirac theory, the conjugation operator C_{Dirac} is taken to

be the real unitary matrix

$$C_{\text{Dirac}} = -\sigma_2 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}. \quad (41.86)$$

The conjugation matrix (41.86) in the Dirac representation translates into the chiral representation (41.13) as $C_{\text{chiral}} = XC_{\text{Dirac}}X^{-*}$, which happens to be the same matrix as in the Dirac representation, $C_{\text{chiral}} = C_{\text{Dirac}}$. However, to compensate for the extra factor of i introduced into the definition (41.38) of the chiral spinor metric $\varepsilon_{\text{chiral}}$, it is necessary to introduce an extra factor of $-i$ in the definition of the chiral conjugation matrix C_{chiral} ,

$$C_{\text{chiral}} = -iXC_{\text{Dirac}}X^{-*} = -iC_{\text{Dirac}} = \begin{pmatrix} 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix}. \quad (41.87)$$

The compatibility of the normalizations of ε and C is necessary to ensure that the scalar product $\bar{\psi} \cdot \psi$ of a spinor with its conjugate is real, equation (41.151). Note that the conjugation matrix (41.87) in the chiral representation (41.13) is $C_{\text{chiral}} = iI\gamma_2$, not $-\sigma_2$.

Notwithstanding the equality of C and $-\sigma_2$ in the Dirac representation, or $-I\gamma_2$ in the chiral representation, the conjugation operator C is defined to transform not as an element of the geometric algebra, but rather as

$$R: C \rightarrow RCR^{-*}, \quad (41.88)$$

in accordance with the defining condition (41.85). Note that if the Lorentz rotor R were unitary, then RCR^{-*} would equal $RCR^{\top}$; but although spatial rotations are unitary, Lorentz boosts are not. The spinor tensor that Lorentz transforms as (41.88) and remains invariant under that transformation is precisely the conjugation operator C .

In both Dirac and chiral representations (41.86) and (41.87), the conjugation operator is symmetric and unitary,

$$C^{\top} = C, \quad CC^{\dagger} = CC^* = 1. \quad (41.89)$$

In both Dirac and chiral representations, commuting the conjugation operator C through the orthonormal basis vectors γ_m converts them to their complex (with respect to i) conjugates,

$$C\gamma_m = \gamma_m^*C. \quad (41.90)$$

In both Dirac and chiral representations, commuting C through the spinor metric ε converts the former to minus its complex conjugate,

$$C\varepsilon = -\varepsilon C^*. \quad (41.91)$$

41.8.2 Conjugate spinor

The complex conjugate ψ^* of a Dirac spinor $\psi = \psi^a \epsilon_a$ is defined to be the spinor with complex-conjugated (with respect to i) coefficients in the Dirac or chiral matrix representation of the spinor,

$$\psi^* \equiv \psi^{a*} \epsilon_a . \quad (41.92)$$

In effect, the basis spinors ϵ_a are taken to be real in the Dirac or chiral representations. The operation (41.92) of complex conjugation of a spinor is representation-dependent, as is evident from the fact that the unitary matrix X , equation (41.12), that transforms between Dirac and chiral representations is complex. By contrast, the conjugation operation (41.94) below is representation-independent. Complex conjugation leaves the boost and spin of a spinor unchanged. Since a spinor ψ Lorentz transforms under a Lorentz rotor R as $\psi \rightarrow R\psi$, its complex conjugate ψ^* transforms according to the complex-conjugate representation of the γ -matrices,

$$R : \psi^* \rightarrow (R\psi)^* = R^* \psi^* . \quad (41.93)$$

The conjugate Dirac spinor $\bar{\psi}$ is defined by

$$\boxed{\bar{\psi} \equiv C\psi^*} , \quad (41.94)$$

where C is the conjugation operator defined in the Dirac or chiral representations by equations (41.86) and (41.87). The conjugation operator C is by construction Lorentz invariant, so the conjugate spinor $\bar{\psi}$ Lorentz transforms as

$$R : \bar{\psi} \equiv C\psi^* \rightarrow CR^*\psi^* = RC\psi^* = R\bar{\psi} , \quad (41.95)$$

that is, the conjugate spinor $\bar{\psi}$ Lorentz transforms in the same way as the spinor ψ . The middle expression of equation (41.95) is $CR^*\psi^* = C(R\psi)^* = \overline{R\psi}$, so

$$\overline{R\psi} = R\bar{\psi} , \quad (41.96)$$

that is, the operations of conjugation and Lorentz transformation commute.

The symmetry of the conjugation operator, $C = C^\top$, implies that conjugating a Dirac spinor ψ twice recovers the original spinor,

$$\bar{\bar{\psi}} = C(C\psi^*)^* = CC^*\psi = CC^{-\top}\psi = \psi . \quad (41.97)$$

If $\psi = \psi^a \epsilon_a$, then the conjugate spinor $\bar{\psi}$ is

$$\bar{\psi} = \psi^{a*} \bar{\epsilon}_a , \quad (41.98)$$

where the conjugate basis spinors $\bar{\epsilon}_a$ are defined by

$$\bar{\epsilon}_a \equiv C\epsilon_a . \quad (41.99)$$

In the Dirac representation the conjugate basis spinors $\bar{\epsilon}_a$ are, from the expression (41.86) for C ,

$$\{\bar{\epsilon}_{\uparrow\uparrow}, \bar{\epsilon}_{\uparrow\downarrow}, \bar{\epsilon}_{\downarrow\uparrow}, \bar{\epsilon}_{\downarrow\downarrow}\} = \{-\epsilon_{\downarrow\downarrow}, \epsilon_{\downarrow\uparrow}, \epsilon_{\uparrow\downarrow}, -\epsilon_{\uparrow\uparrow}\} . \quad (41.100)$$

In the chiral representation the conjugate basis spinors $\bar{\epsilon}_a$ are, from the expression (41.87) for C ,

$$\{\bar{\epsilon}_{\uparrow\uparrow}, \bar{\epsilon}_{\downarrow\downarrow}, \bar{\epsilon}_{\downarrow\uparrow}, \bar{\epsilon}_{\uparrow\downarrow}\} = \{i\epsilon_{\uparrow\downarrow}, -i\epsilon_{\downarrow\uparrow}, -i\epsilon_{\downarrow\downarrow}, i\epsilon_{\uparrow\uparrow}\} . \quad (41.101)$$

Equation (41.101) shows that conjugation flips spin, but leaves boost unchanged.

41.8.3 Row conjugate spinor

In both Dirac (14.102) and chiral (41.13) representations, the row conjugate spinor $\bar{\psi}$ corresponding to the column conjugate spinor $\bar{\psi}$ is

$$\bar{\psi} \cdot \equiv \bar{\psi}^\top \varepsilon = \psi^\dagger C^\top \varepsilon = -i\psi^\dagger \gamma_0 . \quad (41.102)$$

The row conjugate spinor $\bar{\psi} \cdot$ is commonly called the adjoint spinor. The row conjugate spinor $\bar{\psi} \cdot$ equals the reverse spinor $\tilde{\psi}$ defined by equation (14.119). Note that the column conjugate spinor $\bar{\psi}$ is *not* the same as the conventional adjoint spinor $\tilde{\psi}$; rather the conventional adjoint spinor $\tilde{\psi}$ is the row conjugate spinor $\bar{\psi} \cdot$.

Equation (41.102) implies that

$$C^\top \varepsilon = -i\gamma_0 . \quad (41.103)$$

Equation (41.103) holds in both Dirac and chiral representations, but in fact it must be true in any satisfactory representation of Dirac spinors governed by the Dirac Lagrangian (44.31), in order that the Dirac number current density $n^0 \equiv i\bar{\psi} \cdot \gamma^0 \psi$, equation (44.20), equal a positive number $\psi^\dagger \psi$.

The spinor metric ε and conjugation operator C may be regarded as being defined by their actions (41.40) and (41.90) on the Minkowski basis vectors γ_m , namely $\gamma_m^\top = -\varepsilon \gamma_m \varepsilon^{-1}$ and $\gamma_m^* = C \gamma_m C^{-1}$. If equation (41.103) holds, as it must do, and if in addition C is symmetric, $C^\top = C$, as it must be if C is unitary and the double conjugate of a spinor is itself, as it is in Dirac representation, then the Hermitian conjugates of the basis vectors γ_m satisfy

$$\gamma_m^\dagger = (\gamma_m^\top)^* = -C \varepsilon \gamma_m \varepsilon^{-1} C^{-1} = -(C^\top \varepsilon) \gamma_m (C^\top \varepsilon)^{-1} = -(\gamma_0)^{-1} \gamma_m (\gamma_0) = \gamma_0 \gamma_m \gamma_0 = \gamma^m . \quad (41.104)$$

Equation (41.104) shows that γ_m are unitary, which is the condition (14.100) originally adopted for the Dirac γ -matrices.

41.8.4 Conjugate multivector

The complex conjugate (with respect to i) $\mathbf{a}^*$ of a multivector $\mathbf{a} = a^M \gamma_M$, of any grade, is defined to be, in either the Dirac or chiral representation (and in either an orthonormal or Newman-Penrose basis),

$$\mathbf{a}^* \equiv a^{M*} \gamma_M^* . \quad (41.105)$$

Since $\mathbf{a}$ Lorentz transforms under a Lorentz rotor R as $\mathbf{a} \rightarrow R \mathbf{a} \tilde{R}$, its complex conjugate $\mathbf{a}^*$ transforms according to the complex conjugate representation of the γ -matrices,

$$R : \mathbf{a}^* \rightarrow (R \mathbf{a} \tilde{R})^* = R^* \mathbf{a} \tilde{R}^* . \quad (41.106)$$

So defined, complex conjugation is multiplicative over multivectors and spinors,

$$(\mathbf{a}\psi)^* = \mathbf{a}^* \psi^* , \quad (41.107)$$

and consistent with the spacetime algebra in the sense that the complex conjugate of a multivector that is an outer product of spinors is the outer product of the complex conjugate spinors,

$$(\psi\chi\cdot)^* = \psi^* \chi^* \cdot . \quad (41.108)$$

Complex conjugation leaves the boost and spin of a multivector unchanged.

The conjugate multivector $\bar{\mathbf{a}}$ (not to be confused with the reverse multivector $\tilde{\mathbf{a}}$) of a multivector $\mathbf{a}$ is defined to be

$$\boxed{\bar{\mathbf{a}} \equiv C \mathbf{a}^* C^{-1}} . \quad (41.109)$$

The conjugate multivector $\bar{\mathbf{a}}$ Lorentz transforms in the same way as the parent multivector $\mathbf{a}$,

$$R : \bar{\mathbf{a}} \rightarrow C R^* \mathbf{a}^* \tilde{R}^* C^{-1} = R C \mathbf{a} C^{-1} \tilde{R} = R \bar{\mathbf{a}} \tilde{R} . \quad (41.110)$$

Conjugation is multiplicative over multivectors and spinors,

$$\overline{\mathbf{a}\psi} = \bar{\mathbf{a}} \bar{\psi} . \quad (41.111)$$

The conjugate of a multivector that is an outer product of spinors is minus the outer product of the conjugate spinors,

$$\overline{\psi\chi\cdot} = -\bar{\psi}\bar{\chi}\cdot . \quad (41.112)$$

The sign comes from the anticommutation of the conjugation operator with the spinor metric tensor, equation (41.91).

If $\mathbf{a} = a^M \boldsymbol{\gamma}_M$, then the conjugate multivector $\bar{\mathbf{a}}$ is

$$\bar{\mathbf{a}} = a^M \bar{\boldsymbol{\gamma}}_M , \quad (41.113)$$

where the conjugate basis elements $\bar{\boldsymbol{\gamma}}_M$ are, in either an orthonormal or Newman-Penrose basis, and in either the Dirac or chiral representations,

$$\bar{\boldsymbol{\gamma}}_M = C \boldsymbol{\gamma}_M^* C^{-1} . \quad (41.114)$$

The conjugates of orthonormal basis elements are equal to themselves,

$$\bar{\boldsymbol{\gamma}}_M = \boldsymbol{\gamma}_M \quad \text{orthonormal basis multivectors} . \quad (41.115)$$

But the conjugates of Newman-Penrose basis vectors are not equal to themselves. Just as conjugation flips the spin but not boost of a spinor ψ , so also conjugation flips the spin but not boost of a multivector $\mathbf{a}$. Conjugation flips the spin indices $+\leftrightarrow -$ of the Newman-Penrose basis vectors $\boldsymbol{\gamma}_m$, while leaving the boost indices v and u unchanged,

$$\bar{\boldsymbol{\gamma}}_v = \boldsymbol{\gamma}_v , \quad \bar{\boldsymbol{\gamma}}_u = \boldsymbol{\gamma}_u , \quad \bar{\boldsymbol{\gamma}}_+ = \boldsymbol{\gamma}_- , \quad \bar{\boldsymbol{\gamma}}_- = \boldsymbol{\gamma}_+ , \quad (41.116)$$

as can be verified by direct calculation from the matrices (41.15) and (41.87). This is true in general: the

conjugate of any Newman-Penrose basis multivector γ_M is obtained by flipping its spin indices $+\leftrightarrow -$. The chiral matrix γ_5 expressed in the Newman-Penrose tetrad is

$$\gamma_5 \equiv -iI = -\frac{i}{4!}\varepsilon^{vu+-}\gamma_v\gamma_u\gamma_+\gamma_- = -\gamma_v\wedge\gamma_u\wedge\gamma_+\wedge\gamma_- , \quad (41.117)$$

where the imaginary factor i in the definition of γ_5 cancels against the imaginary determinant of the transformation from Minkowski to Newman-Penrose tetrad, leaving a real factor in the rightmost expression of equations (41.117). Conjugation flips the sign of the chiral operator γ_5 ,

$$\bar{\gamma}_5 = -\gamma_5 . \quad (41.118)$$

41.8.5 Real multivector

Conventionally, a multivector $\mathbf{a} = a^M\gamma_M$ is said to be real if its conjugate is itself,

$$\bar{\mathbf{a}} = \mathbf{a} . \quad (41.119)$$

In an orthonormal basis, the conjugates of the basis elements are themselves, $\bar{\gamma}_M = \gamma_M$, and a multivector $\mathbf{a}$ is then real if and only if the coefficients a^M of its expansion $\mathbf{a} = a^M\gamma_M$ in the orthonormal basis are real. The pseudoscalar I is real. The chiral operator γ_5 is pure imaginary.

Most classical multivectors are real. For example, derivatives are real, Lorentz rotors are real, the classical electromagnetic field is real.

41.9 Anticommutation of Dirac spinors

The Dirac spinor Lagrangian (44.3) involves a mass term $m\bar{\psi}\cdot\psi$. The complex conjugate (with respect to i) of the Dirac mass term is, Exercise 41.4,

$$(m\bar{\psi}\cdot\psi)^* = -m\psi\cdot\bar{\psi} . \quad (41.120)$$

Requiring that the Dirac mass term be real, as required for a real Lagrangian, then imposes the condition that the scalar product of the Dirac spinors $\bar{\psi}$ and ψ be antisymmetric,

$$\bar{\psi}\cdot\psi = -\psi\cdot\bar{\psi} . \quad (41.121)$$

More generally, the scalar product of any two Dirac spinors is antisymmetric,

$$\psi\cdot\chi = -\chi\cdot\psi . \quad (41.122)$$

The spinor scalar product is non-vanishing only between like-chiral components. Since the conjugate of a right-handed chiral spinor is left-handed, and vice versa, the scalar product of a pure right- or left-handed spinor (a Weyl spinor) with its conjugate is necessarily zero,

$$\bar{\psi}\cdot\psi = \bar{\psi}\cdot I\psi = 0 \quad (\text{Weyl}) . \quad (41.123)$$

Thus Weyl spinors are necessarily massless.

If a Dirac spinor ψ is decomposed into its right- and left-handed chiral parts ψ_R and ψ_L , equation (41.32), then since conjugation flips chirality, the scalar product is non-vanishing only between like-chiral spinors. The scalar and pseudoscalar products of $\bar{\psi}$ and ψ are

$$\bar{\psi} \cdot \psi = \bar{\psi}_L \cdot \psi_R + \bar{\psi}_R \cdot \psi_L, \quad \bar{\psi} \cdot I\psi = i(\bar{\psi}_L \cdot \psi_R - \bar{\psi}_R \cdot \psi_L). \quad (41.124)$$

Note that $\bar{\psi}_L$ is right-handed, and $\bar{\psi}_R$ is left-handed.

Exercise 41.4 Complex conjugate of a product of spinors and multivectors.

1. What is the complex conjugate (with respect to i) of a product $\chi \cdot \mathbf{a}\psi$ of a row spinor χ , a multivector $\mathbf{a}$ of grade p , and a column spinor ψ ?
2. If $\mathbf{a}$ is a real multivector of grade p , is the product $\bar{\psi} \cdot \mathbf{a}\psi$ real or imaginary?

Solution.

1. The complex conjugate of $\chi \cdot \mathbf{a}\psi$ is

$$(\chi \cdot \mathbf{a}\psi)^* = \chi^* \cdot \mathbf{a}^* \psi^* = \bar{\chi}^\top C^{-\top} \varepsilon \mathbf{a}^* C^{-1} \bar{\psi} = -\bar{\chi}^\top \varepsilon C \mathbf{a}^* C^{-1} \bar{\psi} = -\bar{\chi} \cdot \bar{\mathbf{a}} \bar{\psi}. \quad (41.125)$$

The sign flip in the penultimate expression occurs because commuting the conjugation operator C through the spinor metric tensor ε converts C to minus its complex conjugate, equation (41.91). An alternative, equivalent expression follows from the antisymmetry of the spinor metric,

$$(\chi \cdot \mathbf{a}\psi)^* = -\bar{\chi} \cdot \bar{\mathbf{a}} \bar{\psi} = (\bar{\mathbf{a}} \bar{\psi}) \cdot \bar{\chi} = \bar{\psi} \bar{\mathbf{a}}^\top \cdot \bar{\chi} = (-)^p (-)^{[p/2]} \bar{\psi} \cdot \bar{\mathbf{a}} \bar{\chi}. \quad (41.126)$$

The first equality is equation (41.125), while the second equality is from the anticommutation of Dirac spinors, equation (41.122). The $(-)^p$ sign in the final expression comes from commuting a grade- p multivector through the spinor metric, equation (41.40), while the $(-)^{[p/2]}$ sign comes from the reversion needed to undo the transposition of a grade- p multivector. The overall $(-)^{p+[p/2]}$ sign is positive for scalars, trivectors, and pseudoscalars, negative for vectors and bivectors.

2. If the multivector $\mathbf{a}$ is real, $\bar{\mathbf{a}} = \mathbf{a}$, then the complex conjugate of $\bar{\psi} \cdot \mathbf{a}\psi$, is, from equation (41.126),

$$(\bar{\psi} \cdot \mathbf{a}\psi)^* = (-)^p (-)^{[p/2]} \bar{\psi} \cdot \mathbf{a}\psi. \quad (41.127)$$

Thus $\bar{\psi} \cdot \mathbf{a}\psi$ is real for scalars, trivectors, and pseudoscalars, imaginary for vectors and bivectors.

41.10 Discrete transformations P, T

Besides conjugation C , the super spacetime algebra contains two other discrete transformations, parity inversion P , and time reversal T . Parity and time reversal are improper Lorentz transformations, which preserve the Minkowski metric, but which cannot be obtained by any continuous Lorentz transformation starting from the identity. Parity and time-reversal are examples of the geometric algebra transformation of reflection through an axis, §13.6.

41.10.1 Parity inversion P

The parity inversion operation P reverses all the spatial axes, while keeping the time axis unchanged¹,

$$P : \boldsymbol{\gamma}_m \rightarrow P\boldsymbol{\gamma}_mP^{-1} = \begin{cases} \boldsymbol{\gamma}_m & m = 0 , \\ -\boldsymbol{\gamma}_m & m = 1, 2, 3 . \end{cases} \quad (41.128)$$

Parity reversal transforms a Dirac spinor ψ as

$$P : \psi \rightarrow P\psi . \quad (41.129)$$

In any representation, the transformation (41.128) requires P to commute with the time axis $\boldsymbol{\gamma}_0$ and anticommute with the spatial axes $\boldsymbol{\gamma}_a$, $a = 1, 2, 3$. The only basis element of the spacetime algebra with the required (anti)commutation properties is the time vector $\boldsymbol{\gamma}_0$, so P must equal $\boldsymbol{\gamma}_0$ up to a possible scalar normalization:

$$P = \boldsymbol{\gamma}_0 . \quad (41.130)$$

If desired, a scalar factor of i could be inserted, $P = i\boldsymbol{\gamma}_0$, so that $P^2 = 1$, but the choice of phase factor is not essential. Parity flips boost $\uparrow \leftrightarrow \downarrow$ while leaving spin unchanged. This makes some physical sense: flipping boost flips the direction of the momentum of the spinor; while spin is a form of angular momentum, which is unchanged by parity inversion. Parity flips chirality, the projection of spin along the direction of momentum.

41.10.2 Time reversal T

The time-reversal operation T reverses the time axis, while keeping all the spatial axes unchanged,

$$T : \boldsymbol{\gamma}_m \rightarrow T\boldsymbol{\gamma}_mT^{-1} = \begin{cases} -\boldsymbol{\gamma}_m & m = 0 , \\ \boldsymbol{\gamma}_m & m = 1, 2, 3 . \end{cases} \quad (41.131)$$

Time reversal transforms a Dirac spinor ψ as

$$T : \psi \rightarrow T\psi . \quad (41.132)$$

In any representation, the transformation (41.131) requires T to anticommute with the time axis $\boldsymbol{\gamma}_0$ and commute with the spatial axes $\boldsymbol{\gamma}_a$, $a = 1, 2, 3$. The only basis element of the spacetime algebra with the required (anti)commutation properties is the time pseudovector $I\boldsymbol{\gamma}_0$, so T must equal that pseudovector up to a possible scalar normalization:

$$T = I\boldsymbol{\gamma}_0 . \quad (41.133)$$

If desired, a scalar factor of $-i$ could be inserted, $T = -iI\boldsymbol{\gamma}_0$, to ensure that $T^2 = 1$ and $PT = I$ (with $P = i\boldsymbol{\gamma}_0$), but again the choice of phase factor is not essential.

¹ Defining parity inversion as a reversal of all spatial axes is convenient when the number of spatial dimensions is odd, as here. In general the spatial rotation group splits into two disjoint parts, a proper group connected continuously to the identity, and an improper group obtained by a reflection through any one spatial axis and a continuous rotation. Parity inversion can be achieved by reflecting through any odd number of spatial axes.

41.10.3 PT

The product PT of the parity and time inversion operators,

$$PT = I , \quad (41.134)$$

reverses all 4 spacetime axes γ_m ,

$$PT : \gamma_m \rightarrow I\gamma_m I^{-1} = -\gamma_m , \quad (41.135)$$

and transforms a Dirac spinor ψ as

$$PT : \psi \rightarrow I\psi . \quad (41.136)$$

The fact that the PT operator equals the pseudoscalar I makes physical sense. The operation of reversing all axes, both space and time, is Lorentz invariant. The only Lorentz-invariant basis multivectors of the spacetime algebra are the unit matrix and the pseudoscalar. The pseudoscalar is related to the chiral matrix by $I = i\gamma_5$, so the basis spinors ϵ_a in the chiral representation are PT -eigenstates.

41.11 The supergeometric algebra in arbitrarily many space and time dimensions

Exercise 41.5 Generalize the super spacetime algebra to an arbitrary number of space and time dimensions. Generalize the super spacetime algebra to an arbitrary number of dimensions, with K spatial dimensions, and M timelike dimensions, and a total of $K+M = N$ dimensions. This is a generalization of Exercise 40.3.

Solution. The construction described in Exercise 40.3, in which all dimensions are spatial, carries through unchanged through parts 1–13. After the construction is completed, modify the matrix representing any timelike orthonormal basis vector γ_m by multiplying the matrix by $-i$ (or i , if preferred),

$$\gamma_m \rightarrow \mp i\gamma_m \quad \text{for timelike orthonormal basis vectors } \gamma_m . \quad (41.137)$$

Propagate that modification through the basis orthonormal multivectors of the spacetime algebra. The spinor metric ε can be left unchanged, so that it remains real.

As an example of this algorithm, the chiral basis vectors γ_i and $\gamma_{\bar{i}}$ for $i = 1 \dots [N/2]$ continue to be defined in terms of orthonormal vectors $\gamma_i^\pm$ by the unchanged equations (40.80), $\gamma_i = \frac{1}{\sqrt{2}}(\gamma_i^+ + i\gamma_i^-)$ and $\gamma_{\bar{i}} = \frac{1}{\sqrt{2}}(\gamma_i^+ - i\gamma_i^-)$. The chiral construction in part 5 of Exercise 40.3 yields unchanged real matrix representations of all chiral basis vectors γ_i and $\gamma_{\bar{i}}$. If in fact γ_i^- (say) is timelike, then replacing $i\gamma_i^- \rightarrow \gamma_i^-$ (after the construction is completed) means that $\gamma_i = \frac{1}{\sqrt{2}}(\gamma_i^+ + \gamma_i^-)$ and $\gamma_{\bar{i}} = \frac{1}{\sqrt{2}}(\gamma_i^+ - \gamma_i^-)$ are really sums of spacelike and timelike vectors γ_i^+ and γ_i^- , like the null vectors γ_v and γ_u in the Newman-Penrose formalism.

A supergeometric algebra with both space and time dimensions differs from a supergeometric algebra with only space (or only time) dimensions in that rotations in a time-space plane are non-compact, whereas

rotations in a space-space (or time-time) plane are compact. Rotations in a time-space plane are called (Lorentz) boosts. For example, if γ_i^+ is spacelike and γ_i^- is timelike, then a rotation by boost angle (rapidity) θ in the $\gamma_i^+ \gamma_i^-$ -plane transforms the the i th pair of chiral basis vectors γ_i and $\gamma_{\bar{i}}$ as, in place of (40.81),

$$\gamma_i \rightarrow e^\theta \gamma_i, \quad \gamma_{\bar{i}} \rightarrow e^{-\theta} \gamma_{\bar{i}}, \quad (41.138)$$

and a basis spinor ϵ_a transforms as, in place of (40.84),

$$\epsilon_{...i...} \rightarrow e^{\theta/2} \epsilon_{...i...}, \quad \epsilon_{...\bar{i}...} \rightarrow e^{-\theta/2} \epsilon_{...\bar{i}...}. \quad (41.139)$$

The chiral representation (41.13) of Dirac γ -matrices is equivalent to the chiral construction in part 5 of Exercise 40.3 with the following rearrangement of indices:

$$\{\gamma_1, \gamma_2, \gamma_3, \gamma_0\}_{\text{Dirac}} = \{\gamma_1^+, \gamma_1^-, \gamma_2^+, i\gamma_2^-\}, \quad (41.140a)$$

$$\{\gamma_+, \gamma_-, \gamma_v, \gamma_u\}_{\text{chiral}} = \{\gamma_1, \gamma_{\bar{1}}, \gamma_2, -\gamma_2\}. \quad (41.140b)$$

9. **Parity and time reversal.** Parity reversal is the operation of reflecting an odd number of spatial axes. Time reversal is the operation of reflecting an odd number of time axes. A reflection of an even number of spatial axes can be accomplished by a continuous rotation in spatial dimensions, while a reflection of an even number of time axes can be accomplished by a continuous rotation in time dimensions.

If the total number $N = K + M$ of spacetime dimensions is even, then parity reversal may be accomplished by setting the parity operator P equal to one of the space dimensions if K is even, or to one of the time dimensions if K is odd, and transforming spinors and multivectors by

$$P: \quad \psi \rightarrow P\psi, \quad \mathbf{a} \rightarrow P\mathbf{a}P^{-1}. \quad (41.141)$$

If desired, a phase factor can be inserted into P to ensure that $P^2 = 1$, but the choice of phase factor is not essential. Again, if the total number $N = K + M$ of spacetime dimensions is even, then the combined operation of parity and time reversal may be accomplished by setting the PT operator equal to the pseudoscalar I_N ,

$$PT: \quad \psi \rightarrow I_N\psi, \quad \mathbf{a} \rightarrow I_N\mathbf{a}I_N^{-1}. \quad (41.142)$$

Time reversal is accomplished by the operator $T = P(PT) = PI_N$.

As in part 9 of Exercise 40.3, if the total number $N = K + M$ of spacetime dimensions is odd, then there is no element of the geometric algebra that accomplishes parity or time reversal by operations like (41.141) and (41.142). The implementation of parity and time reversal in odd N dimensions is described in the next part.

10. **Super spacetime algebra in odd dimensions, version 2.** As described in parts 7 and 10 of Exercise 40.3, there are two ways to construct the supergeometric algebra in odd $N = K + M$ dimensions, the first being to project the algebra into one dimension lower, the second to embed the algebra in one dimension higher, and to treat either the final (odd) dimension γ_N or the extra (even) dimension γ_{N+1} as a scalar. The vectors γ_N or γ_{N+1} have the usual property of anticommuting with all orthonormal vectors γ_m other than themselves. If the number K of time dimensions is odd, then the scalar axis γ_N

or γ_{N+1} serves as a time-reversal operator T , while if the number of time dimensions is even, then the scalar axis serves as a parity-reversal operator P . If the number of time dimensions is odd, a suitable parity operator is $P = \gamma_a T$, where γ_a is any spatial vector; while if the number of time dimensions is even, a suitable time-reversal operator is $T = \gamma_k P$ where γ_k is any time vector.

14. **Conjugation.** Part 14 of Exercise 40.3 mostly carries through, but the condition that the Lorentz-invariant conjugation operator C commute with all real orthonormal bivectors, and anticommute with all imaginary orthonormal bivectors translates into the condition that, in place of expression (40.148), C equals, modulo a normalization factor, the product of the spinor metric tensor ε (or the alternative spinor metric tensor ε_{alt}) with the product of all timelike orthonormal basis vectors,

$$C = \varepsilon \mathbf{\Gamma}^\top, \quad \mathbf{\Gamma} \equiv c \prod_m \gamma_m(\text{timelike}), \quad (41.143)$$

where the phase factor c is chosen such that the square of $\mathbf{\Gamma}$ is the unit matrix,

$$\mathbf{\Gamma}^2 = 1. \quad (41.144)$$

The conjugation operator C is unitary, as is $\mathbf{\Gamma}$. The condition (41.144) implies that the eigenvalues of $\mathbf{\Gamma}$ are ± 1 , which ensures that $\bar{\psi} \cdot \psi$ is real, equation (41.149). If there is at least one time dimension, then the trace of $\mathbf{\Gamma}$ is zero, implying that there are equal numbers of $+1$ and -1 eigenvalues. If there is just one time dimension γ_0 , then $\mathbf{\Gamma}$ is (the minus sign is optional, but conventional in the Dirac algebra, equation (41.103)),

$$\mathbf{\Gamma} = -i\gamma_0. \quad (41.145)$$

Notwithstanding equation (41.143), the conjugation operator C is defined to transform not as an element of the geometric algebra, but rather as a spinor tensor that is invariant under Lorentz transformations.

Conjugation flips all space-space and time-time bits of a spinor, while keeping all space-time bits unflipped. The chirality of a spinor is its sign under the chiral operator $\varkappa_N$. For even $K-M$, conjugation flips the chirality of a spinor if $(K-M)/2$ is odd, and leaves the chirality unchanged if $(K-M)/2$ is even. For odd $K-M$, if the path proposed in part 7 is followed, where the odd- N algebra is projected into one lower dimension, which requires identifying $\varkappa_N$ with unity, then chirality is not a rotationally invariant property of spinors. If on the other hand the path proposed in part 10 is followed, where the odd- N algebra is projected into one higher dimension, then chirality is the sign under $\varkappa_{N+1}$.

The double conjugate of a spinor is

$$\bar{\bar{\psi}} = C(C\psi^*)^* = CC^*\psi = \pm\psi, \quad (41.146)$$

where the sign is $+$ or $-$ depending on whether the conjugation operator is symmetric, $C = C^\top$, or antisymmetric (the symmetry condition $C = C^\top$ is equivalent to $CC^* = 1$ in view of the unitarity of C , equation (41.89)). Table 41.1 shows the symmetry of conjugation operator C for the standard and alternative spinor metrics, including the tilde'd versions (40.90) for odd $K-M$. Table 41.1 is essentially identical to the earlier Table 40.1, except that the number N of spatial dimensions is changed to the difference $K-M$ of numbers of space and time dimensions. For Dirac spinors in 3+1 dimensions, the

Table 41.1: Symmetry of the conjugation operator C

$(K - M) \pmod{8}$	C	C_{alt}	$\tilde{C}$	$\tilde{C}_{\text{alt}}$
1	+	+	+	−
2	+	−		
3	−	−	+	−
4	−	−		
5	−	−	−	+
6	−	+		
7	+	+	−	+
8	+	+		

conventional choice is the standard spinor metric (40.88), which ensures that the conjugation operator is symmetric, hence that the double conjugate of a spinor is itself, $\bar{\bar{\psi}} = \psi$.

The scalar product of a conjugate spinor $\bar{\psi}$ with a spinor χ is (compare equation (41.102))

$$\bar{\psi} \cdot \chi = \psi^\dagger C^\top \varepsilon \chi = \psi^\dagger \mathbf{\Gamma} \chi , \quad (41.147)$$

which is a complex (with respect to i) number. In particular, in a basis with respect to which $\mathbf{\Gamma}$ is diagonal, the scalar product of a conjugate basis spinor $\bar{\epsilon}_a$ with a basis spinor ϵ_b is plus or minus the Kronecker delta,

$$\bar{\epsilon}_a \cdot \epsilon_b = \pm \delta_{ab} , \quad (41.148)$$

the sign being that of the eigenvalue of $\mathbf{\Gamma}$. The scalar product of a spinor ψ with its conjugate is

$$\bar{\psi} \cdot \psi = \psi^\dagger \mathbf{\Gamma} \psi , \quad (41.149)$$

which is real given that the eigenvalues of $\mathbf{\Gamma}$ are real. In zero time dimensions the scalar product of a spinor with its conjugate was always positive, equation (40.67), but with one or more time dimensions the scalar product of a spinor with its conjugate can be either positive or negative.

The scalar product $\bar{\psi} \cdot \mathbf{\Gamma} \chi$ is

$$\bar{\psi} \cdot \mathbf{\Gamma} \chi = \psi^\dagger \chi . \quad (41.150)$$

In particular, $\bar{\psi} \cdot \mathbf{\Gamma} \psi$ is real and positive,

$$\bar{\psi} \cdot \mathbf{\Gamma} \psi = \psi^\dagger \psi . \quad (41.151)$$

The conjugate $\bar{\mathbf{a}}$ of an outer product $\mathbf{a} \equiv \psi \bar{\chi} \cdot$ of spinors is related to the outer product $\bar{\psi} \chi \cdot$ by

$$\mathbf{a} \equiv \psi \bar{\chi} \cdot , \quad \bar{\mathbf{a}} \equiv \overline{\psi \bar{\chi} \cdot} = \pm \bar{\psi} \chi \cdot , \quad (41.152)$$

where the $\pm$ sign is the symmetry of the spinor metric, Table 40.1.

The conjugate $\bar{\gamma}_A$ of a basis multivector is defined by equation (40.163). The conjugate of an orthonormal basis vector γ_m is, in place of equation (40.164),

$$\bar{\gamma}_m = \pm(-)^M \gamma_m , \quad (41.153)$$

where the $\pm$ sign is as given in Table 40.2. For the (3+1)-dimensional Dirac algebra, the $\pm$ sign is $-$, and the number of time dimensions is $M = 1$, so $\bar{\gamma}_m = \gamma_m$, in agreement with equation (41.115).

15. **Real subalgebra.** As in part 15 of Exercise 40.3, a real subalgebra of the complex geometric algebra may be obtained by restricting to multivectors satisfying the reality condition that they are their own conjugates,

$$\bar{a} = a . \quad (41.154)$$

Conjugates of orthonormal basis vectors are plus or minus themselves per equation (41.153). If the overall sign $\pm(-)^M$ in equation (41.153) is $+$, as it is for example in the (3+1)-dimensional Dirac algebra, then the real subalgebra consists of real linear combinations of orthonormal basis multivectors. If the sign $\pm(-)^M$ in equation (41.153) is $-$, then the real subalgebra consists of linear combinations of odd-grade orthonormal multivectors with pure imaginary coefficients and even-grade orthonormal multivectors with pure real coefficients.

Part 15 of Exercise 40.3 showed that a real super spacetime subalgebra could be obtained as the algebra of outer products of self-conjugate spinors,

$$\bar{\psi} = \psi , \quad (41.155)$$

which worked provided that the conjugation operator is symmetric. If there are time as well as space dimensions, then the algebra of outer products of self-conjugate spinors is real, satisfying condition (41.154), only if both the spinor metric ε and the conjugation operator C are symmetric, that is, the sign is $+$ in both Tables 40.1 and 41.1. This is not true for example in the (3+1)-dimensional Dirac algebra, where the spinor metric is antisymmetric. Suppose that $\psi = \bar{\psi}$ and $\chi = \bar{\chi}$ are self-conjugate spinors. The conjugate of their multivector outer product $a = \psi\chi \cdot$ satisfies

$$\bar{a} = \overline{\psi\chi \cdot} = C(\psi\chi \cdot)^* C^{-1} = \pm(C\psi^*)\chi^\dagger C^\top \varepsilon = \pm(C\psi^*)(C\chi^*)^\top \cdot = \pm\bar{\psi}\bar{\chi} \cdot = \pm\psi\chi \cdot = \pm a , \quad (41.156)$$

the $\pm$ sign at the third step coming from commuting the conjugation operator through the spinor metric. Given that the conjugation operator must be symmetric for spinors to be self-conjugate, so $C = C^\top = \Gamma\varepsilon^\top$, it follows that

$$C\varepsilon C = \Gamma\varepsilon^\top \varepsilon \Gamma\varepsilon^\top = \Gamma^2 \varepsilon^\top = \varepsilon^\top = \pm\varepsilon , \quad (41.157)$$

the $\pm$ sign being the symmetry of the spinor metric. The sign is positive, yielding a real geometric algebra satisfying equation (41.154), only if the spinor metric is symmetric.

16. **Transformations that leave the spinor scalar product unchanged.** The first half of part 16 of Exercise 40.3 carries through. The list (40.171) of grades of multivectors that generate transformations that preserve the spinor scalar product remains unchanged.

But the condition for the scalar product of spinors and conjugate spinors to be preserved under a

transformation $R = e^{-\theta \gamma_A/2}$ generated by a grade- p multivector γ_A is modified. The commutation rule (40.173) is modified to

$$CR^* = Ce^{-\theta^* \gamma_A^*/2} = e^{-\theta^* \gamma_A/2} C = e^{-(\pm)^p (-)^{Mp} \theta^* \gamma_A/2} C, \quad (41.158)$$

where the $\pm$ sign in $(\pm)^p$, from equation (41.153), is as given in Table 40.2. A scalar product $\bar{\psi} \cdot \chi$ of a conjugate spinor with a spinor transforms under R to, in place of equation (40.174),

$$(C(R\psi)^*) \cdot (R\chi) = \bar{\psi} e^{-(\pm)^p (-)^{Mp} \theta^* \gamma_A^*/2} \varepsilon R\chi = \bar{\psi} \varepsilon e^{-(-)^{[p/2]} (-)^{Mp} \theta^* \gamma_A/2} R\chi, \quad (41.159)$$

where the sign $(-)^{[p/2]} (-)^{Mp}$ in the final expression is the product of $(\pm)^p (-)^{Mp}$ and the sign $(\pm)^p (-)^{[p/2]}$ in the commutation rule (40.98) of a multivector γ_A through the spinor metric ε . The spinor product is preserved provided that $e^{-(-)^{[p/2]} (-)^{Mp} \theta^* \gamma_A/2} = R^{-1}$, which is to say provided that

$$-(-)^{[p/2]} (-)^{Mp} \theta^* = \theta. \quad (41.160)$$

If the number of time dimensions is $M = 1$, or more generally if the number M of time dimensions is odd, then the scalar product of spinors and conjugate spinors is preserved under transformations generated by multivectors of grade p provided that the coefficient θ satisfies

$$\theta \text{ real} \quad \text{grades } (1 \text{ or } 2) \bmod 4 \quad (\text{thus } 1, 2, 5, 6, \dots), \quad (41.161a)$$

$$\theta \text{ imaginary} \quad \text{grades } (0 \text{ or } 3) \bmod 4 \quad (\text{thus } 0, 3, 4, 7, \dots). \quad (41.161b)$$

If the number M of time dimensions is even, then $(-)^{Mp} = 1$, and the earlier condition (40.176) holds.

17. **Rotor group.** The rotor group is generated by the basis of orthonormal bivectors. Bivectors that are the wedge product of a timelike vector and a spacelike vector are multiplied by i , so that rotations in a time-space plane take the exponential form $e^{\theta/2}$, rather than being rotations by a phase, $e^{-i\theta/2}$. The orthonormal basis bivectors remain traceless and unitary, but whereas time-time and space-space bivectors remain skew-Hermitian, the time-space bivectors become Hermitian. The rotor group in K spatial dimensions and M time dimensions is called $\text{Spin}(K, M)$. The construction (40.108) described in Exercise 40.3 embeds $\text{Spin}(K, M)$ as a subgroup of the group $\text{SL}(2^{[N/2]}, \mathbb{C})$, where $N \equiv K+M$, of complex $2^{[N/2]} \times 2^{[N/2]}$ matrices of unit determinant,

$$\text{Spin}(K, M) \subseteq \text{SL}(2^{[N/2]}, \mathbb{C}), \quad N \equiv K + M. \quad (41.162)$$

The group is not unitary, since bivector generators that are wedge products of a timelike vector and a spacelike vector are Hermitian, whereas unitarity requires all generators to be skew-Hermitian (compare Exercise 14.17). Switching time and space dimensions leaves the group unchanged, so $\text{Spin}(K, M)$ is isomorphic to $\text{Spin}(M, K)$.

18. **Grade-preserving subgroup of $\text{Spin}(K, M)$.** As in part 18 of Exercise 40.3, there exists a subgroup of $\text{Spin}(K, M)$ that preserves the grade (number of up bits) of the spinor. The construction (40.180) runs into an obstacle because mixed space-time bivectors cannot be combined in real linear combinations with space-space or time-time bivectors to form bivectors of zero spin (complex linear combinations yes, but not real linear combinations). The best that can be done is to minimize the number of mixed space-time

bivectors, by grouping spatial dimensions into pairs, and time dimensions into pairs, leaving at most one pair of dimensions a mixed combination of a space and a time dimension. The mixed pair is needed only if both space and time dimensions K and M are odd. The construction (40.180) then yields $[K/2]^2$ skew-Hermitian space-space generators, $[M/2]^2$ skew-Hermitian time-time generators, and $2[K/2][M/2]$ Hermitian space-time generators. If there is a mixed space-time pair of dimensions, then there is 1 extra Hermitian space-time generator. Altogether the grade-preserving subgroup of $\text{Spin}(K, M)$ has dimension $[(K+M)/2]^2$ if at most one of K or M is odd, or $([K/2] + [M/2])^2 + 1$ if K and M are both odd. The largest unitary subgroup of $\text{Spin}(K, M)$ is the direct product $U([K/2]) \times U([M/2])$ of unitary groups generated by the $[K/2]^2$ skew-Hermitian space-space generators and the $[M/2]^2$ skew-Hermitian time-time generators.

Quantum field theory of a real scalar field

“Shut up and calculate!” – Mermin (1989).

Quantum mechanics is motivated by the observation of wave-particle duality. A particle behaves like a particle when you look at it, but like a wave when you don’t. Quantum mechanics must explain four observed facts: (1) in units $\hbar = 1$, the energy of a particle equals its frequency, and its momentum equals its wavevector, equations (42.3); (2) particles of the same species are indistinguishable; (3) fermions and boson satisfy an exclusion and anti-exclusion principle; (4) particles can be created and destroyed.

Traditional nonrelativistic quantum mechanics solves the first problem by describing a particle by a wavefunction, and postulating that energy and momentum are operators that act on the wavefunction. It solves the second and third problems by describing multiple particles by (anti)symmetrized products of single-particle wavefunctions. But the description by wavefunctions does not allow for the creation and destruction of particles. Experiment shows that collisions of sufficient energy can create particle-antiparticle pairs that were not there before.

Quantum field theory solves all four problems by asserting the existence of fields, which are operators that act on a state, creating and destroying particles into and out of that state. According to Feynman’s prescription, a field operator is a sum of both creation and destruction parts: the field operator pairs a forward-in-time operator that destroys a particle with a backward-in-time operator that creates the corresponding antiparticle, as in equation (42.41) for the simple harmonic oscillator, equations (42.61) for a real scalar field, (44.58) for a Dirac spinor field, or (45.8) for the electromagnetic field. The prescription enforces so-called *CPT* symmetry as an exact symmetry of nature. The resulting Feynman propagator (42.105) of the real scalar field, (44.83) of the Dirac spinor field, or (45.15) of the electromagnetic field, propagates positive-energy waves into the asymptotic future and negative energy waves into the asymptotic past. Negative energy particles propagating into the past are equivalent to positive energy antiparticles propagating into the future.

States are conveniently written in Dirac’s elegant bra-ket $\langle | \rangle$ notation, in which $|in\rangle$ denotes an in-state, and $\langle out|$ denotes an out-state. States are vectors in a vector space of (possibly) infinite dimension called Hilbert space. Hilbert space is equipped with a complex-valued inner product denoted $\langle out|in\rangle$. When an in-state $|in\rangle$ is ‘measured,’ acted on by some operator observable such as position or momentum, the probability that the outcome will be in an eigenstate $\langle out|$ of the observable equals the magnitude squared of the complex

amplitude $\langle \text{out} | \text{in} \rangle$ between the in- and out-states,

$$P_{\text{in} \rightarrow \text{out}} = |\langle \text{out} | \text{in} \rangle|^2 . \quad (42.1)$$

The mathematics of quantum field theory is encapsulated by a powerful and conceptually appealing approach invented by Feynman: Feynman diagrams, and the Feynman rules associated with them. Feynman diagrams contain two basic building blocks: vertices where quantized interactions occur; and lines joining vertices. The vertices represent unresolved points of spacetime where interactions occur. The lines represent the evolution of fields propagating between interaction points. In basic quantum field theory, the background spacetime is taken to be flat Minkowski space, and the propagation of fields between points is governed by free wave equations, which are the free Klein-Gordon equation for scalars, the free Dirac equation for spinors, and the free Maxwell equations for the electromagnetic field. The Feynman diagram approach is essentially perturbative, with diagrams with more vertices describing higher order perturbative contributions.

A jarring feature of quantum field theory is that interactions are asserted to take place at points. This is built into the assumed Lagrangians of the theory. For example, the interaction between a Dirac spinor ψ of charge e and the electromagnetic field $\mathbf{A} \equiv A^m \boldsymbol{\gamma}_m$ is described by the term

$$ie\bar{\psi} \cdot \mathbf{A}\psi , \quad (42.2)$$

or more explicitly $ie\bar{\psi}(x) \cdot \mathbf{A}(x)\psi(x)$, in the Dirac Lagrangian (44.30), an interaction illustrated by the Feynman diagram 47.1. This forces quantum field theory to deal with the notion of a field operator $\hat{\psi}(x)$ that creates and destroys a field at a spacetime point x .

That quantum field theory must be doing something right is evidenced by the fact that it has proven spectacularly successful at reproducing experimental measurements (e.g. Langacker, 2004). However, an apparent flaw of quantum field theory is that, if it is supposed to hold to arbitrarily tiny scales, then interactions at tiny scales diverge. For the three forces of the Standard Model, the divergences are logarithmic, and can be cured by a process called renormalization. Renormalization cuts off calculations at some tiny scale, takes bare masses, charges, and field normalizations to be different from observed quantities, and concludes that observables can be arranged to be independent of the small-scale cutoff. The cutoff independence is associated with the fact that SM couplings are dimensionless; for example, the strength of the electromagnetic coupling is $e^2 = 4\pi\alpha \approx 0.1$, equation (44.2). The mathematics of quantum field theory cannot be blamed for its divergences. The divergences indicate that some new physics must enter at tiny scales.

The fourth force, gravity, is not renormalizable, because the coupling strength of gravity, proportional to Newton's gravitational constant G , is not dimensionless, but rather has dimension length^2 (in units $c = \hbar = 1$). Coupling constants with positive length dimension give nonrenormalizable theories; coupling constants with negative length dimension give superrenormalizable theories (Gross and Wilczek, 1973a; Peskin and Schroeder, 1995). The positive length dimension means that higher order terms of a perturbative expansion are increasingly divergent to smaller scales, preventing the theory from being truncated at a tiny scale in a manner that allows observables to be independent of the cutoff scale. As a consequence, each term of the perturbative expansion of gravitational quantum field theory has a different dimension, each of which must be renormalized separately, requiring an infinite number of renormalization parameters. This contrasts with the 3 SM forces, where renormalization reduces to the behaviour of just three parameters, mass, charge,

and field normalization, as a function of cutoff scale. An alternative view (Burgess, 2004; Donoghue, 2019) is that the difficulties of the non-renormalizability of gravity are overstated. Almost all the renormalization parameters of gravity affect its behaviour only at energies approaching the Planck scale (or whatever might be the unification scale of gravity). The effects of quantum gravity at energies much less than the Planck scale are calculable, and unaffected by the high-energy behaviour. From this perspective the gravitational force is no different from the 3 SM forces: some new physics must enter at tiny scales. It is possible that the new physics is the same for all 4 forces.

A widespread modern view is that quantum field theory is, like classical thermodynamics, an effective theory (Weinberg, 1995). This view is encouraged by mathematical difficulties in putting quantum field theory on a rigorous mathematical foundation (Kronz and Lupher, 2019, §5); the difficulties stem from the need to treat spacetime and energy-momentum in the continuous limit, and the need to treat interactions at unresolved points.

This chapter follows a standard physics approach to introducing quantum field theory, which is to start by quantizing a classical simple harmonic oscillator, §42.2, and use that as a template to quantize a real scalar field, also known as a Klein-Gordon field, §42.3. If the approach comes across as hand-wavy, yes it is. The chapter concludes, §42.4, with an exposition of the Klein-Gordon propagator, a key ingredient of the Feynman rules.

42.1 Quanta

Experiment shows that quantum mechanical waves are quantized into quanta that carry various discrete properties, such as spin and charge. Among those discrete properties are energy and momentum. The energy E of a particle is proportional to the angular frequency ω of the corresponding wave, and its spatial momentum p_a is proportional to the wavevector k_a , with constant of proportionality the reduced Planck constant $\hbar$,

$$E = \hbar\omega, \quad p_a = \hbar k_a, \quad (42.3)$$

or in four-dimensional notation,

$$p_n = \hbar k_n. \quad (42.4)$$

Equations (42.3), or (42.4), are exact, not uncertain. The left hand sides are particle properties; the right hand sides are wave properties. Equations (42.3) could hardly be simpler; yet this particle-wave schizophrenia is profoundly perplexing to our human intuition.

In quantum mechanics, the relations (42.3) or (42.4) are reproduced by asserting that a particle of energy E and momentum p_a is described by a complex-valued wavefunction ϕ of the form

$$\phi \propto e^{-i\omega t + i k_a x^a} \propto e^{i k_n x^n}. \quad (42.5)$$

The particle described by the wavefunction (42.5) has a definite energy and momentum, but it does not have a definite location in time t or space x^a . In quantum mechanics, classical observables such as position and

momentum cease to be mere numbers, and are interpreted instead as operators that act on the wavefunction. In particular, the energy and momentum operators $\hat{E}$ and $\hat{p}_a$ are

$$\hat{E} = i\hbar \frac{\partial}{\partial t} , \quad \hat{p}_a = -i\hbar \frac{\partial}{\partial x^a} , \quad (42.6)$$

which, when acting on the wavefunction (42.5), reproduce correctly its energy and momentum. In four-dimensional notation, the energy and momentum operators (42.6) are

$$\hat{p}_n = -i\hbar \frac{\partial}{\partial x^n} . \quad (42.7)$$

The relation (42.7) between energy-momenta p_n and spacetime derivatives ∂_n is the standard quantum mechanics convention. Beware that this convention is opposite to the standard cosmological convention, §26.8.2, adopted in Chapters 26–37.

42.2 Quantization of an elementary simple harmonic oscillator

Plane waves undergo pure periodic evolution. The classical system that undergoes pure periodic motion is the simple harmonic oscillator. The properties of the quantum simple harmonic oscillator reproduce at least some of the observed properties of quantum systems of bosons, such as photons in a laser. The energy levels of the quantum simple harmonic oscillator are discrete, and equally spaced, equation (42.42). The lowest energy state, $n = 0$, the ground state, can be interpreted as the state containing 0 photons, while the n th energy state can be interpreted as the state containing n photons. The n photons are automatically indistinguishable.

This section 42.2 reviews how an elementary simple harmonic oscillator is quantized. The procedure provides a template for how to define a field operator $\hat{\phi}$.

42.2.1 Classical simple harmonic oscillator

The Hamiltonian $H(q, p)$ of a classical simple harmonic oscillator with coordinate q and canonically conjugate momentum p is a sum of a kinetic energy $p^2/(2m)$ and a quadratic potential energy $m\omega^2 q^2/2$

$$H = \frac{p^2}{2m} + \frac{m\omega^2 q^2}{2} . \quad (42.8)$$

The Hamiltonian of a real scalar field, equation (42.53), is similar, differing in that the real scalar field involves a continuum of simple harmonic oscillators, and its coordinate ϕ and conjugate momenta π do not have the mass m factor of equation (42.8).

Consider then the Hamiltonian

$$H = \frac{p^2}{2} + \frac{\omega^2 q^2}{2} . \quad (42.9)$$

The classical Hamilton's equations are, with respect to some time parameter t ,

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = p, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\omega^2 q. \quad (42.10)$$

Combining Hamilton's equations (42.9) gives the classical equations of motion for the coordinate q ,

$$\frac{d^2 q}{dt^2} = -\omega^2 q, \quad (42.11)$$

which describes simple harmonic motion at frequency ω . The solutions of the equation of motion (42.11) are waves,

$$q \propto e^{\pm i\omega t}, \quad p \propto \pm i\omega e^{\pm i\omega t}. \quad (42.12)$$

42.2.2 Reinterpret coordinates and momenta as operators

The solution (42.12) is for a classical simple harmonic oscillator, not a quantum simple harmonic oscillator. The simple harmonic oscillator is quantized by reinterpreting the coordinate and momentum as operators $\hat{q}$ and $\hat{p}$ that act on some wavefunction ϕ . The expansions of the wavefunction ϕ in coordinate and momentum space at some time t are related by a Fourier transform,

$$\phi(p) = \int \phi(q) e^{-ipq/\hbar} dq, \quad \phi(q) = \int \phi(p) e^{ipq/\hbar} \frac{dp}{2\pi\hbar}. \quad (42.13)$$

Do not be confused by the fact that $\phi(q)$ and its Fourier transform $\phi(p)$ are written with the same symbol ϕ . The wavefunction ϕ is a vector in a Hilbert space, with components $\phi_q = \phi(q)$ expressed with respect to a coordinate basis, and components $\phi_p = \phi(p)$ expressed with respect to a momentum basis. The momentum operator $\hat{p}$ acting on the wavefunction $\phi(p)$ in the momentum representation multiplies it by p , while acting on the wavefunction $\phi(q)$ in the coordinate representation differentiates it by $-i\hbar\partial/\partial q$,

$$\hat{p}\phi(p) = p\phi(p), \quad \hat{p}\phi(q) = -i\hbar \frac{\partial}{\partial q} \phi(q), \quad (42.14)$$

as follows from

$$\hat{p}\phi(q) = \int p\phi(p) e^{ipq/\hbar} \frac{dp}{2\pi\hbar} = -i\hbar \frac{\partial}{\partial q} \int \phi(p) e^{ipq/\hbar} \frac{dp}{2\pi\hbar}. \quad (42.15)$$

Similarly the coordinate operator $\hat{q}$ acting on the wavefunction $\phi(q)$ in the coordinate representation multiplies it by q , while acting on the wavefunction $\phi(p)$ in the momentum representation differentiates it by $i\hbar\partial/\partial p$,

$$\hat{q}\phi(q) = q\phi(q), \quad \hat{q}\phi(p) = i\hbar \frac{\partial}{\partial p} \phi(p). \quad (42.16)$$

A wavefunction ϕ cannot be an eigenfunction of both coordinate and momentum simultaneously. An eigenfunction of the coordinate operator $\hat{q}$ is a delta-function at some definite coordinate q_0 , in which case the wavefunction is distributed over all momenta,

$$\phi(q) = \delta(q - q_0), \quad \phi(p) = e^{-ipq_0/\hbar}. \quad (42.17)$$

Likewise an eigenfunction of the momentum operator $\hat{p}$ is a delta-function at some definite momentum p_0 , in which case the wavefunction is distributed over all coordinates,

$$\phi(q) = e^{ip_0 q/\hbar} , \quad \phi(p) = 2\pi\hbar\delta(p - p_0) . \quad (42.18)$$

42.2.3 Lowering and raising operators

The essence of the operator relations (42.14) or (42.16) is expressed compactly by the commutators

$$[\hat{q}, \hat{p}] = i\hbar , \quad [\hat{q}, \hat{q}] = [\hat{p}, \hat{p}] = 0 . \quad (42.19)$$

The commutators (42.19) are all that is needed to construct the complete solution of the quantized simple harmonic oscillator. To carry out the solution, introduce dimensionless lowering and raising operators a and $a^\dagger$ defined in terms of the coordinate and momentum operators $\hat{q}$ and $\hat{p}$ by

$$a \equiv \sqrt{\frac{\omega}{2\hbar}} \hat{q} + i\sqrt{\frac{1}{2\hbar\omega}} \hat{p} , \quad a^\dagger \equiv \sqrt{\frac{\omega}{2\hbar}} \hat{q} - i\sqrt{\frac{1}{2\hbar\omega}} \hat{p} , \quad (42.20)$$

where ω is the frequency of the harmonic oscillator, equation (42.9). Inverting the relations (42.20) yields the coordinate and momentum operators in terms of the lowering and raising operators,

$$\hat{q} = \sqrt{\frac{\hbar}{2\omega}} (a + a^\dagger) , \quad \hat{p} = -i\sqrt{\frac{\hbar\omega}{2}} (a - a^\dagger) . \quad (42.21)$$

The lowering and raising operators satisfy the commutation relations

$$[a, a^\dagger] = 1 , \quad [a, a] = [a^\dagger, a^\dagger] = 0 . \quad (42.22)$$

42.2.4 States of n quanta

Quantum field theory postulates the existence of a special state, the vacuum, denoted $|0\rangle$, that contains no particles. As long as the vacuum is stable, the vacuum remains the vacuum, so the vacuum-to-vacuum amplitude is one,

$$\langle 0|0\rangle = 1 . \quad (42.23)$$

The vacuum is interpreted as containing zero quanta. The in-vacuum state $|0\rangle$ gives zero when acted on by the lowering operator,

$$a|0\rangle = 0 , \quad (42.24)$$

while the out-vacuum state $\langle 0|$ gives zero when acted on by the raising operator,

$$\langle 0|a^\dagger = 0 . \quad (42.25)$$

The raising operator $a^\dagger$ creates one quantum out of the vacuum,

$$a^\dagger|0\rangle = |1\rangle , \quad (42.26)$$

while the lowering operator a removes the single quantum

$$a|1\rangle = |0\rangle . \quad (42.27)$$

The one-quantum-to-one-quantum amplitude is one,

$$\langle 1|1\rangle = \langle 0|aa^\dagger|0\rangle = \langle 0|[a, a^\dagger] + a^\dagger a|0\rangle = 1 . \quad (42.28)$$

States $|n\rangle$ containing n particles are obtained by acting on the vacuum with the raising operator n times. Raising and lowering operators $a^\dagger$ and a acting on n -particle states increase and decrease the number of particles by one,

$$\langle n|a = \langle n+1|\sqrt{n+1} , \quad a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle , \quad (42.29a)$$

$$\langle n|a^\dagger = \langle n-1|\sqrt{n} , \quad a|n\rangle = \sqrt{n}|n-1\rangle . \quad (42.29b)$$

The ensemble of n -particle states form an orthonormal set,

$$\langle m|n\rangle = \delta_{mn} . \quad (42.30)$$

The operator $a^\dagger a$ counts the number n of particles in an n -particle state, while the operator $aa^\dagger$ counts one more, consistent with the commutation rule (42.22),

$$a^\dagger a|n\rangle = n|n\rangle , \quad aa^\dagger|n\rangle = (n+1)|n\rangle . \quad (42.31)$$

The operator $a^\dagger a$ is said to be normal-ordered, meaning that all creation operators are to the left of all destruction operators.

Out- and in-states $\langle n|$ and $|n\rangle$ of n particles, correctly normalized to $\langle n|n\rangle = 1$, are

$$\langle n| = \frac{1}{\sqrt{n!}} \langle 0|a^n , \quad \frac{1}{\sqrt{n!}} (a^\dagger)^n |0\rangle = |n\rangle . \quad (42.32)$$

Similarly

$$\langle 0| = \frac{1}{\sqrt{n!}} \langle n|(a^\dagger)^n , \quad \frac{1}{\sqrt{n!}} a^n |n\rangle = |0\rangle . \quad (42.33)$$

Concept question 42.1 Is vacuum nothing, or is it something? Quantum field theory postulates the existence of out- and in-vacua $\langle 0|$ and $|0\rangle$. Are these vacuum states nothing, or are they something?

42.2.5 Hamiltonian operator

Since $\hat{p}$ and $\hat{q}$ are now operators, the Hamiltonian (42.9) can likewise be reinterpreted as an operator $\hat{H}$,

$$\hat{H} = \frac{\hat{p}^2}{2} + \frac{\omega^2 \hat{q}^2}{2} . \quad (42.34)$$

In terms of products of the lowering and raising operators, equations (42.21), the Hamiltonian operator (42.34) is

$$\hat{H} = \frac{\hbar\omega}{2} (aa^\dagger + a^\dagger a) = \hbar\omega \left(aa^\dagger - \frac{1}{2} \right) = \hbar\omega \left(a^\dagger a + \frac{1}{2} \right) . \quad (42.35)$$

The commutators of the Hamiltonian operator with the lowering and raising operators are

$$[\hat{H}, a] = -\hbar\omega a, \quad [\hat{H}, a^\dagger] = \hbar\omega a^\dagger, \quad (42.36)$$

and with the coordinate and momentum operators are

$$[\hat{H}, \hat{q}] = -i\hbar\hat{p}, \quad [\hat{H}, \hat{p}] = i\omega^2\hbar\hat{q}. \quad (42.37)$$

The operator equivalents of the classical Hamilton's equations (42.10) are

$$\frac{d\hat{q}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{q}] = \hat{p}, \quad \frac{d\hat{p}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{p}] = -\omega^2\hat{q}. \quad (42.38)$$

If $\hat{f}$ is an operator that is some function of (for example, a polynomial in) the coordinate and momentum operators $\hat{q}$ and $\hat{p}$, then $\hat{f}$ evolves according to

$$\frac{d\hat{f}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{f}]. \quad (42.39)$$

Inserting the commutators (42.36) into the evolution equation (42.39) implies that the lowering and raising operators evolve as

$$a(t) = a(0)e^{-i\omega t}, \quad a^\dagger(t) = a^\dagger(0)e^{i\omega t}. \quad (42.40)$$

With the time dependence explicit, equations (42.21) are

$$\boxed{\hat{q}(t) = \sqrt{\frac{\hbar}{2\omega}} (a(0)e^{-i\omega t} + a^\dagger(0)e^{i\omega t}), \quad \hat{p}(t) = -i\sqrt{\frac{\hbar\omega}{2}} (a(0)e^{-i\omega t} - a^\dagger(0)e^{i\omega t})}. \quad (42.41)}$$

Equations (42.41) constitute the prototype for expressions for a bosonic field and its conjugate momentum, such as those (42.61) for a real scalar field.

The Hamiltonian operator acting on the n -quantum state $|n\rangle$ gives

$$\hat{H}|n\rangle = \hbar\omega \left(n + \frac{1}{2}\right) |n\rangle. \quad (42.42)$$

Thus the Hamiltonian counts the energy of the n -quantum state, plus an extra $\frac{1}{2}\hbar\omega$. The $\frac{1}{2}\hbar\omega$ contribution, which is present even in the vacuum, is interpreted as the zero-point fluctuation of the vacuum. Equation (42.42) exhibits the characteristic feature of a quantized simple harmonic oscillator, that the energy levels are equally spaced, and that the energy level of the ground state is $\frac{1}{2}$ a unit above the minimum of the potential.

42.2.6 Eigenmodes of the quantum simple harmonic oscillator

As differential operators in the coordinate q , the lowering and raising operators are, equation (42.20),

$$a = \sqrt{\frac{\omega}{2\hbar}}q + \sqrt{\frac{\hbar}{2\omega}}\frac{\partial}{\partial q}, \quad a^\dagger = \sqrt{\frac{\omega}{2\hbar}}q - \sqrt{\frac{\hbar}{2\omega}}\frac{\partial}{\partial q}. \quad (42.43)$$

Let ϕ_0 denote the ground state wavefunction of the quantum simple harmonic oscillator. The lowering operator a acting on the ground state yields zero, $a\phi_0 = 0$, equation (42.24), which solves to

$$\phi_0 = \left(\frac{\omega}{\pi\hbar}\right)^{1/4} \exp\left(-\frac{\omega q^2}{2\hbar}\right), \quad (42.44)$$

normalized to $\int |\phi_0|^2 dq = 1$. Excited states ϕ_n are obtained by acting on the ground state ϕ_0 n times with the raising operator,

$$\phi_n = \frac{1}{\sqrt{n!}} (a^\dagger)^n \phi_0 = \frac{1}{\sqrt{n!2^n}} \left(\frac{\omega}{\pi\hbar}\right)^{1/4} \exp\left(-\frac{\omega q^2}{2\hbar}\right) H_n\left(\sqrt{\frac{\omega}{\hbar}}q\right), \quad (42.45)$$

where $H_n(x)$ are Hermite polynomials. The wavefunctions ϕ_n form an orthonormal set,

$$\int_{-\infty}^{\infty} \phi_m \phi_n dq = \delta_{mn}. \quad (42.46)$$

42.3 Real scalar field

With the exception of the electroweak Higgs field, no fundamental scalar (spin 0) field has been observed in nature. However, scalar fields are predicted by theoretical physicists and cosmologists to play a central role in the Universe. The field that drives inflation in cosmology is postulated to be a scalar field, Chapter 38. The cold dark matter that dominates matter in the Universe could perhaps be a scalar particle. String theory predicts that the coupling of gravity is mediated by a scalar particle called the dilaton. Supersymmetry predicts that fundamental spin- $\frac{1}{2}$ fields should have fundamental spin-0 superpartners.

This section deals with a real scalar field, as an exemplar of how a field is quantized. A real scalar field is one that carries no charge of any kind. The electroweak Higgs field is, before electroweak symmetry breaking, part of a 4-component multiplet carrying electroweak charges. After electroweak symmetry breaking, 3 of the 4 components are absorbed into generating masses for the electroweak bosons $W^\pm$ and Z , leaving the electroweak Higgs boson, which carries broken weak charge. Thus the notion of a real scalar field may be an idealized fiction; but it is a guidepost to quantization.

This section uses natural units $c = \hbar = 1$.

42.3.1 Real scalar Lagrangian

For a scalar field φ , the torsion-free covariant derivative of the field is the same as its directed derivative, $\mathring{D}_\mu \varphi = \partial_\mu \varphi$. The general relativistic Lagrangian $L(\varphi, \partial_\mu \varphi)$ of a real scalar field φ with potential energy $V(\varphi)$ is

$$L = -\frac{1}{2} g^{\mu\nu} (\partial_\mu \varphi) (\partial_\nu \varphi) - V(\varphi), \quad (42.47)$$

where $g^{\mu\nu}$ is the inverse metric. If the scalar field is a free field with mass m , then the potential energy $V(\varphi)$ is that of a simple harmonic oscillator, quadratic in the field φ , with energy levels spaced by the mass m ,

$$V(\varphi) = \frac{1}{2}m^2\varphi^2 . \quad (42.48)$$

In dimensionful units, the mass m in equation (42.48) would be replaced by $mc/\hbar$.

The momentum π^μ conjugate to the field φ is defined to be the derivative of the Lagrangian with respect to the velocity $\partial_\mu\varphi$ of the field,

$$\pi^\mu \equiv \frac{\delta L}{\delta(\partial_\mu\varphi)} = -\partial^\mu\varphi . \quad (42.49)$$

The Euler-Lagrange equations of motion are, equation (16.7),

$$\dot{D}_\mu\pi^\mu = \frac{\delta L}{\delta\varphi} = -m^2\varphi , \quad (42.50)$$

which, when combined with (42.49), yield the Klein-Gordon wave equation

$$(\Box - m^2)\varphi = 0 , \quad (42.51)$$

where $\Box \equiv \dot{D}_\mu\dot{D}^\mu$ is the torsion-free d'Alembertian.

Concept question 42.2 **What are the units of the scalar field?** **Answer.** The action of a field is defined to be $S = \int L d^4x$, equation (16.1). In quantum field theory, the action must be dimensionless (in natural units $c = \hbar = 1$), which is true provided that the scalar field φ has units 1/length. The conjugate momentum π , equation (42.49) has units 1/length².

42.3.2 Real scalar super-Hamiltonian

The super-Hamiltonian $H(\varphi, \pi^\mu)$ of a real scalar field φ is defined by, equation (16.8),

$$H \equiv \pi^\mu\partial_\mu\varphi - L , \quad (42.52)$$

which, with the relation (42.49) between the momentum and velocity of the field, yields a super-Hamiltonian that resembles the Hamiltonian of a simple harmonic oscillator,

$$H = \frac{1}{2}(-\pi^\mu\pi_\mu + m^2\varphi^2) . \quad (42.53)$$

Hamilton's equations are, equations (16.11),

$$\dot{D}_\mu\pi^\mu = -\frac{\delta H}{\delta\varphi} = -m^2\varphi , \quad \partial_\mu\varphi = \frac{\delta H}{\delta\pi^\mu} = -\pi_\mu . \quad (42.54)$$

Combining equations (42.54) recovers the Klein-Gordon wave equation (42.51).

42.3.3 Real scalar conventional Hamiltonian

The conventional Hamiltonian, as opposed to super-Hamiltonian, of a real scalar field φ is defined to be $H \equiv \pi^t \partial_t \varphi - L$, equation (16.13). The momentum π^t conjugate to the coordinate φ is $\pi^t = \delta L / \delta(\partial_t \varphi) = -\partial^t \varphi$, and the conventional Hamiltonian is

$$H = \frac{1}{2} (-\pi^t \pi_t + (\partial^a \varphi)(\partial_a \varphi) + m^2 \varphi^2) , \quad (42.55)$$

which differs from the super-Hamiltonian (42.53) in the sign of the spatial derivative term $\frac{1}{2}(\partial^a \varphi)(\partial_a \varphi)$. The variation of the spatial derivatives $\partial_a \varphi$ in the conventional Hamiltonian (42.55) is given by $\delta(\partial_a \varphi) = \partial_a(\delta \varphi)$, so that the variation of the spatial derivative term in the Hamiltonian is

$$\delta\left(\frac{1}{2}(\partial^a \varphi)(\partial_a \varphi)\right) = (\partial^a \varphi)\partial_a(\delta \varphi) = \mathring{D}_a((\partial^a \varphi)(\delta \varphi)) - (\mathring{D}_a \partial^a \varphi)(\delta \varphi) , \quad (42.56)$$

the first term on the right hand side of which integrates to a surface term. Hamilton's equations are

$$\mathring{D}_t \pi^t = -\frac{\delta H}{\delta \varphi} = \mathring{D}_a \partial^a \varphi - m^2 \varphi , \quad \partial_t \varphi = \frac{\delta H}{\delta \pi^t} = -\pi_t , \quad (42.57)$$

which again recover the Klein-Gordon wave equation (42.51).

42.3.4 Space+time (3+1) split of the scalar equations

Quantization of fields requires identifying the physical degrees of freedom in a matching number of fields and their conjugate momenta, but, as discussed in §16.5.8 and §16.6.6 for the electromagnetic field and in §16.15 for the gravitational field, the super-Hamiltonian approach does not yield equal numbers of coordinates and momenta, because the covariant approach admits extra gauge degrees of freedom that have no physical significance. The conventional Hamiltonian approach does yield equal numbers of fields and conjugate momenta, but at the price of sacrificing general covariance.

Dirac's (1964) solution is to adhere to the covariant super-Hamiltonian approach, but to interpret only the Hamilton's equations involving time derivatives as genuine equations of motion, while other equations are interpreted either as constraints, in which the equation must be satisfied on the initial conditions, but whose preservation is then guaranteed by conservation laws, or else as identities, which define redundant fields in terms of others (in electromagnetism, for example, the magnetic field is defined as the spatial curl of the electric potential, equation (16.47b)).

For a real scalar field, the time-dependent equations of motion are

$$1 \text{ equation of motion: } \mathring{D}_t \pi^t + \mathring{D}_\alpha \pi^\alpha = -m^2 \varphi , \quad (42.58a)$$

$$1 \text{ equation of motion: } \partial_t \varphi = -\pi_t . \quad (42.58b)$$

The remaining scalar Hamilton's equations (42.54), those not involving time derivatives of the coordinates and momenta, form 3 identities

$$3 \text{ identities: } \partial_\alpha \varphi = -\pi_\alpha . \quad (42.59)$$

In Minkowski (flat) spacetime, spatial translation invariance implies that the covariant spatial momentum

π_α is a constant of motion (and time translation invariance implies that π_t is a constant of motion). But in curved (nonflat) spacetime, π_α (and π_t) are not constants of motion. Particle quantum field theory posits that an interaction, thus the creation and destruction of particles, takes place at a point x of spacetime, and the field operator $\hat{\varphi}(x)$ at that point must be expanded as a superposition of plane waves with respect to a locally inertial frame at that point. Thus even though in curved spacetime the solutions to the Klein-Gordon wave equation (42.51) are not global plane waves, nevertheless the solution in the infinitesimal vicinity of a point x is a superposition of plane waves. The Klein-Gordon wave equation (42.51) implies that each plane wave component $e^{ip_\mu x^\mu}$ of the superposition satisfies $p_\mu p^\mu = -m^2$.

42.3.5 Quantization of the real scalar field

The scalar field is quantized by reinterpreting the scalar field and its conjugate momenta as operators $\hat{\varphi}$ and $\hat{\pi}$, and imposing equal-time quantum commutation relations analogous to those (42.19) of the simple harmonic oscillator,

$$[\hat{\varphi}(t, x'^\alpha), \hat{\pi}(t, x^\alpha)] = i \delta^3(t, x'^\alpha - x^\alpha) , \quad (42.60a)$$

$$[\hat{\varphi}(t, x'^\alpha), \hat{\varphi}(t, x^\alpha)] = [\hat{\pi}(t, x'^\alpha), \hat{\pi}(t, x^\alpha)] = 0 , \quad (42.60b)$$

in which the delta-function in equation (42.60a) is to be interpreted as defined with respect to a locally inertial frame at the point where x^α and x'^α coincide. Per Concept question 42.2, $\hat{\varphi}$ and $\hat{\pi}$ have units 1/length and 1/length² respectively, so the commutator $[\hat{\varphi}, \hat{\pi}]$ has units of 1/length³, in agreement with the 3-dimensional delta-function on its right hand side. The commutation relation (42.60a) for $[\hat{\varphi}, \hat{\pi}]$ is not Lorentz-invariant. The scalar field φ is by assumption a Lorentz scalar, therefore Lorentz-invariant, but its conjugate momentum is the time component π_t of a vector π_μ , and is not Lorentz-invariant. The Feynman rules work with the field operator $\hat{\varphi}$, equation (42.61a) or equivalently equation (42.65), which *is* Lorentz-invariant.

Equations (42.41) for the simple harmonic oscillator suggest that, at least in Minkowski (flat) spacetime, the field operator $\hat{\varphi}$ and its conjugate momentum operator $\hat{\pi}$ should be expanded in destruction and creation operators a_p and $a_p^\dagger$ as

$$\hat{\varphi}(x^\mu) \equiv \int \sqrt{\frac{1}{2E_p}} \left(a_p e^{ip_\mu x^\mu} + a_p^\dagger e^{-ip_\mu x^\mu} \right) \frac{d^3 p}{(2\pi)^3} \Big|_{p^t=E_p} , \quad (42.61a)$$

$$\hat{\pi}(x^\mu) \equiv -i \int \sqrt{\frac{E_p}{2}} \left(a_p e^{ip_\mu x^\mu} - a_p^\dagger e^{-ip_\mu x^\mu} \right) \frac{d^3 p}{(2\pi)^3} \Big|_{p^t=E_p} = \partial_t \hat{\varphi}(x^\mu) . \quad (42.61b)$$

The conjugate momentum operator (42.61b) is the time component of a contravariant 4-vector,

$$\hat{\pi}^\nu(x^\mu) = -i \int p^\nu \sqrt{\frac{1}{2E_p}} \left(a_p e^{ip_\mu x^\mu} - a_p^\dagger e^{-ip_\mu x^\mu} \right) \frac{d^3 p}{(2\pi)^3} \Big|_{p^t=E_p} = -\partial^\nu \hat{\varphi}(x^\mu) . \quad (42.62)$$

In curved (nonflat) spacetime, the plane-wave eigenfunctions $e^{\pm ip_\mu x^\mu}$ should be replaced by eigenfunctions of the Klein-Gordon equation (42.51) having the same $e^{\pm ip_\mu x^\mu}$ behaviour in an infinitesimal interval of spacetime

around the point $\{t, x\}$. The destruction and creation operators a_p and $a_p^\dagger$ themselves satisfy commutation relations analogous to those (42.22) of the simple harmonic oscillator,

$$[a_{p'}, a_p^\dagger] = (2\pi)^3 \delta^3(p'_\alpha - p_\alpha) , \quad (42.63a)$$

$$[a_{p'}, a_p] = [a_{p'}^\dagger, a_p^\dagger] = 0 . \quad (42.63b)$$

Taking the equal-time commutator of the expressions (42.61) for $\hat{\varphi}$ and $\hat{\pi}$, and eliminating the resulting commutators of a and $a^\dagger$ using equations (42.63), recovers the desired commutation rules (42.60), notably the rule (42.60a),

$$[\hat{\varphi}(t, x'^\alpha), \hat{\pi}(t, x^\alpha)] = i \int e^{ip_\alpha(x'^\alpha - x^\alpha)} \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} = i \delta^3(t, x'^\alpha - x^\alpha) . \quad (42.64)$$

Like the commutator $[\hat{\varphi}, \hat{\pi}]$ of equation (42.60a), the commutator $[a_{p'}, a_p^\dagger]$ of equation (42.63a) is not Lorentz-invariant, and the destruction and creation operators a_p and $a_p^\dagger$ are likewise not Lorentz-invariant. But $\sqrt{2E_p}a_p$ and $\sqrt{2E_p}a_p^\dagger$ are Lorentz-invariant, and the field operator $\hat{\varphi}$ can be written in the Lorentz-invariant form

$$\hat{\varphi}(x^\mu) = \int \left((\sqrt{2E_p}a_p) e^{ip_\mu x^\mu} + (\sqrt{2E_p}a_p^\dagger) e^{-ip_\mu x^\mu} \right) \frac{d^3p}{2E_p(2\pi)^3} \Big|_{p^t=E_p} . \quad (42.65)$$

The 3-dimensional integral can be replaced by a 4-dimensional integral over the positive energy $p^t > 0$ solution of the wave equation,

$$\int \frac{d^3p}{2E_p(2\pi)^3} \Big|_{p^t=E_p} = \int 2\pi \delta(p^2 + m^2) \frac{d^4p}{(2\pi)^4} \Big|_{p^t>0} . \quad (42.66)$$

At least in Minkowski space, quantum field theory postulates the existence of a special state, the vacuum, which contains no particles. The out-vacuum is denoted $\langle 0|$, and the in-vacuum $|0\rangle$, and the vacuum-to-vacuum amplitude is one, $\langle 0|0\rangle = 1$. Analogously to the raising and lowering operators (42.25) and (42.24) of the simple harmonic oscillator, the creation and destruction operators of the scalar field are defined to satisfy

$$\langle 0|a_p^\dagger = 0 , \quad a_p|0\rangle = 0 . \quad (42.67)$$

The state consisting of zero particles in the momentum state p , without regard to the possible presence of particles in other momentum states, is denoted $|0_p\rangle$. Out- and in-states containing n particles in the momentum state p are denoted $\langle n_p|$ and $|n_p\rangle$. Similarly to the simple harmonic oscillator (42.29), creation and destruction operators $a_p^\dagger$ and a_p acting on n -particle states increase and decrease the number of particles by one,

$$\langle n_p|a_p = \langle (n+1)_p|\sqrt{n+1} , \quad a_p^\dagger|n_p\rangle = \sqrt{n+1}|(n+1)_p\rangle , \quad (42.68a)$$

$$\langle n_p|a_p^\dagger = \langle (n-1)_p|\sqrt{n} , \quad a_p|n_p\rangle = \sqrt{n}|(n-1)_p\rangle . \quad (42.68b)$$

Similarly to the simple harmonic oscillator (42.31), the operator $a_p^\dagger a_p$ counts the number n of particles in an

n -particle momentum state $|n_p\rangle$, while the operator $a_p a_p^\dagger$ counts one more,

$$a_p^\dagger a_p |n_p\rangle = n |n_p\rangle, \quad a_p a_p^\dagger |n_p\rangle = (n+1) |n_p\rangle. \quad (42.69)$$

Similarly to the case of the simple harmonic oscillator, equations (42.32) and (42.33), out- and in-states $\langle n_p|$ and $|n_p\rangle$, correctly normalized to $\langle n_p|n_p\rangle = 1$, are

$$\langle n_p| = \frac{1}{\sqrt{n!}} \langle 0_p| (a_p)^\dagger{}^n, \quad \frac{1}{\sqrt{n!}} (a_p^\dagger)^n |0_p\rangle = |n_p\rangle, \quad (42.70a)$$

$$\langle 0_p| = \frac{1}{\sqrt{n!}} \langle n_p| (a_p^\dagger)^n, \quad \frac{1}{\sqrt{n!}} (a_p)^n |n_p\rangle = |0_p\rangle. \quad (42.70b)$$

42.3.6 Klein-Gordon Hamiltonian operator

As in the case of the simple harmonic oscillator, §42.2.5, since the scalar field $\hat{\varphi}$ and its conjugate momentum $\hat{\pi}$ are now operators, the Hamiltonian can likewise be reinterpreted as an operator $\hat{H}$. The super-Hamiltonian (42.53) becomes the operator

$$\hat{H} = \frac{1}{2} (-\hat{\pi}^\mu \hat{\pi}_\mu + m^2 \hat{\varphi}^2), \quad (42.71)$$

while the conventional Hamiltonian (42.55) is

$$\hat{H} = \frac{1}{2} (-\hat{\pi}^t \hat{\pi}_t + (\partial_a \hat{\varphi})(\partial_a \hat{\varphi}) + m^2 \hat{\varphi}^2). \quad (42.72)$$

To find an operator equation of motion analogous to that (42.39) for the simple harmonic oscillator, define the Hamiltonian operator $\hat{\mathcal{H}}$ integrated over all space at $t = 0$,

$$\hat{\mathcal{H}} \equiv \int \hat{H} d^3x \Big|_{t=0}. \quad (42.73)$$

The super-Hamiltonian operator $\hat{\mathcal{H}}$ from equation (42.71) is

$$\hat{\mathcal{H}} = \int m^2 (a_p a_p^\dagger + a_p^\dagger a_p) \frac{d^3p}{2E_p (2\pi)^3} \Big|_{p^t=E_p}, \quad (42.74)$$

while the conventional Hamiltonian operator $\hat{\mathcal{H}}$ from equation (42.72) is

$$\hat{\mathcal{H}} = \int E_p^2 (a_p a_p^\dagger + a_p^\dagger a_p) \frac{d^3p}{2E_p (2\pi)^3} \Big|_{p^t=E_p}. \quad (42.75)$$

The commutators of the conventional Hamiltonian operator (42.75) with the destruction and creation operators a_p and $a_p^\dagger$ are

$$[\hat{\mathcal{H}}, a_p] = -E_p a_p, \quad [\hat{\mathcal{H}}, a_p^\dagger] = E_p a_p^\dagger, \quad (42.76)$$

and with the coordinate and momentum operators $\hat{\varphi}$ and $\hat{\pi}$ are

$$[\hat{\mathcal{H}}, \hat{\varphi}] = \int \sqrt{\frac{1}{2E_p}} \left([\hat{\mathcal{H}}, a_p] e^{ip_\mu x^\mu} + [\hat{\mathcal{H}}, a_p^\dagger] e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} = -i\partial_t \hat{\varphi} , \quad (42.77a)$$

$$[\hat{\mathcal{H}}, \hat{\pi}] = -i \int \sqrt{\frac{E_p}{2}} \left([\hat{\mathcal{H}}, a_p] e^{ip_\mu x^\mu} - [\hat{\mathcal{H}}, a_p^\dagger] e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} = -i\partial_t \hat{\pi} . \quad (42.77b)$$

The operator equivalents of Hamilton's equations (42.57) are

$$\partial_t \hat{\varphi} = i[\hat{\mathcal{H}}, \hat{\varphi}] , \quad \partial_t \hat{\pi} = i[\hat{\mathcal{H}}, \hat{\pi}] . \quad (42.78)$$

If $\hat{f}$ is an operator that is some function of (for example, a polynomial in) the coordinate and momentum operators $\hat{\varphi}$ and $\hat{\pi}$, then $\hat{f}$ evolves according to

$$\partial_t \hat{f} = i[\hat{H}, \hat{f}] . \quad (42.79)$$

The conventional Hamiltonian operator $\hat{\mathcal{H}}$ acting on an eigenstate $|n_p\rangle$, equation (42.70a), containing n particles in momentum state p yields its energy $E_{n,p}$,

$$\hat{\mathcal{H}}|n_p\rangle = E_{n,p}|n_p\rangle . \quad (42.80)$$

The following manipulation shows that the state $|(n+1)_p\rangle$ has energy E_p above that of the state $|n_p\rangle$:

$$\begin{aligned} E_{n+1,p}|(n+1)_p\rangle &= \hat{\mathcal{H}}|(n+1)_p\rangle = \frac{1}{\sqrt{n+1}} \hat{\mathcal{H}} a_p^\dagger |n_p\rangle = \frac{1}{\sqrt{n+1}} (a_p^\dagger \hat{\mathcal{H}} + [\hat{\mathcal{H}}, a_p^\dagger]) |n_p\rangle = \frac{1}{\sqrt{n+1}} (E_{n,p} + E_p) a_p^\dagger |n_p\rangle \\ &= (E_{n,p} + E_p) |(n+1)_p\rangle . \end{aligned} \quad (42.81)$$

Thus, like a simple harmonic oscillator, the energy levels are equally spaced, with the n th level lying nE_p above the energy $E_{0,p}$ of the ground state $|0_p\rangle$. The energy of the ground state $|0_p\rangle$ itself satisfies

$$\hat{\mathcal{H}}|0_p\rangle = \int \frac{E_{p'}}{2} (2\pi)^3 \delta^3(p'-p) \frac{d^3p'}{(2\pi)^3} \Big|_{p'^t=E_{p'}} |0_p\rangle = \frac{E_p}{2} |0_p\rangle , \quad (42.82)$$

so is not zero, but rather $E_p/2$.

The vacuum state $|0\rangle$ satisfies $a_p a_p^\dagger |0\rangle = |0\rangle$ for any momentum p , equation (42.31), so the conventional Hamiltonian operator $\hat{\mathcal{H}}$ acting on the vacuum state is an integral over all momenta,

$$\hat{\mathcal{H}}|0\rangle = \int \frac{E_p}{2} \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} |0\rangle , \quad (42.83)$$

which is nominally infinite. This is one of the many infinities that dog quantum field theory. The standard physics approach is to discard the infinity, and hope that one day it might all make sense.

42.4 The Klein-Gordon propagator in Minkowski spacetime

A central component of the Feynman rules is the amplitude for a field to propagate from one spacetime point x to another x' . This amplitude is called the propagator, and may be evaluated as the Green's function of the relevant wave equation. The Klein-Gordon propagator $D(x' - x)$ is the solution of

$$(-\square' + m^2)D(x' - x) = -i\delta^4(x' - x) \quad (42.84)$$

with appropriate boundary conditions, which are discussed in §42.4.1.

In Minkowski (flat) space, the Klein-Gordon propagator $D(x' - x)$ can be evaluated explicitly. The Green's function equation (42.84) solves to

$$D(x' - x) = -i \int \frac{1}{p^2 + m^2} e^{ip \cdot (x' - x)} \frac{d^4 p}{(2\pi)^4} . \quad (42.85)$$

The defining equation (42.84) for the Klein-Gordon propagator $D(x' - x)$ is Lorentz invariant in Minkowski space, and it has a solution $D(s)$ that is Lorentz invariant, where

$$s \equiv \sqrt{(x' - x)^2} \quad (42.86)$$

is the Lorentz-invariant spacetime separation from x to x' . The spacetime separation s lies in three disconnected regions that cannot be rotated into each other by any continuous Lorentz transformation, namely spacelike (x' outside the lightcone of x), positive timelike (x' inside the future lightcone of x), and negative timelike (x' inside the past lightcone of x). In the spacelike region, the spacetime separation is real and positive, $s > 0$, and equal to the proper distance between x and x' . In the timelike regions, the spacetime separation is pure imaginary, $s = i\tau$ where τ is the proper time from x to x' . The proper time τ is positive in the future lightcone, negative in the past lightcone.

Rewritten in terms of the Lorentz-invariant spacetime separation s , the defining equation (42.84) for the Klein-Gordon propagator is

$$\left(-\frac{\partial^2}{\partial s^2} - \frac{3}{s} \frac{\partial}{\partial s} + m^2 \right) D(s) = -i\delta^4(x' - x) = \frac{\delta(s)}{2\pi^2 s^3} . \quad (42.87)$$

The expression for the delta-function in s on the right hand side of equation (42.87) comes from Euclideanizing time, $x'^0 - x^0 \rightarrow -is^0$, and from the fact that the Euclideanized 4-volume element integrated over angles is $2\pi^2 s^3 ds$. The homogeneous solution of equation (42.87) that converges at spacelike infinity is the Bessel function $s^{-1}K_1(ms)$. Integrating the Euclideanized Green's function equation (42.87) over the origin at $s = 0$ yields the change in the derivative of the propagator at the origin,

$$\left[-s^3 \frac{dD}{ds} \right] = \int \frac{\delta(s)}{2\pi^2 s^3} s^3 ds = \frac{1}{2\pi^2} . \quad (42.88)$$

The spacetime delta-function $\delta^4(x' - x)$ in equation (42.87) is supposed to be nonzero only at the origin $x' - x = 0$. However, the spacetime separation s vanishes not only at the origin but everywhere along the lightcone $s^2 = 0$. To ensure that there is no delta-function across the lightcone (except at the origin),

$[s^3 dD/ds]_{s^2=-0}^{s^2=+0}$ must vanish across the lightcone, which requires that $D \propto s^{-2}$ as $s^2 \rightarrow 0$, with the same constant of proportionality for either sign of s^2 . The constant of proportionality comes from the integral (42.88) over the origin in 4-dimensional Euclidean space, yielding the Lorentz-invariant Klein-Gordon propagator $D(s)$,

$$D(s) = \frac{m}{4\pi^2 s} K_1(ms) \rightarrow \frac{1}{4\pi^2 s^2} \begin{cases} 1 & s \rightarrow 0, \\ \sqrt{\frac{\pi ms}{2}} e^{-ms} & s \rightarrow \infty. \end{cases} \quad (42.89)$$

The derivative at the origin satisfies $-s^3 dD/ds \rightarrow 1/(2\pi^2)$ as $s \rightarrow 0$, in agreement with equation (42.88). At spacelike separations $s^2 > 0$, the propagator (42.89) is pure real. A more pedantic derivation of the expression (42.89) for the Klein-Gordon propagator is given in §42.4.1.

At timelike separations $s^2 < 0$, the spacetime separation $s = i\tau$ is imaginary, with the proper time τ being positive in the future lightcone, negative in the past lightcone. The expression (42.89) for the Klein-Gordon propagator holds also for timelike separations, but it is not obvious whether the correct continuation of the expression across the lightcone at $s^2 = 0$ should be $D(i\tau)$ or $D(-i\tau)$. The correct choice, Feynman's choice, derived in the next section 42.4.1 by contour integration of the integral expression (42.85), proves to be $D(i|\tau|)$, that is $D(i\tau)$ if $\tau > 0$, and $D(-i\tau)$ if $\tau < 0$,

$$D(s) = D(i|\tau|) = \frac{m}{4\pi^2(i|\tau|)} K_1(im|\tau|) \rightarrow -\frac{1}{4\pi^2 \tau^2} \begin{cases} 1 & \tau \rightarrow 0, \\ \sqrt{\frac{i\pi m|\tau|}{2}} e^{-im|\tau|} & \tau \rightarrow \pm\infty. \end{cases} \quad (42.90)$$

Note that the Bessel function of imaginary argument can also be expressed as a Hankel function of real argument, $K_1(im|\tau|) = -\frac{\pi}{2} H_1^{(1)}(-m|\tau|) = -\frac{\pi}{2} H_1^{(2)}(m|\tau|)$. The Feynman-Klein-Gordon propagator (42.89) or (42.90) has the property that $D(x' - x)$ is identical to $D(x - x')$, for both spacelike and timelike separations,

$$D(x - x') = D(x' - x). \quad (42.91)$$

The equality of the forwards and backwards propagators expresses the fact that the scalar field φ is its own antiparticle.

In the massless limit $m = 0$, the Klein-Gordon propagator reduces to

$$D(s) = \frac{1}{4\pi^2 s^2} \quad (m = 0), \quad (42.92)$$

at both spacelike and timelike separations.

The Klein-Gordon propagator (42.89) or (42.90) is nonvanishing for all separations $x' - x$, including not only future timelike separations, but also spacelike separations, and past timelike separations. This means that an event that occurs at x has an effect on what happens not only on points x' in the future of x , but also on points x' spacelike separated from x , and on points x' in the past of x . But the propagator is Lorentz invariant, and is furthermore invariant under re-ordering of the times t and t' at which the timelike- or spacelike-separated events x and x' occur. If the time-ordering of events makes no difference, then x “causing” what happens at x' can equally well be interpreted as x' “causing” x . This is the strange conundrum of quantum mechanics.

42.4.1 Klein-Gordon propagator by contour integration

The correctness of the expressions (42.89) and (42.90) for the Klein-Gordon propagator $D(x'-x)$ in Minkowski spacetime at respectively spacelike and timelike separations can be established by carrying out the energy p^0 part of the integral (42.85) over d^4p by contour integration. The integral over energy p^0 , which is over all energies including negative as well as positive energy, is

$$-i \int_{-\infty}^{\infty} \frac{1}{p^2 + m^2} e^{-ip^0(t'-t)} \frac{dp^0}{2\pi} = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \left(\frac{1}{p^0 + E_p} - \frac{1}{p^0 - E_p} \right) \frac{e^{-ip^0(t'-t)}}{2E_p} dp^0, \quad (42.93)$$

where $E_p = \sqrt{p_a p^a + m^2}$ is the positive energy of a particle of spatial momentum p^a and mass m . The integrand has two poles, at $p^0 = \pm E_p$, through both of which the integral nominally passes. To obtain a definite result, the contour of integration must be perturbed above or below each of the poles, and completed in the complex plane in a way that ensures convergence. The value of the integral depends on which way the contour is perturbed.

If $t' - t > 0$, then the contour must be completed in the lower complex plane, and the integral picks up a residue $\pm e^{\mp i E_p(t'-t)}$ for each pole $p^0 = \pm E_p$ that the contour circulates. If on the other hand $t' - t < 0$, then the contour must be completed in the upper complex plane, and the integral picks up a residue $\mp e^{\mp i E_p(t'-t)}$ for each pole $p^0 = \pm E_p$ that the contour circulates. In all,

$$\begin{aligned} -i \int_{-\infty}^{\infty} \frac{1}{p^2 + m^2} e^{-ip^0(t'-t)} \frac{dp^0}{2\pi} &= \frac{1}{2E_p} \begin{cases} e^{-iE_p(t'-t)} & (t' - t > 0, + \text{ pole}), \\ -e^{iE_p(t'-t)} & (t' - t > 0, - \text{ pole}), \\ -e^{-iE_p(t'-t)} & (t' - t < 0, + \text{ pole}), \\ e^{iE_p(t'-t)} & (t' - t < 0, - \text{ pole}) \end{cases} \\ &= \frac{1}{2E_p} \begin{cases} e^{-iE_p|t'-t|} & (t' - t > 0 \text{ and } + \text{ pole, or } t' - t < 0 \text{ and } - \text{ pole}), \\ -e^{iE_p|t'-t|} & (t' - t > 0 \text{ and } - \text{ pole, or } t' - t < 0 \text{ and } + \text{ pole}). \end{cases} \end{aligned} \quad (42.94)$$

If the separation $x' - x$ is spacelike, then since the Klein-Gordon propagator $D(x' - x)$ is Lorentz invariant, the separation can be Lorentz transformed without loss of generality so that $t' - t = 0$. As $t' - t \rightarrow 0$, the right hand sides of equations (42.94) go over to $\pm 1/(2E_p)$. The contribution $D_+(x' - x)$ to the Klein-Gordon propagator (42.85) from the upper (+) option in the final equality of equations (42.94) is then, at spacelike separations,

$$\begin{aligned} D_+(x' - x) &= \int \frac{1}{2E_p} e^{ip_a(x'^a - x^a)} \frac{d^3p}{(2\pi)^3} = \int_{p=0}^{\infty} \int_{\theta=0}^{\pi} \frac{1}{2\sqrt{p^2 + m^2}} e^{ips \cos \theta} \frac{2\pi p^2 dp \sin \theta d\theta}{(2\pi)^3} \\ &= \frac{1}{4\pi^2 s} \int_0^{\infty} \frac{1}{\sqrt{p^2 + m^2}} \sin(ps) p dp = -\frac{1}{4\pi^2 s} \frac{\partial}{\partial s} \int_0^{\infty} \frac{1}{\sqrt{p^2 + m^2}} \cos(ps) dp \\ &= -\frac{1}{4\pi^2 s} \frac{\partial}{\partial s} K_0(ms) = \frac{m}{4\pi^2 s} K_1(ms). \end{aligned} \quad (42.95)$$

The integral that yields the Bessel function is integral 3.754.2 of Gradshteyn and Ryzhik (1980). The contribution $D_-(x' - x)$ to the Klein-Gordon propagator from the lower option in the final line of equations (42.94)

is minus the upper option,

$$D_-(x' - x) = -D_+(x' - x) . \quad (42.96)$$

If the separation $x' - x$ is timelike, then again because the Klein-Gordon propagator $D(x' - x)$ is Lorentz invariant, the separation can be Lorentz transformed without loss of generality so that $x'^a - x^a = 0$. The contribution $D_+(x' - x)$ to the Klein-Gordon propagator (42.85) from the upper option in the final equality of equations (42.94) is then, at timelike separations, with $\tau \equiv t' - t$,

$$\begin{aligned} D_+(x' - x) &= \int \frac{1}{2E_p} e^{-iE_p|\tau|} \frac{d^3p}{(2\pi)^3} = \int_0^\infty \frac{1}{2E_p} e^{-iE_p|\tau|} \frac{4\pi p^2 dp}{(2\pi)^3} \\ &= \int_m^\infty \frac{1}{2E_p} e^{-iE_p|\tau|} \frac{4\pi p E_p dE_p}{(2\pi)^3} = \frac{1}{4\pi^2} \int_m^\infty \sqrt{E_p^2 - m^2} e^{-iE_p|\tau|} dE_p \\ &= \frac{m}{4\pi^2 i|\tau|} K_1(im|\tau|) . \end{aligned} \quad (42.97)$$

The integral that yields the Bessel function is integral 3.387.3 of Gradshteyn and Ryzhik (1980). The contribution $D_-(x' - x)$ to the Klein-Gordon propagator from the lower option in the final equality of equations (42.94) is minus the complex conjugate of the upper option,

$$D_-(x' - x) = -D_+(x' - x)^* . \quad (42.98)$$

Equation (42.98) holds at spacelike as well as timelike separations $x' - x$, since the propagator at spacelike separations is real and satisfies equation (42.96).

Different choices of contour yield different propagators. If the contour is perturbed above both poles, then the integral will vanish for $t' - t < 0$ and be finite for $t' - t > 0$, yielding the retarded Klein-Gordon propagator $D_{\text{ret}}(x' - x)$. For $t' - t > 0$, the retarded propagator is a sum of the contributions from both poles, $D_{\text{ret}} = D_+ + D_-$. Since D_- is minus the complex conjugate of D_+ , equation (42.98), and D_+ is real at spacelike separations, the retarded propagator vanishes at spacelike separations for positive as well as negative values of $t' - t$. The retarded Klein-Gordon propagator is

$$D_{\text{ret}}(x' - x) = \begin{cases} \frac{im}{4\pi\tau} J_1(m\tau) & \text{timelike, } \tau \geq 0 , \\ 0 & \text{timelike, } \tau \leq 0 , \\ 0 & \text{spacelike ,} \end{cases} \quad (42.99)$$

where J_1 is the Bessel function, related to the modified Bessel function K_1 by $-\pi J_1(z) = K_1(iz) + K_1(-iz)$.

Similarly, if the contour is perturbed below both poles, then the integral will vanish for $t' - t > 0$ and be finite for $t' - t < 0$, yielding the advanced Klein-Gordon propagator $D_{\text{adv}}(x' - x)$. The advanced Klein-Gordon propagator is

$$D_{\text{adv}}(x' - x) = \begin{cases} 0 & \text{timelike, } \tau \geq 0 , \\ \frac{im}{4\pi|\tau|} J_1(m|\tau|) & \text{timelike, } \tau \leq 0 , \\ 0 & \text{spacelike .} \end{cases} \quad (42.100)$$

The retarded and advanced propagators (42.99) and (42.100) are suitable as classical solutions to the Green's function equation (42.87), but they are unsatisfactory in quantum field theory because, as found in §§47.3, 44.5, the propagators that describe the evolution of particles between interaction vertices of a Feynman diagram necessarily extend outside the lightcone, and can go backwards as well as forwards in time.

The choice of contour adopted in quantum field theory is Feynman's prescription, which chooses the contour that goes below the negative pole $p^0 = -E_p$ and above the positive pole $p^0 = E_p$. A convenient way to implement the Feynman prescription is to perturb the poles of p^0 to $\pm(E_p - i\epsilon)$ where ϵ is a small positive quantity, yielding the Feynman-Klein-Gordon propagator

$$D(x' - x) = -i \int \frac{1}{p^2 + m^2 - i\epsilon} e^{ip \cdot (x' - x)} \frac{d^4 p}{(2\pi)^4} . \quad (42.101)$$

If $t' - t > 0$, then the contour must be completed in the lower complex plane, circulating (clockwise) the positive pole $p^0 = E_p$, yielding $D_+(x' - x)$, the upper option in the final equality of equations (42.94). If on the other hand $t' - t < 0$, then the contour must be completed in the upper complex plane, circulating (anticlockwise) the negative pole $p^0 = -E_p$, again yielding $D_+(x' - x)$, the upper option in the final equality of equations (42.94). The Feynman-Klein-Gordon propagator (42.101) thus integrates to

$$D(x' - x) = D_+(x' - x) = \int e^{\pm ip \cdot (x' - x)} \Big|_{p^0 = E_p} \frac{d^3 p}{2E_p (2\pi)^3} , \quad (42.102)$$

where the $\pm$ sign in the exponent is positive for $t' - t > 0$, negative for $t' - t < 0$. The integral on the right hand side of equation (42.102) is Lorentz invariant as it should be. Thanks to the $\pm$ sign in the exponential, the propagator is unchanged by flipping the sign of $x' - x$, which implies the equality of forwards and backwards propagators, equation (42.91).

The Feynman propagator (42.102) propagates positive-energy waves $\sim e^{-im\tau}$ into the asymptotic future ($\tau > 0$), and negative energy waves $\sim e^{im\tau}$ into the asymptotic past ($\tau < 0$), equation (42.90). If the contour had been chosen instead to go above the negative pole and below the positive pole, then the result would have been $D_-(x' - x)$, minus the complex conjugate of the Feynman propagator (42.102), which would propagate negative-energy waves into the asymptotic future and positive-energy waves into the asymptotic past. Of course, the choice that $e^{-im\tau}$ represents positive energy while $e^{im\tau}$ represents negative energy is just a matter of convention; if the sign convention were opposite, then the flipped contour would be correct.

42.4.2 The Klein-Gordon propagator as a two-point function

The propagators of quantum field theory can be expressed as expectation values of products of field operators at two points. Inserting the expression (42.61a) for the Klein-Gordon field operator $\hat{\varphi}$ in terms of destruction and creation operators, and bearing in mind that only $\langle 0 | a_p a_p^\dagger | 0 \rangle$ is nonzero, yields the two-point function

$$\langle 0 | \hat{\varphi}(x') \hat{\varphi}(x) | 0 \rangle = \int e^{ip \cdot (x' - x)} \Big|_{p^0 = E_p} \frac{d^3 p}{2E_p (2\pi)^3} . \quad (42.103)$$

With the various signs from equations (42.94), expressions for the retarded and advanced propagators in terms of two-point functions are

$$D_{\text{ret}}(x' - x) \equiv \Theta(t' - t) (\langle 0 | \hat{\varphi}(x') \hat{\varphi}(x) | 0 \rangle - \langle 0 | \hat{\varphi}(x) \hat{\varphi}(x') | 0 \rangle) = \Theta(t' - t) \langle 0 | [\hat{\varphi}(x'), \hat{\varphi}(x)] | 0 \rangle, \quad (42.104a)$$

$$D_{\text{adv}}(x' - x) \equiv \Theta(t - t') (\langle 0 | \hat{\varphi}(x') \hat{\varphi}(x) | 0 \rangle - \langle 0 | \hat{\varphi}(x) \hat{\varphi}(x') | 0 \rangle) = \Theta(t - t') \langle 0 | [\hat{\varphi}(x'), \hat{\varphi}(x)] | 0 \rangle, \quad (42.104b)$$

while the expression for the Feynman propagator relevant to quantum field theory is

$$D(x' - x) \equiv \Theta(t' - t) \langle 0 | \hat{\varphi}(x') \hat{\varphi}(x) | 0 \rangle + \Theta(t - t') \langle 0 | \hat{\varphi}(x) \hat{\varphi}(x') | 0 \rangle = \langle 0 | T \hat{\varphi}(x') \hat{\varphi}(x) | 0 \rangle, \quad (42.105)$$

where $\Theta(t)$ is the Heaviside function, defined to be 0 for $t < 0$ and 1 for $t \geq 0$, and the symbol T is the **time-ordering** symbol, which instructs to order the operators in the usual right-to-left physics convention, with the earliest on the right and the latest on the left.

Quantization from the perspective of an accelerating observer

A core feature of quantum field theory is that waves are quantized into quanta of definite frequency, §42.1. However, if an observer sees an emitter to be accelerating (redshifting or blueshifting), then the emitter's pure modes of frequency do not appear to be pure to the observer. In quantum field theory, the resolution of this conundrum is that the observer and emitter perceive different sets of quanta.

A takeaway from this chapter is that an observer quantizes waves according to their own perception of frequencies and wavenumbers, regardless of the observer's state of motion. A second takeaway is that an observer accelerating uniformly in a straight line sees pair production, §43.1. A third takeaway is that acceleration does not by itself imply pair production; for example, an observer uniformly accelerating in a circle does *not* see pair production, Exercise 43.4. A fourth takeaway is that pair creation appears to be associated with the development of an observer-dependent horizon that separates spacetime into a part that observer can see and a part that they cannot see, at least for some period of time.

43.1 Unruh radiation

Consider an observer accelerating uniformly in a straight line in Minkowski space, a Rindler observer, §7.30. As first shown persuasively by Unruh (1976), an observer undergoing proper acceleration at rate κ in empty Minkowski space will perceive themselves to be engulfed in a bath of thermal radiation at temperature

$$T = \frac{\kappa}{2\pi} \quad (43.1)$$

in Planck units $c = \hbar = G = 1$. For accelerations κ much less than a Planck acceleration, the Unruh temperature is tiny, and the predicted thermal radiation is practically imperceptible.

The Rindler line-element (4.57) (with $\alpha \rightarrow \kappa\tau$, $\xi \rightarrow \kappa\xi - \ln \kappa$) can be written

$$ds^2 = e^{2\kappa\xi} (-d\tau^2 + d\xi^2) + dy^2 + dz^2, \quad (43.2)$$

where τ is the proper time of an observer accelerating at rate κ along the worldline $\xi = 1/\kappa$ in the right quadrant of Rindler space, Figure 4.2. The Rindler line-element (43.2) is stationary, the metric coefficients being independent of the time τ . Therefore the time coordinate τ is a Killing coordinate, §7.32.1, and wave

equations of free fields have solutions $\sim e^{-i\omega\tau}$ that the accelerating observer perceives to have pure positive frequency ω . These are the quanta of the accelerating observer.

Minkowski space contains four separate Rindler quadrants, right, top, left, and bottom, Figure 7.25, analogous to the four quadrants of the Kruskal diagram 7.11 of the analytically extended Schwarzschild geometry. The relation between Minkowski coordinates $\{t, x\}$ and Rindler coordinates $\{\tau, \xi\}$ in the various quadrants is

$$\kappa t = e^{\kappa\xi} \sinh(\kappa\tau) , \quad \kappa x = e^{\kappa\xi} \cosh(\kappa\tau) \quad \text{right} , \quad (43.3a)$$

$$\kappa t = e^{\kappa\xi} \cosh(\kappa\tau) , \quad \kappa x = e^{\kappa\xi} \sinh(\kappa\tau) \quad \text{top} , \quad (43.3b)$$

$$\kappa t = -e^{\kappa\xi} \sinh(\kappa\tau) , \quad \kappa x = -e^{\kappa\xi} \cosh(\kappa\tau) \quad \text{left} , \quad (43.3c)$$

$$\kappa t = -e^{\kappa\xi} \cosh(\kappa\tau) , \quad \kappa x = -e^{\kappa\xi} \sinh(\kappa\tau) \quad \text{bottom} . \quad (43.3d)$$

Analogously to Kruskal coordinates, the Rindler time coordinate τ is timelike in the right and left quadrants, but spacelike in the top and bottom quadrants (and the Rindler space coordinate ξ is spacelike where τ is timelike, and vice versa). Define Minkowski null coordinates $\{v, u\}$ and Rindler null coordinates $\{V, U\}$ by

$$v \equiv t + x , \quad u \equiv t - x , \quad (43.4a)$$

$$V \equiv \tau + \xi , \quad U \equiv \tau - \xi . \quad (43.4b)$$

The relation between Minkowski and Rindler null coordinates is

$$\kappa v = \begin{smallmatrix} + \\ - \end{smallmatrix} e^{\kappa V} , \quad \kappa u = \begin{smallmatrix} + \\ - \end{smallmatrix} e^{-\kappa U} , \quad (43.5a)$$

$$\kappa V = \frac{1}{\kappa} \ln(\begin{smallmatrix} + \\ - \end{smallmatrix} \kappa v) = \frac{1}{\kappa} \ln|\kappa v| , \quad \kappa U = -\ln(\begin{smallmatrix} + \\ - \end{smallmatrix} \kappa u) = -\ln|\kappa u| , \quad (43.5b)$$

where the asterism of signs pertains to each of the quadrants.

In Rindler coordinates $\{\tau, \xi, y, z\}$, the wave equation for a scalar field φ of rest mass m is

$$(\square - m^2)\varphi = \left[\pm e^{-2\kappa\xi} \left(-\frac{\partial^2}{\partial\tau^2} + \frac{\partial^2}{\partial\xi^2} \right) + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - m^2 \right] \varphi = 0 , \quad (43.6)$$

where the $\pm$ sign is $+$ in the right and left quadrants, $-$ in the top and bottom quadrants. The wave equation (43.6) has two linearly independent solutions, outgoing ($+$) and ingoing ($-$),

$$\varphi_{\omega k}^{\pm} = A_{\mp\omega, k} e^{i(-\omega\tau + \mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp})} I_{\mp i\omega/\kappa}(ke^{\kappa\xi}/\kappa) , \quad (43.7)$$

where $\mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp} = k_y y + k_z z$, $k \equiv \sqrt{k_{\perp}^2 + m^2}$ (with positive sign, without loss of generality), and $A_{\mp\omega, k}$ is a normalization constant independent of the Killing coordinates. The Bessel functions $I_{\mp i\omega/\kappa}(ke^{\kappa\xi}/\kappa)$ in equation (43.7) are valid in the right and left quadrants of Rindler space; they should be replaced by $J_{\mp i\omega/\kappa}(ke^{\kappa\xi}/\kappa)$ in the upper and lower quadrants. Rindler horizons are located where $\xi \rightarrow -\infty$. On the horizons, the Bessel functions $I_{\mp i\omega/\kappa}(ke^{\kappa\xi}/\kappa)$ simplify, up to a ξ -independent normalization factor, to

$$I_{\mp i\omega/\kappa}(ke^{\kappa\xi}/\kappa) \rightarrow e^{\mp i\omega\xi} \quad \text{as} \quad \xi \rightarrow -\infty , \quad (43.8)$$

which determines the corresponding normalization factor $A_{\mp\omega,k}$. The solutions $\varphi_{\omega k}^{\pm}$ on the horizons are then

$$\begin{cases} \varphi_{\omega k}^+ \rightarrow e^{i(-\omega V + \mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp})} = |\kappa v|^{-i\omega/\kappa} e^{i\mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp}} \\ \varphi_{\omega k}^- \rightarrow e^{i(-\omega U + \mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp})} = |\kappa u|^{i\omega/\kappa} e^{i\mathbf{k}_{\perp} \cdot \mathbf{x}_{\perp}} \end{cases} \quad \text{as } \xi \rightarrow -\infty. \quad (43.9)$$

Minkowski modes, whether positive or negative frequency, must be analytic functions of the outgoing and ingoing null Minkowski coordinates v and u . To Minkowski observers, modes $(\kappa v)^{-i\omega/\kappa}$ and $(\kappa u)^{i\omega/\kappa}$ with $\omega > 0$ are positive frequency, forward-in-time, analytic modes throughout Minkowski space. But to Rindler observers, those same modes are discontinuous across Rindler horizons:

$$(i\kappa v)^{-i\omega/\kappa} = \begin{cases} e^{\pi\omega/(2\kappa)} |\kappa v|^{-i\omega/\kappa} & v > 0 \\ e^{-\pi\omega/(2\kappa)} |\kappa v|^{-i\omega/\kappa} & v < 0 \end{cases}, \quad (43.10a)$$

$$(i\kappa u)^{i\omega/\kappa} = \begin{cases} e^{-\pi\omega/(2\kappa)} |\kappa u|^{i\omega/\kappa} & u > 0 \\ e^{\pi\omega/(2\kappa)} |\kappa u|^{i\omega/\kappa} & u < 0 \end{cases}. \quad (43.10b)$$

A Minkowski observer defines their field operator to be a linear combination of forward-in-time destruction a operators and backward-in-time creation $a^{\dagger}$ operators, with coefficients that are complex conjugate solutions of the Minkowski wave equation,

$$a_{\omega k} (i\kappa v)^{-i\omega/\kappa} + a_{\omega k}^{\dagger} (-i\kappa v)^{i\omega/\kappa}. \quad (43.11)$$

A Rindler observer similarly defines their field operator to be a linear combination of forward-in-time destruction b operators and backward-in-time creation $b^{\dagger}$ operators, with coefficients that are complex conjugates solutions of the Rindler wave equation,

$$b_{\omega k} |\kappa v|^{-i\omega/\kappa} + b_{\omega k}^{\dagger} |\kappa v|^{i\omega/\kappa}. \quad (43.12)$$

The relation (43.10) between Minkowski and Rindler modes implies that the destruction and creation operators are related by

$$\begin{pmatrix} b_{\omega k} \\ b_{\omega k}^{\dagger} \end{pmatrix} = \frac{1}{2 \sinh(\pi\omega/\kappa)} \begin{pmatrix} e^{\pi\omega/(2\kappa)} & e^{-\pi\omega/(2\kappa)} \\ e^{-\pi\omega/(2\kappa)} & e^{\pi\omega/(2\kappa)} \end{pmatrix} \begin{pmatrix} a_{\omega k} \\ a_{\omega k}^{\dagger} \end{pmatrix}, \quad (43.13)$$

the $(2 \sinh(\pi\omega/\kappa))^{-1}$ factor ensuring that the relation between destruction and creation operators is unitary. The coefficients of the matrix (43.13) relating and destruction and creation operators in different frames are called **Bogoliubov coefficients**.

The operator $a_{\omega k}^{\dagger} a_{\omega k}$ counts the number of particles in the state ωk , while the operator $a_{\omega k} a_{\omega k}^{\dagger}$ counts one more, equations (42.69). The Minkowski vacuum contains no particles,

$$\langle 0 | a_{\omega k}^{\dagger} a_{\omega k} | 0 \rangle = 0. \quad (43.14)$$

But the Rindler observer perceives the Minkowski vacuum to be nonempty; rather, each state of frequency (energy) ω has an occupation number of

$$\langle 0 | b_{\omega k}^{\dagger} b_{\omega k} | 0 \rangle = \frac{e^{-\pi\omega/\kappa}}{2 \sinh(\pi\omega/\kappa)} \langle 0 | a_{\omega k} a_{\omega k}^{\dagger} | 0 \rangle = \frac{1}{e^{2\pi\omega/\kappa} - 1}, \quad (43.15)$$

which is that of a thermal distribution at temperature $T = \kappa/(2\pi)$, as anticipated in equation (43.1).

43.2 Plane wave seen by a uniformly accelerating observer

A core message of the previous section 43.1 is that an observer counts quanta with respect to their own proper time, regardless of their state of motion. It might seem then that a shortcut to determining whether an observer watching an emitter sees pair creation would be to Fourier transform the emitted wavetrain into the observer's frame, and see whether the Fourier transform contains negative as well as positive frequencies. As will be seen in this section 43.2, the Fourier approach works for an observer uniformly accelerating in a straight line in Minkowski space, because the observed ratio of negative to positive frequencies, equation (43.23), proves to be independent of the energy and momentum of the plane wave (43.16) in Minkowski space. In general it is necessary to be more careful, and use the apparatus of quantum field theory to determine whether an observer sees pair creation.

For example, Exercise 43.4 considers an observer who accelerates along a Killing trajectory that combines uniform acceleration in the x -direction with uniform rotation in the transverse yz -plane. A conclusion is that an observer who accelerates uniformly in a circle does not see pair creation. Exercise 43.5 Fourier transforms a plane wave seen by the circling observer, finding that the observer sees negative as well as positive frequencies, but the ratio of negative to positive frequencies depends on the energy and momentum of the plane wave being observed, so pair creation cannot be inferred.

Consider then an observer who is uniformly accelerating in a straight line in Minkowski space. It is not actually necessary that the observer themselves be accelerating. It could be that the observer is watching an emitter that appears to the observer to be uniformly accelerating (redshifting or blueshifting) for one reason or another, perhaps because the emitter is at cosmological distances in an accelerating universe, perhaps because the emitter is falling through the horizon of a black hole. The emitter does not actually have to be emitting anything. Indeed the interesting case is where the 'emitter' is actually vacuum, disappearing over the horizon of an accelerating universe, or over the horizon of a black hole.

For definiteness, consider the situation of the previous section 43.1, an observer accelerating at rate κ watching a plane wave of a field of mass m in Minkowski space. In the Minkowski frame, the plane wave takes the form

$$\varphi(t, x, \mathbf{x}_\perp) = e^{i(-Et + px + \mathbf{p}_\perp \cdot \mathbf{x}_\perp)} , \quad (43.16)$$

with energy E and momentum p satisfying

$$E = \sqrt{p^2 + p_\perp^2 + m^2} . \quad (43.17)$$

The coordinates t and x along the path of an observer accelerating uniformly at rate κ in the direction x are related to the proper time τ of the observer by

$$t = \kappa^{-1} \sinh(\kappa\tau) , \quad x = \kappa^{-1} \cosh(\kappa\tau) . \quad (43.18)$$

In terms of the accelerating observer's proper time τ , the wave (43.16) is

$$\varphi(t, x, \mathbf{x}_\perp) = \varphi(\tau) e^{i\mathbf{p}_\perp \cdot \mathbf{x}_\perp} , \quad \varphi(\tau) \equiv \exp \left\{ \frac{i}{2\kappa} [(E+p)\sigma - (E-p)\sigma^{-1}] \right\} , \quad (43.19)$$

where σ is

$$\sigma \equiv e^{-\kappa\tau} . \quad (43.20)$$

The Fourier transform of the wave (43.19) is

$$\begin{aligned} \varphi_\omega &\equiv \int_{-\infty}^{\infty} e^{i\omega\tau} \varphi(\tau) d\tau \\ &= \frac{1}{\kappa} \int_0^\infty \sigma^{-i\omega/\kappa} \exp \left\{ \frac{i}{2\kappa} [(E+p)\sigma - (E-p)\sigma^{-1}] \right\} \frac{d\sigma}{\sigma} \\ &= \frac{2}{\kappa} e^{\pi\omega/(2\kappa)} \left(\frac{E+p}{E-p} \right)^{i\omega/(2\kappa)} K_{i\omega/\kappa} \left(\frac{\sqrt{p_\perp^2 + m^2}}{\kappa} \right) , \end{aligned} \quad (43.21)$$

where $K_{i\omega/\kappa}$ is the modified Bessel function of the second kind. In the limiting case $E \rightarrow p$, which occurs for a massless field, $m = 0$, with no transverse motion, $p_\perp = 0$, the Fourier modes φ_ω are

$$\varphi_\omega = \frac{1}{\kappa} e^{\pi\omega/(2\kappa)} E^{i\omega/\kappa} \Gamma(-i\omega/\kappa) . \quad (43.22)$$

As seen in the previous section 43.1, equation (43.13), the emitter's vacuum appears to the observer to contain particles if modes of positive frequency in the emitter's frame become combinations of positive and negative frequency in the observer's frame. The Bessel function $K_{i\omega/\kappa}(\sqrt{E^2 - p^2}/\kappa)$ in equation (43.21) is real, and unchanged by $\omega \rightarrow -\omega$. The negative frequency component $\varphi_{-\omega}$ corresponding to a positive frequency component φ_ω is

$$\varphi_{-\omega} = e^{-\pi\omega/\kappa} \varphi_\omega^* , \quad (43.23)$$

the $e^{-\pi\omega/\kappa}$ coming from the ratio of $e^{\pi\omega/(2\kappa)}$ factors in equation (43.21). The ratio $|\varphi_{-\omega}/\varphi_\omega| = e^{-\pi\omega/\kappa}$ is the same as that (43.10) found in the previous section 43.1, and it can be concluded that the emitter's and observer's creation and annihilation operators are related in the same fashion (43.13) as before. The observer perceives a thermal distribution at temperature $\kappa/(2\pi)$, equation (43.1).

Formally, the integral (43.21) does not converge at its limits $\tau \rightarrow \mp\infty$, but it can be made to do so by introducing smooth cutoffs at each end. Physically, the observer has only a finite interval of time over which to measure the Fourier modes. The details of the cutoffs are not important. Multiplication of the field $\varphi(\tau)$ by a window in τ is equivalent to convolving the Fourier transform φ_ω with the Fourier transform of the window. If the window is broad, then the convolution kernel is narrow. As long as the Fourier transform is already smooth, smoothing it further makes little difference. One way to impose cutoffs is to give $E \pm p$ small positive and negative imaginary parts, and then set those small parts to zero after evaluating the resulting convergent integral.

Concept question 43.1 Does an observer in the uniform gravitational field of the Earth perceive Unruh radiation? You, standing on the Earth, experience constant acceleration from the gravitational field of the Earth. Should you (in principle) see Unruh radiation associated with that acceleration? Why, or why not? **Answer.** No.

Concept question 43.2 Does an observer inside a uniformly accelerating box perceive Unruh radiation? Put yourself inside a box that is uniformly accelerating in otherwise empty (Minkowski) space. Do you perceive the walls of the box to emit Unruh radiation?

Concept question 43.3 Does an observer uniformly accelerating in a circle perceive Unruh radiation? Consider an observer who accelerates uniformly so as to keep themselves on a circular trajectory in Minkowski space. Do they perceive Unruh radiation? **Answer.** No. See Exercise 43.4.

Exercise 43.4 Pair production seen by an observer moving along a Killing trajectory. A Rindler trajectory is a Killing trajectory generated by a Lorentz boost in Minkowski space. In Minkowski space, the most general timelike Killing trajectory is generated by a composition of a Lorentz boost and a spatial rotation, §14.6. By a suitable Lorentz transformation, the direction of the boost can be taken to be the x -direction, and the rotation can be taken in the yz -plane. Generalize the derivation of Unruh radiation in §43.1 to such a Killing trajectory.

Solution. The Minkowski line-element written in Killing coordinates is, equation (14.91) with $l \equiv e^\xi$ and $r \equiv e^\rho$,

$$ds^2 = -e^{2\xi}(d\alpha^2 + d\xi^2) + e^{2\rho}(d\rho^2 + d\phi^2) , \quad (43.24)$$

generalizing the Rindler line-element (43.2). As in Rindler space, acceleration divides the Killing trajectories into four separate Rindler quadrants of Minkowski space, right, top, left, and bottom, Figure 7.25. In the right quadrant of Rindler space, Minkowski coordinates are related to Killing coordinates by, equation (14.88),

$$\{t, x, y, z\} = \{e^\xi \sinh \alpha, e^\xi \cosh \alpha, e^\rho \cos \phi, e^\rho \sin \phi\} , \quad (43.25)$$

generalizing equation (43.3a). The relation between Minkowski t, x and Killing α, ξ coordinates in the various Rindler quadrants are given by equations (43.3) (with κ set equal to 1). The relation between transverse Minkowski y, z and Killing ρ, ϕ coordinates is the same in all quadrants. Define Minkowski null coordinates $\{v, u, x_+, x_-\}$ and Killing null coordinates $\{V, U, X_+, X_-\}$ by, generalizing equations (43.4),

$$v \equiv t + x , \quad u \equiv t - x , \quad x_+ \equiv y + iz , \quad x_- \equiv y - iz , \quad (43.26a)$$

$$V \equiv \alpha + \xi , \quad U \equiv \alpha - \xi , \quad X_+ \equiv \rho + i\phi , \quad X_- \equiv \rho - i\phi . \quad (43.26b)$$

The relation between Minkowski and Killing null coordinates is, generalizing equations (43.5),

$$v = -\frac{t+x}{\kappa} e^V , \quad u = \frac{t-x}{\kappa} e^{-U} , \quad x_+ = e^{X_+} , \quad x_- = e^{X_-} , \quad (43.27a)$$

$$V = \ln\left(-\frac{t+x}{\kappa}\right) = \ln|v| , \quad U = -\ln\left(\frac{t-x}{\kappa}\right) = -\ln|u| , \quad X_+ = \ln x_+ , \quad X_- = \ln x_- , \quad (43.27b)$$

where the asterism of signs pertains to each of the quadrants of Rindler space. In Killing coordinates the wave equation for a scalar field φ of mass m is, generalizing equation (43.6),

$$(\square - m^2)\varphi = \left[e^{-2\xi} \left(-\frac{\partial^2}{\partial \alpha^2} + \frac{\partial^2}{\partial \xi^2} \right) + e^{-2\rho} \left(\frac{\partial^2}{\partial \rho^2} + \frac{\partial^2}{\partial \phi^2} \right) - m^2 \right] \varphi = 0 . \quad (43.28)$$

The scalar wave equation (43.28) has outgoing (+) and ingoing (−) solutions, generalizing equation (43.7),

$$\varphi^\pm = A_{\mp\nu, \mp\mu, k, k_\perp} e^{-i(\nu\alpha + \mu\phi)} I_{\mp i\nu}(ke^\xi) J_{\mp\mu}(k_\perp e^\rho) , \quad (43.29)$$

where $k = \sqrt{k_\perp^2 + m^2}$, and $A_{\mp\nu, \mp\mu, k, k_\perp}$ is a normalization factor independent of the Killing coordinates. The Bessel functions $I_{\mp i\nu}(ke^\xi)$ in equation (43.29) are valid in the right and left quadrants of Rindler space; they should be replaced by $J_{\mp i\nu}(ke^\xi)$ in the upper and lower quadrants. Each of the frequencies ν and μ can independently be either positive or negative. The transverse factor $e^{-i\mu\phi} J_{\mp\mu}(k_\perp e^\rho)$ is the same analytic function throughout Minkowski space, and has no effect on the Bogoliubov coefficients. Rindler horizons are located where $\xi \rightarrow -\infty$. On the horizons, the function $I_{\mp i\nu}(ke^\xi)$ simplifies, up to a ξ -independent normalization factor, to, generalizing equation (43.8),

$$I_{\mp i\nu}(ke^\xi) \rightarrow e^{\mp i\nu\xi} \quad \text{as} \quad \xi \rightarrow -\infty . \quad (43.30)$$

The solutions $\varphi^\pm$ on the horizons are then, up to a phase, generalizing equations (43.9),

$$\begin{cases} \varphi^+ \rightarrow e^{-i\nu V} = |v|^{-i\nu} \\ \varphi^- \rightarrow e^{-i\nu U} = |u|^{i\nu} \end{cases} \quad \text{as} \quad \xi \rightarrow -\infty . \quad (43.31)$$

The outgoing v modes are discontinuous across the $v = 0$ horizon, while the ingoing u modes are discontinuous across the $u = 0$ horizon, generalizing equations (43.10),

$$\left| \frac{\varphi_{v>0}^+}{\varphi_{v<0}^+} \right| = e^{\pi\nu} = \left| \frac{\varphi_{u<0}^-}{\varphi_{u>0}^-} \right| . \quad (43.32)$$

In the right Rindler quadrant, the proper time $d\tau$ along a Killing trajectory $d\xi = d\rho = 0$ is

$$d\tau = \sqrt{e^{2\xi} d\alpha^2 - e^{2\rho} d\phi^2} = e^\xi \sqrt{1 - v_\perp^2} d\alpha , \quad (43.33)$$

where the constant $v_\perp$ is the circular velocity

$$v_\perp \equiv \frac{e^\rho d\phi}{e^\xi d\alpha} . \quad (43.34)$$

The frequency ω measured by the Killing observer is related to the dimensionless frequency ν by

$$\frac{\omega}{\nu} = \frac{d\alpha}{d\tau} = e^{-\xi} \gamma_\perp , \quad (43.35)$$

where $\gamma_\perp = 1/\sqrt{1 - v_\perp^2}$ is the Lorentz gamma-factor associated with the velocity $v_\perp$. Thus the ratio of positive to negative frequencies perceived by the observer is, from equation (43.32), $e^{\pi\omega/(e^{-\xi}\gamma_\perp)}$. The Killing observer perceives themselves to be in a thermal bath at temperature

$$T = \frac{e^{-\xi} \gamma_\perp}{2\pi} . \quad (43.36)$$

The case of pure circular motion corresponds to $\xi \rightarrow \infty$, for which the temperature is zero. Thus an observer who accelerates uniformly in a circle does not see Unruh radiation.

Exercise 43.5 Plane wave seen by an observer moving in a circle. Fourier transform the wavetrain of a plane wave in Minkowski space with respect to a circulating observer's proper time τ . Does the Fourier transform include negative as well as positive frequencies? Does the circulating observer see pair production?

Solution. A plane wave in Minkowski space takes the form

$$\varphi(t, x, y, z) = e^{i(-Et + p_x x + p_y y + p_z z)} , \quad (43.37)$$

with $E = \sqrt{p^2 + m^2}$. Without loss of generality, take the circulating observer to be moving in the yz -plane, with $x = 0$ and $y = r \cos \phi$, $z = r \sin \phi$. Let the observer's constant circular velocity be v in the Minkowski frame,

$$v = r \frac{d\phi}{dt} . \quad (43.38)$$

The relation between the observer's proper time τ and Minkowski time t is

$$d\tau = dt \sqrt{1 - v^2} . \quad (43.39)$$

The plane wave (43.37) seen by the observer is

$$\varphi(\tau) = \exp \left\{ -iE\gamma\tau + \frac{ir}{2} [(p_y - ip_z)\sigma + (p_y + ip_z)\sigma^{-1}] \right\} , \quad (43.40)$$

where

$$\sigma \equiv e^{i\phi} = e^{i\gamma v \tau / r} . \quad (43.41)$$

The Fourier transform of the wave (43.40) is

$$\begin{aligned} \varphi_\omega &\equiv \int_{-\infty}^{\infty} e^{i\omega\tau} \varphi(\tau) d\tau \\ &= -\frac{ir}{\gamma v} \int_0^\infty \sigma^\nu \exp \left\{ \frac{ir}{2} [(p_y - ip_z)\sigma + (p_y + ip_z)\sigma^{-1}] \right\} \frac{d\sigma}{\sigma} \\ &= -\frac{2ir}{\gamma v} \left(\frac{p_y + ip_z}{p_y - ip_z} \right)^{\nu/2} K_{-\nu}(-2ip_\perp) , \end{aligned} \quad (43.42)$$

where $p_\perp \equiv \sqrt{p_y^2 + p_z^2}$, and $K_{-\nu}$ is the modified Bessel function of the second kind, and ν is

$$\nu \equiv \frac{(\omega - E\gamma)r}{\gamma v} . \quad (43.43)$$

The Fourier transform (43.42) contains negative $\omega < 0$ as well as positive $\omega > 0$ frequencies. However, unlike the case of an observer accelerating in a straight line, equation (43.21), the Bessel function $K_{-\nu}$ with ν given by equation (43.43) is different for $\omega \rightarrow -\omega$, so one cannot conclude that pair creation occurs. In fact, as seen in Exercise 43.4, a circulating observer does not see pair creation.

Exercise 43.6 Pair production seen by an adiabatically accelerating observer. Generalize to an observer who accelerates along the x -direction, at a rate $\kappa(\tau)$ that is slowly varying as opposed to constant. Use a saddle-point approximation to find the approximate Fourier transform, valid as long as the acceleration varies slowly over several acceleration timescales,

$$\left| \frac{d \ln \kappa}{\kappa d\tau} \right| \ll 1, \quad (43.44)$$

which is called the **adiabatic condition**. Note that the observer need not be accelerating themselves, but merely perceives an emitter to be accelerating (redshifting or blueshifting), but it is simpler to deal with an accelerating observer since if the observer accelerates the effect of acceleration is immediate in the observer's frame, whereas if the emitter accelerates the observer sees the acceleration delayed by the time it takes for the wave to travel to the observer.

Solution. A plane wave moving in Minkowski space in the x -direction ($\mathbf{p}_\perp = 0$) satisfies

$$\varphi(t, x, \mathbf{x}_\perp) = e^{i(-Et+px)}, \quad (43.45)$$

the energy E being related to the momentum p and mass m by

$$E = \sqrt{p^2 + m^2}. \quad (43.46)$$

Suppose that the observer accelerates (or decelerates) along the x -direction directly away from the emitter. The observer's 4-velocity $u^m(\tau)$ in the Minkowski frame is

$$u^m \equiv \frac{dx^m}{d\tau} = \{u^t, u^x\} = \{\cosh \alpha, \sinh \alpha\}, \quad (43.47)$$

where $\alpha(\tau)$ is the rapidity. The observer's 4-acceleration κ^m in the Minkowski frame is

$$\kappa^m \equiv \kappa \{\sinh \alpha, \cosh \alpha\}, \quad (43.48)$$

where $\kappa \equiv \sqrt{\kappa_m \kappa^m}$ is the proper acceleration experienced by the observer,

$$\kappa \equiv \frac{d\alpha}{d\tau}. \quad (43.49)$$

The Minkowski coordinates along the path of the observer are

$$x^m \equiv \{t, x\} = \int \frac{dx^m}{d\tau} d\tau = \int \{\cosh \alpha, \sinh \alpha\} d\tau. \quad (43.50)$$

The plane wave (43.45) appears to the observer to be

$$\varphi(\tau) = e^{-i\psi(\tau)}, \quad (43.51)$$

where the phase $\psi(\tau)$ along the path of the observer is

$$\begin{aligned} \psi(\tau) &= \int (E \cosh \alpha - p \sinh \alpha) d\tau \\ &= \int \left(\frac{E-p}{2} e^\alpha + \frac{E+p}{2} e^{-\alpha} \right) \frac{d\alpha}{\kappa(\alpha)}. \end{aligned} \quad (43.52)$$

The Fourier transform of the observed wave (43.51) with respect to the observer's frequencies ω is

$$\varphi_\omega = \int_{-\infty}^{\infty} e^{i(-\psi + \omega\tau)} d\tau . \quad (43.53)$$

The Fourier transform (43.53) involves an integral over all time. In reality, the observer observes Fourier modes over only a finite time, with consequently finite resolution in frequency.

An approximate way to evaluate the integral (43.53) over the observer's proper time is to shift the contour into the complex plane, and to integrate over a saddle point where the phase $\psi - \omega\tau$ is stationary. The physical idea behind the approximation is that the integral over rapidly oscillating regions tends to cancel. An advantage of the approach is that it points to which parts of the domain of integration dominate the integral. The time derivative of the phase $\psi - \omega\tau$ is

$$\frac{d(\psi - \omega\tau)}{d\tau} = \frac{E - p}{2} e^\alpha + \frac{E + p}{2} e^{-\alpha} - \omega . \quad (43.54)$$

The phase is stationary where the derivative vanishes, $d(\psi - \omega\tau)/d\tau = 0$, which happens where

$$e^\alpha = \frac{E + p}{\omega + k} , \quad (43.55)$$

where k is the wavenumber in the observer's frame, satisfying

$$\omega^2 - k^2 = m^2 . \quad (43.56)$$

The stationary-phase condition (43.55) has real solutions for α in which the observed frequency ω is positive and the observed wavenumber k is positive (that is, the observed wavevector is pointed in the positive x -direction, the direction of motion of the wave). However, the Fourier decomposition (43.53) of the wave $\varphi(\tau)$ includes contributions from negative frequencies $\omega < 0$ with corresponding negative wavenumbers $k < 0$. If the positive frequency solution of the stationary-phase condition (43.55) is denoted by the real number α_+ , then the corresponding negative frequency solution α_- is, from equation (43.55) with $\omega \rightarrow -\omega$ and $k \rightarrow -k$, the complex number

$$\alpha_- = \alpha_+ + i\pi . \quad (43.57)$$

As a consequence, the phase of the negative frequency solution in the vicinity of the stationary points approximates minus the phase of the positive frequency solution, but with an extra imaginary part,

$$(\psi - \omega\tau)_- - (\psi - \omega\tau)_+ \approx \omega(\tau_- - \tau_+) \approx \frac{\pi(\alpha_- - \alpha_+)}{\kappa} \approx \frac{i\pi\omega}{\kappa} . \quad (43.58)$$

The second proper time derivative of the phase $\psi - \omega\tau$ is

$$\frac{d^2(\psi - \omega\tau)}{d\tau^2} = \kappa \left(\frac{E - p}{2} e^\alpha - \frac{E + p}{2} e^{-\alpha} \right) , \quad (43.59)$$

which at the stationary-phase point (43.55) is, for either positive k or negative $-k$ frequencies,

$$\frac{d^2(\psi - \omega\tau)_\pm}{d\tau^2} = \mp k\kappa . \quad (43.60)$$

The method of steepest descent approximates the integral by

$$\varphi_{\pm\omega} \approx e^{i(\psi-\omega\tau)_{\pm}} \int_{-\infty}^{\infty} e^{\pm i k \kappa \tau^2} d\tau = e^{i(\psi-\omega\tau)_{\pm}} \sqrt{\frac{\pm i\pi}{k\kappa}} . \quad (43.61)$$

As previously seen in equations (43.10) and (43.23), the key question for particle production is what is the relative amplitude φ_-/φ_+ of negative to positive modes? The ratio is determined by the difference in the imaginary parts of the phases $(\psi - \omega\tau)_{\pm}$, which is given by equation (43.58),

$$\left| \frac{\varphi_-}{\varphi_+} \right| \approx e^{-\pi\omega/\kappa} . \quad (43.62)$$

Once again, the observer perceives a thermal distribution at a temperature $\approx \kappa/(2\pi)$.

Exercise 43.7 Pair production seen by an observer in de Sitter space.

1. Show that an inertial observer in de Sitter space, §10.28, sees pair production with a temperature

$$T = \frac{H}{2\pi} , \quad (43.63)$$

where H is the Hubble parameter.

2. Show that an observer accelerating uniformly at rate κ in a straight line in de Sitter space sees pair production with a temperature

$$T = \frac{\sqrt{H^2 + \kappa^2}}{2\pi} . \quad (43.64)$$

Quantum field theory of Dirac spinors

The units of charge adopted in this Chapter and in the rest of this book are Heaviside, as opposed to the Gaussian units used previously in this book. In natural units $c = \hbar = 1$, electric charge e is dimensionless. The classical interaction (47.1) between a spinor of charge e and the electromagnetic field is proportional to e . Heaviside units scale the unit of charge e so that when the classical interaction (47.1) is reinterpreted as a quantized interaction (47.2), no additional factor appears. Whereas in Gaussian units electric charge e is related to the fine-structure constant α by $e^2/(\hbar c) = \alpha$, in Heaviside units the relation is

$$\frac{e^2}{4\pi\hbar c} = \alpha = 7.2973525643(11) \times 10^{-3} = \frac{1}{137.035999177(21)} , \quad (44.1)$$

where the value is from the NIST 2022 CODATA (NIST, 2022) recommended value of the fine-structure constant, which has been measured experimentally to a part in 10^{10} . The dimensionless value of charge e in Heaviside units is

$$e \equiv \sqrt{4\pi\alpha} = 0.302822120768(23) . \quad (44.2)$$

The value (44.2) is as measured in today's low-energy world. In quantum field theory, the charge e measures the strength with which charges (electrons, specifically) couple to the electromagnetic field. The coupling strength e increases with the energy scale at which it is measured (Mele, 2006) up to the energy of electroweak unification, when the electromagnetic interaction $U(1)_Q$ enlarges to the electroweak group $U(1)_Y \times SU(2)_L$.

As expounded in Chapter 15, d^4x denotes the invariant scalar 4-volume, equation (15.102), not the pseudoscalar 4-volume. The units are $c = \hbar = 1$.

44.1 Dirac spinor field

44.1.1 Dirac Lagrangian

The general relativistic scalar Lagrangian L of a free Dirac spinor field ψ of mass m is

$$L = \bar{\psi} \cdot (\boldsymbol{D} + m) \psi . \quad (44.3)$$

Here $\mathbf{D} \equiv \boldsymbol{\gamma}^n D_n$ is the (torsion-full, in general¹) covariant spacetime derivative, equation (15.31), and $\bar{\psi}$ is the conjugate field defined by equation (41.94). In flat (Minkowski) space the justification for the Dirac Lagrangian (44.3) is that it leads to equations that reproduce ample experiment. Equation (44.3) is the covariant generalization of the flat space Lagrangian of a Dirac field. If units are restored, then the mass is $m/(\hbar c)$. The spinor field ψ has units of $\text{length}^{-3/2}$.

As it stands, the Lagrangian (44.3) is strangely asymmetric in the fields, as it depends only on the velocity $\mathbf{D}\psi$ of the field, not on the velocity $\mathbf{D}\bar{\psi}$ of the conjugate field. Moreover the Lagrangian (44.3) is complex, not real. Symmetry and reality can be restored by symmetrizing the Lagrangian (44.3) with its complex conjugate. The covariant spacetime derivative $\mathbf{D} \equiv \boldsymbol{\gamma}_n D^n$ has real coefficients D^n in an orthonormal basis $\boldsymbol{\gamma}_n$. For any multivector $\mathbf{a}$ whose coefficients are real in an orthonormal basis, the complex conjugate of $\bar{\psi} \cdot \mathbf{a}\psi$ is, equation (41.125),

$$(\bar{\psi} \cdot \mathbf{a}\psi)^* = -\psi \cdot \mathbf{a}\bar{\psi} . \quad (44.4)$$

The symmetrized, real Lagrangian is thus

$$L = \frac{1}{2} \bar{\psi} \cdot (\mathbf{D} + m) \psi - \frac{1}{2} \psi \cdot (\mathbf{D} + m) \bar{\psi} . \quad (44.5)$$

Despite being asymmetric and complex, the original Dirac Lagrangian (44.3) does yield the correct Dirac equations because the imaginary part of the Lagrangian integrates to a surface term, by Gauss' theorem (15.164),

$$i \int \text{Im}(\bar{\psi} \cdot \mathbf{D}\psi) d^4x = \frac{1}{2} \int (\bar{\psi} \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}\bar{\psi}) d^4x = \frac{1}{2} \oint \bar{\psi} \cdot \boldsymbol{\gamma}_n \psi d^3x^n , \quad (44.6)$$

and therefore has no effect on the equations of motion.

The original complex Dirac Lagrangian (44.3) is in (super-)Hamiltonian form $p \cdot \mathbf{D}q - H$ with coordinates $q = \psi$, momenta $p = \bar{\psi}$, and (super-)Hamiltonian

$$H = -m\bar{\psi} \cdot \psi . \quad (44.7)$$

Varying the action with complex Dirac Lagrangian (44.3) with respect to the field ψ and its conjugate momentum $\bar{\psi}$ yields, with the help of Gauss' theorem (15.164) to integrate $\delta(\mathbf{D}\psi) = \mathbf{D}(\delta\psi)$ by parts,

$$\delta S = \oint \bar{\psi} \cdot \boldsymbol{\gamma}_n \delta\psi d^3x^n + \int [\delta\bar{\psi} \cdot (\mathbf{D}\psi + m\psi) + (\mathbf{D}\bar{\psi} + m\bar{\psi}) \cdot \delta\psi] d^4x . \quad (44.8)$$

The resulting Hamilton equations of motion are the Dirac equations

$$(\mathbf{D} + m)\psi = 0 , \quad (44.9a)$$

$$(\mathbf{D} + m)\bar{\psi} = 0 . \quad (44.9b)$$

In flat (Minkowski) space, the solutions of the free Dirac equations (44.9) are plane waves. The solutions are most straightforward to obtain in the rest frame, where the spinor ψ is one of the Dirac basis spinors $\psi_{\uparrow}$ or $\psi_{\downarrow}$ (with spin either up $\uparrow$ or down $\downarrow$), equations (14.108), and the covariant derivative reduces to the time

¹ Gauge fields such as electromagnetism are necessarily defined in terms of torsion-free derivatives, §16.5, but spinor fields contribute to, and experience, torsion, Exercises 16.7 and 16.9.

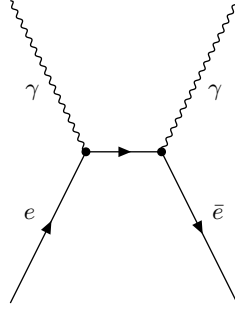


Figure 44.1 Feynman diagram illustrating the Stueckelberg-Feynman interpretation of antiparticles as negative mass particles moving backwards in time (Stueckelberg, 1941; Feynman, 1949). The diagram shows an electron e and positron $\bar{e}$ annihilating into two photons (conservation of energy-momentum prohibits annihilation into one photon). The arrows represent the direction of charge.

derivative $\mathbf{D} \rightarrow \boldsymbol{\gamma}^0 \partial_0$. The conjugate of a spinor is $\bar{\psi} \equiv C\psi^*$, equation (41.94), and the expression (41.86) for C says that conjugation flips the $\psi_{\top}$ and $\psi_{\perp}$ states. The Dirac equations (44.9) in the rest frame become

$$(-i\partial_0 + m)\psi_{\top} = 0, \quad (i\partial_0 + m)\psi_{\perp} = 0, \quad (44.10a)$$

$$(-i\partial_0 + m)\psi_{\perp}^* = 0, \quad (i\partial_0 + m)\psi_{\top}^* = 0, \quad (44.10b)$$

whose solutions are

$$\psi_{\top} \propto e^{-imt}, \quad \psi_{\perp} \propto e^{imt}, \quad (44.11a)$$

$$\psi_{\perp}^* \propto e^{-imt}, \quad \psi_{\top}^* \propto e^{imt}. \quad (44.11b)$$

While the solution for $\psi_{\top}$ has positive mass m , the solution for $\psi_{\perp}$ appears to have negative mass m . This is Dirac's celebrated problem of negative mass states (Bjorken and Drell, 1964). On the other hand, the complex conjugate $\psi_{\perp}^*$ of the negative mass state $\psi_{\perp}$ has positive mass.

The idea that the negative mass states are antiparticles dates to Stueckelberg (1941), who proposed that **an antiparticle is a negative mass particle moving backwards in time**, as illustrated in Figure 44.1. To ensure that the propagator propagates positive-energy particles forward in time and negative-energy particles backward in time, positive-energy solutions are associated with operators that destroy particles, while negative-energy solutions are associated with operators that create particles.

44.1.2 Dirac (super-)Hamiltonian

Although the Dirac Lagrangian (44.3) yields the correct equations of motion (44.9) (and the symmetrized Lagrangian (44.5) yields the same equations), it is not altogether satisfactory. The problem is that the Lagrangians (44.3) or (44.5) assume a priori that the momentum conjugate to ψ is its conjugate $\bar{\psi}$. In

a “correct” Hamiltonian approach, the coordinates and momenta are independent fields, and any relation between them should emerge as an equation of motion.

The solution to the problem is to introduce a momentum π conjugate to the field ψ , with no a priori relation between π and ψ , and to treat the fields ψ and π and their conjugates $\bar{\psi}$ and $\bar{\pi}$ as 4 independent fields. In terms of the 4 fields, the Dirac Lagrangian, symmetrized with its complex conjugate so as to make it real, is

$$L = \frac{1}{2}\pi \cdot \mathbf{D}\psi - \frac{1}{2}\bar{\pi} \cdot \mathbf{D}\bar{\psi} - H , \quad (44.12)$$

with a (super-)Hamiltonian H that resembles the Hamiltonian (42.9) of a simple harmonic oscillator,

$$\boxed{H = \frac{1}{2}m(-\bar{\psi} \cdot \psi + \bar{\pi} \cdot \pi)} . \quad (44.13)$$

The momentum conjugate to ψ is $\frac{1}{2}\pi$, while the momentum conjugate to $\bar{\psi}$ is $-\frac{1}{2}\bar{\pi}$. The Dirac Hamiltonian (44.13) is consistent with, though does not follow uniquely from, the original Hamiltonian (44.7). The justification for the Hamiltonian (44.13) is that it yields the correct Dirac equations of motion, along with $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ as constraint equations, equations (44.15).

Varying the Dirac action with Lagrangian (44.12) with respect to the coordinates ψ and $\bar{\psi}$ and their conjugate momenta π and $-\bar{\pi}$ yields

$$\begin{aligned} \delta S = & \frac{1}{2} \oint (\pi \cdot \boldsymbol{\gamma}_n \delta\psi - \bar{\pi} \cdot \boldsymbol{\gamma}_n \delta\bar{\psi}) d^3x^n \\ & + \frac{1}{2} \int [\delta\pi \cdot (\mathbf{D}\psi + m\bar{\pi}) - \delta\bar{\pi} \cdot (\mathbf{D}\bar{\psi} + m\pi) + (\mathbf{D}\pi + m\bar{\psi}) \cdot \delta\psi - (\mathbf{D}\bar{\pi} + m\psi) \cdot \delta\bar{\psi}] d^4x . \end{aligned} \quad (44.14)$$

The resulting Hamilton’s equations can be written

$$(\mathbf{D} + m)(\bar{\pi} + \psi) = 0 , \quad (\mathbf{D} + m)(\pi + \bar{\psi}) = 0 , \quad (44.15a)$$

$$(\mathbf{D} - m)(\bar{\pi} - \psi) = 0 , \quad (\mathbf{D} - m)(\pi - \bar{\psi}) = 0 . \quad (44.15b)$$

Hamilton’s equations (44.15) appear to describe solutions with both signs of mass m . If the standard choices $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ are imposed initially, then the $-m$ Dirac equations (44.15b) ensure that $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ thereafter. The $+m$ Dirac equations (44.15a) then reproduce the usual Dirac equations (44.9). The conditions $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ thus emerge as constraint equations. The original Hamiltonian (44.7) can be interpreted as an effective Hamiltonian, valid after the solution $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ to the equation of motion is imposed.

44.1.3 Conserved Dirac number current

The Dirac Lagrangians (44.3), (44.5), or (44.12) are unchanged if the field and its conjugate are changed by a global (independent of spacetime) transformation by opposing complex phases, $\psi \rightarrow e^{-i\theta}\psi$ and $\bar{\psi} \rightarrow e^{i\theta}\bar{\psi}$, and likewise the conjugate momenta are changed as $\pi \rightarrow e^{i\theta}\pi$ and $\bar{\pi} \rightarrow e^{-i\theta}\bar{\pi}$. In infinitesimal form, this transformation is

$$\psi \rightarrow \psi - i\theta\psi , \quad \bar{\psi} \rightarrow \bar{\psi} + i\theta\bar{\psi} , \quad \pi \rightarrow \pi + i\theta\pi , \quad \bar{\pi} \rightarrow \bar{\pi} - i\theta\bar{\pi} . \quad (44.16)$$

The corresponding conserved Noether current, equation (16.17), is

$$n^m = \frac{1}{2}i (\pi \cdot \gamma^m \psi + \bar{\pi} \cdot \gamma^m \bar{\psi}) . \quad (44.17)$$

The relative sign of the two terms on the right hand side of equation (44.17) is positive because the fields vary with opposite sign under the transformation (44.16), $\delta \bar{\psi} = -\delta \psi$. Imposing the positive mass conditions $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ brings the Noether current to

$$n^m = \frac{1}{2}i (\bar{\psi} \cdot \gamma^m \psi + \psi \cdot \gamma^m \bar{\psi}) . \quad (44.18)$$

The two terms on the right hand side of equation (44.18) (or on the right hand side of equation (44.25)) are the same since

$$\psi \cdot \gamma^m \bar{\psi} = -(\gamma^m \bar{\psi}) \cdot \psi = \bar{\psi} \cdot \gamma^m \psi , \quad (44.19)$$

the last equality following from equation (41.40), so equation (44.18) simplifies to

$$n^m = i \bar{\psi} \cdot \gamma^m \psi . \quad (44.20)$$

The Dirac current (44.20) is covariantly conserved in accordance with Noether's theorem, equation (16.18),

$$\mathring{D}_m n^m = 0 . \quad (44.21)$$

The factor i in the Dirac current (44.20) is introduced so that the time component n^0 is a positive number,

$$n^0 = \psi^\dagger \psi = \bar{\psi}^\dagger \bar{\psi} , \quad (44.22)$$

where, in accordance with equation (41.102), $\psi^\dagger \equiv \psi^{*\top}$ is the Hermitian conjugate of ψ . The Dirac current n^m is interpreted as a conserved probability number current with a positive density n^0 .

If the current (44.20) is written

$$\mathbf{n} = \gamma_m n^m = i \gamma_m \bar{\psi} \cdot \gamma^m \psi , \quad (44.23)$$

then the probability conservation equation (44.21) is

$$\mathring{D} \cdot \mathbf{n} = 0 . \quad (44.24)$$

If conversely to equation (44.18) the negative mass conditions $\pi = -\bar{\psi}$ and $\bar{\pi} = -\psi$ were imposed, then the sign of the Noether current (44.18) would flip,

$$n^m = -\frac{1}{2}i (\bar{\psi} \cdot \gamma^m \psi + \psi \cdot \gamma^m \bar{\psi}) , \quad (44.25)$$

and the number density n^0 would be negative instead of positive. In the real world, spinors always carry charge (of one kind or another, electric or otherwise). For charged spinors, §44.2, the positive mass conditions can be imposed on spinors of opposite charges, ensuring positive density of both. Ultimately, the problem of negative mass particles is resolved in quantum field theory, §44.4.3, where a negative-mass charged spinor going backwards in time is indistinguishable from a positive-mass oppositely-charged antispinor going forwards in time.

If the Dirac spinor is null, $\bar{\psi} \cdot \psi = \bar{\psi} \cdot I\psi = 0$, then the free Dirac equations preserve chirality. In this case

the right- and left-handed components of the current $\mathbf{n}$ are separately conserved. It follows that, for a free null spinor in the absence of interactions, the pseudovector current

$$n_5^m \equiv i \bar{\psi} \cdot \gamma_5 \boldsymbol{\gamma}^m \psi \quad (44.26)$$

is also conserved.

44.2 Dirac field with electromagnetism

Electromagnetism emerges from the hypothesis that the Lagrangian is invariant under a symmetry that rotates the Dirac field ψ by a complex phase proportional to the electric charge e of the field. This kind of transformation is called a **gauge transformation**. The three forces of the Standard Model, §48.1, the electromagnetic, weak, and strong forces, all emerge from gauge transformations. Electromagnetism is the simplest gauge field, based on the 1-dimensional unitary group $U(1)_Q$ of rotations about a circle.

Under an electromagnetic gauge transformation, a Dirac field ψ of charge e , and its conjugate field $\bar{\psi}$, which is proportional to the complex conjugate of the field, equation (41.94), and likewise their conjugate momenta π and $\bar{\pi}$, transform as

$$\psi \rightarrow e^{-ie\theta} \psi, \quad \bar{\psi} \rightarrow e^{ie\theta} \bar{\psi}, \quad \pi \rightarrow e^{ie\theta} \pi, \quad \bar{\pi} \rightarrow e^{-ie\theta} \bar{\pi}, \quad (44.27)$$

where the phase $\theta(x)$ is some arbitrary function of spacetime. The charge e is dimensionless, and the charge $-e$ of the conjugate field $\bar{\psi}$ must be minus that of the field. To ensure that the Dirac Lagrangian (44.5) remains invariant under the gauge transformation, the derivative $\boldsymbol{D}$ must be replaced by a gauge-covariant derivative $\boldsymbol{D} \pm ie\mathbf{A}$ which, when acting on the field and its conjugate, transforms under the gauge transformation (44.27) as

$$(\boldsymbol{D} + ie\mathbf{A})\psi \rightarrow e^{-ie\theta}(\boldsymbol{D} + ie\mathbf{A})\psi, \quad (\boldsymbol{D} - ie\mathbf{A})\bar{\psi} \rightarrow e^{ie\theta}(\boldsymbol{D} - ie\mathbf{A})\bar{\psi}. \quad (44.28)$$

The conjugate momenta π and $\bar{\pi}$ transform respectively as $\bar{\psi}$ and ψ . The gauge-covariant derivative transforms correctly provided that the gauge field $\mathbf{A}$ transforms under the electromagnetic gauge transformation (44.27) as

$$\mathbf{A} \rightarrow \mathbf{A} + \boldsymbol{D}\theta. \quad (44.29)$$

The gauge field $\mathbf{A}$ is the electromagnetic potential.

The general relativistic scalar Lagrangian L of a Dirac spinor field ψ of mass m and charge e is obtained from the uncharged Dirac Lagrangian (44.3) by changing the (torsion-full, in general) covariant derivative $\boldsymbol{D}$ to the gauge covariant derivative $\boldsymbol{D} + ie\mathbf{A}$,

$$L = \bar{\psi} \cdot (\boldsymbol{D} + ie\mathbf{A} + m) \psi. \quad (44.30)$$

Symmetrized with its complex conjugate, the charged Dirac Lagrangian (44.30) is

$$L = \frac{1}{2} \bar{\psi} \cdot (\boldsymbol{D} + ie\mathbf{A} + m) \psi - \frac{1}{2} \psi \cdot (\boldsymbol{D} - ie\mathbf{A} + m) \bar{\psi}. \quad (44.31)$$

If the momentum π conjugate to ψ is treated as a distinct field as in §44.1.2, then the charged Dirac Lagrangian is

$$L = \frac{1}{2}\pi \cdot (\mathbf{D} + ie\mathbf{A})\psi - \frac{1}{2}\bar{\pi} \cdot (\mathbf{D} - ie\mathbf{A})\bar{\psi} + \frac{1}{2}m(-\bar{\pi} \cdot \pi + \bar{\psi} \cdot \psi) . \quad (44.32)$$

Varying the action with Lagrangian (44.32) yields Hamilton's equations for a charged Dirac field,

$$(\mathbf{D} + ie\mathbf{A} + m)(\bar{\pi} + \psi) = 0 , \quad (\mathbf{D} - ie\mathbf{A} + m)(\pi + \bar{\psi}) = 0 , \quad (44.33a)$$

$$(\mathbf{D} + ie\mathbf{A} - m)(\bar{\pi} - \psi) = 0 , \quad (\mathbf{D} - ie\mathbf{A} - m)(\pi - \bar{\psi}) = 0 , \quad (44.33b)$$

generalizing the earlier uncharged equations (44.15). Once again, the $+m$ conditions $\pi = \bar{\psi}$ and $\bar{\pi} = \psi$ emerge as constraint equations if the conjugate momenta π and $\bar{\pi}$ are treated as fields independent from ψ and $\bar{\psi}$. Under the $+m$ conditions, the Dirac equations (44.33) reduce to

$$(\mathbf{D} + ie\mathbf{A} + m)\psi = 0 , \quad (44.34a)$$

$$(\mathbf{D} - ie\mathbf{A} + m)\bar{\psi} = 0 . \quad (44.34b)$$

The Dirac equation (44.34b) for the conjugate field $\bar{\psi}$ looks like that (44.34a) for the field ψ with the same mass but with opposite charge e .

The charged Dirac field has an electric current $\mathbf{j}$ given by the product of the charge e and the conserved number current $\mathbf{n} \equiv \boldsymbol{\gamma}_m n^m$, equation (44.20),

$$\mathbf{j} \equiv e\mathbf{n} . \quad (44.35)$$

Like the number current, equation (44.21), the electric current $\mathbf{j}$ is covariantly conserved,

$$\mathring{\mathbf{D}} \cdot \mathbf{j} = 0 . \quad (44.36)$$

Current conservation (44.36) is a consequence of the invariance of the Lagrangian (44.31) under an electromagnetic gauge transformation (44.27). The electromagnetic contribution to the Dirac Lagrangian (44.31) can be interpreted as describing the interaction between the electromagnetic field $\mathbf{A}$ and the Dirac electric current $\mathbf{j}$,

$$L_{\text{int}} = ie\bar{\psi} \cdot \mathbf{A}\psi = \mathbf{A} \cdot \mathbf{j} . \quad (44.37)$$

Resolved into components $\psi_{\top}$ and $\psi_{\perp}$ in the Dirac representation, §14.8, the Dirac equations (44.34) become, generalizing equations (44.10),

$$(-iD_0 + eA_0 + m)\psi_{\top} = -\sigma_a(D_a + ieA_a)\psi_{\perp} , \quad (iD_0 - eA_0 + m)\psi_{\perp} = -\sigma_a(D_a + ieA_a)\psi_{\top} , \quad (44.38a)$$

$$(-iD_0 - eA_0 + m)\psi_{\perp}^* = -\sigma_a^*(D_a - ieA_a)\psi_{\top}^* , \quad (iD_0 + eA_0 + m)\psi_{\top}^* = -\sigma_a^*(D_a - ieA_a)\psi_{\perp}^* . \quad (44.38b)$$

The charge-conjugate Dirac equations (44.38b) are complex conjugates (with respect to i) of the parent equations (44.38a). As discussed in §14.8, a Dirac spinor ψ contains two components, which in the rest frame are $\psi_{\top}$ and $\psi_{\perp}$, that cannot be transformed into each other by any proper Lorentz transformation. The two components describe particles and antiparticles. Lorentz-transformed out of the rest frame, particles and antiparticles are each linear combinations of both $\psi_{\top}$ and $\psi_{\perp}$, but still those combinations cannot be transformed into each other: for particles, $\psi_{\top}$ dominates, while for antiparticles $\psi_{\perp}$ dominates. The

first pair (44.38a) of Dirac equations describes the evolution of particles, where $\psi_{\top}$ dominates; the pair of equations are coupled first-order differential equations for $\psi_{\top}$ and $\psi_{\perp}$, which combine to yield a second-order equation for $\psi_{\top}$. Likewise the second pair (44.38b) describes the evolution of antiparticles, where the negative-mass component $\psi_{\perp}$, or physically its positive-mass complex conjugate $\psi_{\perp}^*$, dominates; the second pair (44.38b) combine to yield a second-order equation for $\psi_{\perp}^*$. The charged Dirac equations (44.38) confirm the earlier inference from equations (44.34) that particles and antiparticles have opposite electric charges.

Resolved instead into chiral components ψ_R and ψ_L , §41.2, the Dirac equations (44.34) are

$$[-D_0 - ieA_0 + \sigma_a(D_a + iA_a)]\psi_L = -m\psi_R, \quad [-D_0 - ieA_0 - \sigma_a(D_a + iA_a)]\psi_R = m\psi_L, \quad (44.39a)$$

$$[-D_0 + ieA_0 - \sigma_a^*(D_a - iA_a)]\psi_R^* = m\psi_L^*, \quad [-D_0 + ieA_0 + \sigma_a^*(D_a - iA_a)]\psi_L^* = -m\psi_R^*. \quad (44.39b)$$

Again, the charge-conjugate Dirac equations (44.39b) are complex conjugates (with respect to i) of the parent equations (44.39a).

44.3 Particles and antiparticles

The question of whether a Dirac spinor ψ describes a particle or antiparticle can be decided from the sign of the effective Dirac Hamiltonian, equation (44.7),

$$H = -m\bar{\psi} \cdot \psi = im\psi^\dagger \gamma_0 \psi = -m(\psi_{\top}^\dagger \psi_{\top} - \psi_{\perp}^\dagger \psi_{\perp}). \quad (44.40)$$

Whereas the number density $n^0 \equiv i\bar{\psi} \cdot \gamma^0 \psi = \psi^\dagger \psi$ of the Dirac field is always positive, equation (44.22), the scalar product $\bar{\psi} \cdot \psi$ can be either positive or negative. If the spinor is a particle ($\psi_{\top}$ dominates), then $\bar{\psi} \cdot \psi$ is positive, and the Hamiltonian (44.40) with positive m describes a timelike field. If on the other hand the spinor is an antiparticle ($\psi_{\perp}$ dominates), then $\bar{\psi} \cdot \psi$ is negative, and the Hamiltonian would appear to describe a spacelike field. (As usual in this book, do not confuse the scalar super-Hamiltonian (44.40) with the conventional Hamiltonian, which is the time component of a 4-vector.)

The antisymmetry of the Dirac spinor scalar product means that the Hamiltonian (44.40) can be rewritten

$$H = m\psi \cdot \bar{\psi}. \quad (44.41)$$

This does not resolve the problem that, if the antiparticle component dominates, the Hamiltonian (44.41) is still positive, for positive m , hence spacelike. A timelike Hamiltonian with positive mass m when the antiparticle field $\bar{\psi}$ dominates can be obtained by taking its PT conjugate, yielding the CPT -conjugate field. Let fields with an underbar $\underline{\psi}$ denote the PT -conjugate fields obtained by pre-multiplying by the pseudoscalar I , equation (41.136),

$$PT: \quad \underline{\psi} = I\psi, \quad CPT: \quad \bar{\underline{\psi}} \equiv I\bar{\psi}. \quad (44.42)$$

The fields $\underline{\psi}$ and $\bar{\underline{\psi}}$ are charge conjugates of each other, equation (41.94), since $CI^* = IC$. Since the pseudoscalar satisfies $I^2 = -1$ and I commutes with the spinor metric ε , the Hamiltonian of the CPT -conjugate field $\bar{\underline{\psi}}$ is

$$H = -m\underline{\psi} \cdot \bar{\underline{\psi}}, \quad (44.43)$$

which is timelike when $\psi \cdot \bar{\psi}$ is positive, that is, when the CPT -conjugate field $\bar{\psi}$ dominates. Note that introducing an additional phase factor, such as i , into the definition of the PT -conjugate fields makes no difference to the Hamiltonian, because the opposing phase factors in $\underline{\psi}$ and $\bar{\underline{\psi}}$ cancel each other.

44.3.1 C , P , and T symmetries

The collection of Dirac equations for a spinor ψ , its charge conjugate $\bar{\psi}$, and their PT conjugates $\underline{\psi}$ and $\bar{\underline{\psi}}$ are

$$(\mathbf{D} + ie\mathbf{A} + m)\psi = 0 , \quad (44.44a)$$

$$PT : (\mathbf{D} + ie\mathbf{A} - m)\underline{\psi} = 0 , \quad (44.44b)$$

$$C : (\mathbf{D} - ie\mathbf{A} + m)\bar{\psi} = 0 , \quad (44.44c)$$

$$CPT : (\mathbf{D} - ie\mathbf{A} - m)\bar{\underline{\psi}} = 0 . \quad (44.44d)$$

The PT -conjugate equations (44.44b) and (44.44d) are obtained by commuting the PT operator I through their parent equations (44.44a) and (44.44c), and noting that I anticommutes with the basis vectors γ^n . The PT -conjugate Dirac equations (44.44b) and (44.44d) appear to be flipped in mass m compared to their parent Dirac equations (44.44a) and (44.44c), but if the equations are expanded in terms of components $\psi_\top$ and $\psi_\perp$, as in equations (44.38), the PT -conjugate equations are identical to their parent counterparts.

Despite the apparently differing signs of charge e and mass m , the four sets of Dirac equations (44.44) are equivalent to each other, an equivalence that is manifest when the equations are expanded in components, equations (44.38). The equivalences express symmetry of the charged Dirac equations with respect to the discrete operations of spacetime reversal PT and charge conjugation C . The PT symmetry says that the transformation $\gamma_m \rightarrow -\gamma_m$ and $\psi \rightarrow I\psi \equiv \underline{\psi}$ leaves the Dirac equation unchanged. The C symmetry says that the transformation $e \rightarrow -e$ and $\psi \rightarrow C\psi^* \equiv \bar{\psi}$ leaves the Dirac equation unchanged.

The Dirac equations are also symmetric with respect to the parity operation P . The P symmetry says that flipping the spatial axes $\gamma_a \rightarrow -\gamma_a$ and transforming $\psi \rightarrow \gamma_0\psi$ leaves the Dirac equation unchanged.

Electromagnetic, colour, and gravitational interactions all respect C , P , and T symmetries, but weak interactions violate them. Weak interactions act only on left-handed particles (and right-handed antiparticles), not their opposite-chiral counterparts. A parity transformation flips chirality (it flips momentum while leaving spin unchanged), so weak interactions violate parity symmetry maximally. The excess of matter (baryons and leptons) over antimatter (antibaryons and antileptons) in the Universe suggests that T -violating processes took place during the early Universe.

Although C , P , and T may be individually violated, the combination CPT appears to be a general symmetry of Nature. There is a CPT theorem premised on the proposition that Lorentz transformations in (3+1)-dimensional spacetime can be analytically continued to spatial rotations in 4 spatial dimensions. A spatial rotation by π in the Euclideanized tz plane sends $t \rightarrow -t$ and $z \rightarrow -z$, equivalent to a combination of time reversal and parity reversal in 3+1 spacetime dimensions. A spatial rotation preserves scalars, in particular the scalar (super-)Hamiltonian (44.40); for the Hamiltonian to remain a scalar in 3+1 dimensions, the transformation must be CPT , equation (44.43).

44.4 Quantization of the Dirac spinor field

44.4.1 Vertices are inner products and propagators are outer products

Feynman diagrams have two basic ingredients, vertices where interactions occur, and lines joining those vertices, as illustrated for example in Figure 44.1. The lines can be either external lines, with prescribed in- or out-states, or else internal lines, called propagators, which propagate an intermediate state between two vertices.

Interaction vertices of Dirac spinors involve inner products of spinors, such as the interaction term $ie\bar{\psi} \cdot \mathbf{A}\psi$, equation (44.30), between a Dirac spinor ψ and the electromagnetic field $\mathbf{A}$. Internal lines of Dirac spinors involve outer products of spinors, as can be inferred from the expression (44.80) for the Dirac propagator, since, as will now be shown, the $\mp i\mathbf{p} + m$ factors in the integrand of (44.80) are outer products of spinors, equations (44.48).

To see that the $\mp i\mathbf{p} + m$ factors in the Dirac propagator (44.80) are outer products of spinors, start with the fact that inner and outer products of Dirac basis spinors satisfy, with indices b and b' running over $\uparrow, \downarrow$, $\uparrow, \downarrow$, $\uparrow, \downarrow$,

$$\bar{\epsilon}_b \cdot \epsilon_{b'} = -i\epsilon_b^\top \gamma_0 \epsilon_{b'} = \pm \delta_{bb'} , \quad (44.45a)$$

$$\epsilon_b \bar{\epsilon}_{b'} \cdot = -i\epsilon_b \epsilon_{b'}^\top \gamma_0 = \pm 1_{bb'} . \quad (44.45b)$$

The inner product (44.45a) is a complex number equal to either 0 or 1. The outer product (44.45b) $1_{bb'}$ is a 4×4 Dirac matrix with 1 in the (bb') th entry (not the unit matrix!); the $\pm$ signs are $+$ and $-$ for states b of respectively positive ($\uparrow$) and negative ($\downarrow$) mass. Recall that conjugation flips Dirac indices, equation (41.100); for example the conjugate of $\epsilon_{\uparrow\uparrow}$ is $\bar{\epsilon}_{\uparrow\uparrow} = -\epsilon_{\downarrow\downarrow}$. Thus negative-mass ($\downarrow$) spinors can be subsumed into conjugate spinors $\bar{\epsilon}_{pb}$. Summing outer products (44.45b) over spins $b = b'$ gives

$$\sum_{b=\uparrow,\downarrow} \epsilon_{\uparrow b} \bar{\epsilon}_{\uparrow b} \cdot = \sum_{b=\uparrow,\downarrow} \bar{\epsilon}_{\downarrow b} \epsilon_{\downarrow b} \cdot = \frac{-i\gamma_0 + 1}{2} , \quad (44.46a)$$

$$- \sum_{b=\uparrow,\downarrow} \bar{\epsilon}_{\downarrow b} \epsilon_{\downarrow b} \cdot = - \sum_{b=\uparrow,\downarrow} \epsilon_{\uparrow b} \bar{\epsilon}_{\uparrow b} \cdot = \frac{i\gamma_0 + 1}{2} , \quad (44.46b)$$

where the 1 on the right hand side denotes the unit multivector. Since not only spin but also energy-momentum is quantized, spinor quanta must be labelled with a momentum index p as well as a spin index b . A spinor quantum with arbitrary nonzero spatial momentum p_a and spin b running over $\uparrow$ and $\downarrow$ along any prescribed direction can be obtained by Lorentz-transforming a rest-frame spinor quantum by a Lorentz rotor R ,

$$\epsilon_{pb} \equiv \sqrt{2m} R \epsilon_{\uparrow\uparrow} , \quad \bar{\epsilon}_{pb} \cdot \equiv \sqrt{2m} \bar{\epsilon}_{\uparrow\uparrow} \cdot \tilde{R} , \quad (44.47a)$$

$$\bar{\epsilon}_{pb} \equiv \sqrt{2m} R \bar{\epsilon}_{\uparrow\uparrow} , \quad \epsilon_{pb} \equiv \sqrt{2m} \epsilon_{\uparrow\uparrow} \cdot \tilde{R} . \quad (44.47b)$$

The factors of $\sqrt{2m}$ are introduced to ensure that the boosted spinors remain finite in the massless limit $m \rightarrow 0$, as in equations (44.48). Basis spinors ϵ_b are dimensionless, and, thanks to the factors of $\sqrt{2m}$, boosted spinors ϵ_{pb} have dimension of $\text{mass}^{1/2}$, or equivalently $\text{length}^{-1/2}$. Boosted into a frame where the

spinor energy-momentum is $p = \{p^0, p^a\}$ satisfying the on-shell condition $p^0 = E_p = \sqrt{p_a p^a + m^2}$ with positive energy E_p , the sum over spins of outer products of boosted spinors is

$$\sum_{b=\uparrow, \downarrow} \epsilon_{pb} \bar{\epsilon}_{pb} \cdot = -i\mathbf{p} + m , \quad (44.48a)$$

$$- \sum_{b=\uparrow, \downarrow} \bar{\epsilon}_{pb} \epsilon_{pb} \cdot = i\mathbf{p} + m , \quad (44.48b)$$

where $\mathbf{p} \equiv \boldsymbol{\gamma}_n p^n$. The expressions (44.48) remain finite in the massless limit $m \rightarrow 0$ thanks to the factor of $\sqrt{2m}$ introduced in the definition (44.47) of boosted spinors. The minus sign in equation (44.48b) ensures that the Dirac propagator (44.81) is Lorentz invariant, as it must be. The minus sign cannot be eliminated by adjusting the basis spinor elements by a phase, because any such phase cancels in the outer product of a spinor with its conjugate.

44.4.2 Spinor destruction and creation operators

The standard formalism of quantum field theory relies on the concept of destruction and creation operators. One of the beautiful features of spinors is that they already behave like destruction and creation operators just as they are: column spinors are destruction operators, and row spinors are creation operators. The correct fermionic exclusion behaviour is built in: the prescription that a column spinor cannot be multiplied by a column spinor prevents two spinors from being destroyed out of the same state; and the prescription that a row spinor cannot be multiplied by a row spinor prevents two spinors from being created in the same state. The two building blocks of Feynman diagrams, vertices and lines, are realised as inner and outer products respectively. A row spinor multiplying a column spinor is an inner product, representing a vertex, while a column spinor multiplying a row spinor is an outer product, representing a line:

$$\left(\begin{array}{c} \end{array} \right) \left(\begin{array}{c} \end{array} \right) = \left(\right) \quad \text{inner product = vertex (interaction) ,} \quad (44.49a)$$

$$\left(\begin{array}{c} \end{array} \right) \left(\begin{array}{c} \end{array} \right) = \left(\begin{array}{c} \end{array} \right) \quad \text{outer product = line (propagator) ,} \quad (44.49b)$$

$$\left(\begin{array}{c} \end{array} \right) \left(\begin{array}{c} \end{array} \right) = 0 \quad \text{forbidden ,} \quad (44.49c)$$

$$\left(\begin{array}{c} \end{array} \right) \left(\begin{array}{c} \end{array} \right) = 0 \quad \text{forbidden .} \quad (44.49d)$$

Quantum field theory postulates a special state, the vacuum, that (at least in Minkowski space) contains no particles, and the vacuum-to-vacuum amplitude is one, $\langle 0|0 \rangle = 1$. Spinor creation and destruction operators $\bar{\epsilon}_{pb} \cdot$ and ϵ_{pb} and their conjugate antispinor creation and destruction operators $\epsilon_{pb} \cdot$ and $\bar{\epsilon}_{pb}$ are defined to

yield zero when acting on respectively the out-vacuum $\langle 0|$ and in-vacuum $|0\rangle$,

$$\langle 0|\bar{\epsilon}_{pb} \cdot = 0, \quad \epsilon_{pb}|0\rangle = 0, \quad (44.50a)$$

$$\langle 0|\epsilon_{pb} \cdot = 0, \quad \bar{\epsilon}_{pb}|0\rangle = 0. \quad (44.50b)$$

For spinors, it is convenient to introduce, in addition to the vacuum state, generic 1-particle out- and in-states $\langle 1|$ and $|1\rangle$ analogous to the 0-particle out- and in-states $\langle 0|$ and $|0\rangle$, with the defining property that, analogously to equations (44.50),

$$\langle 1|\epsilon_{pb} = 0, \quad \bar{\epsilon}_{pb} \cdot |1\rangle = 0, \quad (44.51a)$$

$$\langle 1|\bar{\epsilon}_{pb} = 0, \quad \epsilon_{pb} \cdot |1\rangle = 0. \quad (44.51b)$$

The device of sandwiching a product of a pair of (anti)spinor operators between 0-particle out- and in-states ensures that only outer products are nonzero; and similarly the device of sandwiching a product of a pair of (anti)spinor operators between 1-particle out- and in-states ensures that only inner products are nonzero. Outer products produce a line in a Feynman diagram; a line begins with nothing and ends with nothing. Inner products produce a vertex in a Feynman diagram; a spinor vertex begins with a spinor and ends with another spinor.

The state comprising zero spinors in momentum state p and spin state b , without regard to the possible presence of particles in other spin states, is denoted $|0_{pb}\rangle$. Similarly the state comprising zero antispinors in momentum state p and spin state b , without regard to the possible presence of particles in other spin states, is denoted $|\bar{0}_{pb}\rangle$. The creation operator $\bar{\epsilon}_{pb} \cdot$ acting on the vacuum in-state $|0_{pb}\rangle$ is defined to create the one-particle in-state $|1_{pb}\rangle$, correctly normalized so that $\langle 1_{pb}|1_{pb}\rangle = \langle 1_{pb}|\bar{\epsilon}_{pb} \cdot |0_{pb}\rangle = 1$. Analogously to the case of a real scalar field, equations (42.70), Dirac destruction and creation operators ϵ_{pb} and $\bar{\epsilon}_{pb} \cdot$ and their conjugates $\epsilon_{pb} \cdot$ and $\bar{\epsilon}_{pb}$ satisfy

$$\langle 1_{pb}| = \langle 0_{pb}|\epsilon_{pb}, \quad \bar{\epsilon}_{pb} \cdot |0_{pb}\rangle = |1_{pb}\rangle, \quad (44.52a)$$

$$\langle 0_{pb}| = \langle 1_{pb}|\bar{\epsilon}_{pb} \cdot, \quad \epsilon_{pb}|1_{pb}\rangle = |0_{pb}\rangle, \quad (44.52b)$$

$$\langle \bar{1}_{pb}| = \langle \bar{0}_{pb}|\bar{\epsilon}_{pb}, \quad \epsilon_{pb} \cdot |\bar{0}_{pb}\rangle = |\bar{1}_{pb}\rangle, \quad (44.52c)$$

$$\langle \bar{0}_{pb}| = \langle \bar{1}_{pb}|\epsilon_{pb} \cdot, \quad \bar{\epsilon}_{pb}|\bar{1}_{pb}\rangle = |\bar{0}_{pb}\rangle. \quad (44.52d)$$

But in contrast to a real scalar field, a Dirac spinor field satisfies an exclusion principle, enforced by requiring that the product of two destruction operators vanishes, and the product of two creation operators vanishes, in accordance with equations (44.49c) and (44.49d):

$$\epsilon_{pb}\epsilon_{pb} = 0, \quad \bar{\epsilon}_{pb} \cdot \bar{\epsilon}_{pb} \cdot = 0, \quad (44.53a)$$

$$\bar{\epsilon}_{pb}\bar{\epsilon}_{pb} = 0, \quad \epsilon_{pb} \cdot \epsilon_{pb} \cdot = 0. \quad (44.53b)$$

The products of creation and destruction operators that yield a nonvanishing result are

$$\epsilon_{pb} \bar{\epsilon}_{pb} \cdot |0_{pb}\rangle = |0_{pb}\rangle, \quad \bar{\epsilon}_{pb} \cdot \epsilon_{pb}|1_{pb}\rangle = |1_{pb}\rangle, \quad (44.54a)$$

$$\bar{\epsilon}_{pb} \epsilon_{pb} \cdot |\bar{0}_{pb}\rangle = |\bar{0}_{pb}\rangle, \quad \epsilon_{pb} \cdot \bar{\epsilon}_{pb}|\bar{1}_{pb}\rangle = |\bar{1}_{pb}\rangle. \quad (44.54b)$$

The content of equations (44.54), coupled with equations (44.50) and (44.53), is embodied in the assertion that the destruction and creation operators ϵ_{pb} and $\bar{\epsilon}_{pb}$ and their conjugates satisfy anticommutation rules

$$\{\epsilon_{pb}, \bar{\epsilon}_{pb} \cdot\} = \{\bar{\epsilon}_{pb}, \epsilon_{pb} \cdot\} = 1, \quad (44.55a)$$

$$\{\epsilon_{pb}, \epsilon_{pb}\} = \{\bar{\epsilon}_{pb} \cdot, \bar{\epsilon}_{pb} \cdot\} = \{\bar{\epsilon}_{pb}, \bar{\epsilon}_{pb}\} = \{\epsilon_{pb} \cdot, \epsilon_{pb} \cdot\} = 0. \quad (44.55b)$$

For example, the first anticommutator $\{\epsilon_{pb}, \bar{\epsilon}_{pb} \cdot\}$ in equations (44.55a) equals 1 because the only options for the state pb are either vacant $|0_{pb}\rangle$ or occupied $|1_{pb}\rangle$, and the anticommutator transforms either option to itself, equations (44.54a).

When the momenta and spins of the two spinors of a product are permitted to be different, $p'b'$ and pb , the anticommutation rules (44.55) become

$$\{\epsilon_{p'b'}, \bar{\epsilon}_{pb} \cdot\} = \{\bar{\epsilon}_{p'b'}, \epsilon_{pb} \cdot\} = (2\pi)^3 \delta^3(p'_\alpha - p_\alpha) \delta_{b'b}, \quad (44.56a)$$

$$\{\epsilon_{p'b'}, \epsilon_{pb}\} = \{\bar{\epsilon}_{p'b'} \cdot, \bar{\epsilon}_{pb} \cdot\} = \{\bar{\epsilon}_{p'b'}, \bar{\epsilon}_{pb}\} = \{\epsilon_{p'b'} \cdot, \epsilon_{pb} \cdot\} = 0. \quad (44.56b)$$

It is necessary to be careful in making the transition from spinors as classical objects, to spinors as destruction and creation operators. It would be possible to add hats to the spinors to emphasize that they are operators, but that adds an unneeded level of distraction. The key difference is that, as operators, inner products are nonzero only when acting on 1-particle states $|1\rangle$, while outer products are nonzero only when acting on 0-particle states $|0\rangle$. In Feynman diagrams, inner products always occur at vertices (interactions), where they always act on 1-particle states, while outer products always occur at lines (propagators), where they always act on 0-particle states. Thus when dealing with Feynman diagrams, the distinction between classical spinors and spinors as quantum operators effectively disappears: the algebra of the spinor operators is identical to the classical algebra.

44.4.3 Quantized Dirac spinor field

The super-Hamiltonian (44.13) of a Dirac spinor field looks similar to that (42.53) of a scalar field, which in turn looks similar to that (42.9) of a simple harmonic oscillator. One difference between the Dirac super-Hamiltonian (44.13) and the scalar super-Hamiltonian (42.53) lies in where the factors of mass m appear. A consequence is that whereas the scalar field and its conjugate had units $1/\text{length}$ and $1/\text{length}^2$ respectively, the Dirac spinor field and its conjugate both have units $1/\text{length}^{3/2}$.

To achieve *CPT*-invariance, quantum field theory postulates that a field is a linear combination of a forward-in-time destruction operator that destroys a particle, and a backward-in-time creation operator that creates an antiparticle of opposite charge. In accordance with the discussion of spinor creation and destruction operators in §44.4.2, the Dirac field operator $\hat{\psi}$ and its conjugate operator $\hat{\pi}$ should be expanded as

$$\hat{\psi}(x^\mu) \equiv \sum_{b=\uparrow, \downarrow} \hat{\psi}_b(x^\mu), \quad \hat{\pi}(x^\mu) \equiv \sum_{b=\uparrow, \downarrow} \hat{\pi}_b(x^\mu), \quad (44.57)$$

whose spin components $\hat{\psi}_b$ and $\hat{\pi}_b$ are

$$\hat{\psi}_b(x^\mu) \equiv \int \sqrt{\frac{1}{2E_p}} \left(\epsilon_{pb} e^{ip_\mu x^\mu} + \epsilon_{pb} \cdot e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} = \hat{\pi}_b(x^\mu) , \quad (44.58a)$$

$$\hat{\pi}_b(x^\mu) \equiv \int \sqrt{\frac{1}{2E_p}} \left(\bar{\epsilon}_{pb} e^{ip_\mu x^\mu} + \bar{\epsilon}_{pb} \cdot e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} = \hat{\psi}_b(x^\mu) . \quad (44.58b)$$

The expressions (44.58) have the correct form to produce the Feynman-Dirac propagator (44.80) as the expectation value (44.81) of an outer product of the field $\hat{\psi}$ and its conjugate $\hat{\pi} = \hat{\bar{\psi}}$. Equations (44.58) can be compared to the corresponding expressions (42.61) for the real scalar field operator $\hat{\phi}$ and its conjugate momentum operator $\hat{\pi}$.

Hermitian conjugates of column and row basis spinors antispinors are, in accordance with equation (41.102),

$$\epsilon_{pb}^\dagger = i\bar{\epsilon}_{pb} \cdot \gamma_0 , \quad (\epsilon_{pb} \cdot)^\dagger = i\gamma_0 \bar{\epsilon}_{pb} , \quad (44.59a)$$

$$\bar{\epsilon}_{pb}^\dagger = i\epsilon_{pb} \cdot \gamma_0 , \quad (\bar{\epsilon}_{pb} \cdot)^\dagger = i\gamma_0 \epsilon_{pb} . \quad (44.59b)$$

It follows that the Hermitian conjugates $\hat{\psi}_b^\dagger$ and $\hat{\pi}_b^\dagger$ of the field operators $\hat{\psi}_b$ and $\hat{\pi}_b$ given by equations (44.60) are

$$\hat{\psi}_b^\dagger(x^\mu) = i \int \sqrt{\frac{1}{2E_p}} \left(\gamma_0 \bar{\epsilon}_{pb} e^{ip_\mu x^\mu} + \bar{\epsilon}_{pb} \cdot \gamma_0 e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} , \quad (44.60a)$$

$$\hat{\pi}_b^\dagger(x^\mu) = i \int \sqrt{\frac{1}{2E_p}} \left(\gamma_0 \epsilon_{pb} e^{ip_\mu x^\mu} + \epsilon_{pb} \cdot \gamma_0 e^{-ip_\mu x^\mu} \right) \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} . \quad (44.60b)$$

Equal-time anticommutators of the field operators $\hat{\psi}$ and $\hat{\pi}$ with their Hermitian conjugates satisfy

$$\langle 1 | \{ \hat{\psi}_{b'}^\dagger(t, x'^\alpha), \hat{\psi}_b(t, x^\alpha) \} | 1 \rangle = \langle 1 | \{ \hat{\pi}_{b'}^\dagger(t, x'^\alpha), \hat{\pi}_b(t, x^\alpha) \} | 1 \rangle = \delta^3(t, x'^\alpha - x^\alpha) \delta_{b'b} , \quad (44.61)$$

with all other anticommutators vanishing. Enclosing the anticommutators between 1-particle out- and in-states $\langle 1 | \dots | 1 \rangle$ ensures that only inner products, not outer products, contribute to the anticommutators. This is similar to the device of enclosing propagators between 0-particle out- and in-states $\langle 0 | \dots | 0 \rangle$, equation (44.82), to ensure that only outer products, not inner products, contribute to the propagator. An intermediate step to arrive at the anticommutation relation (44.61) is

$$i\bar{\epsilon}_{pb} \cdot \gamma^0 \epsilon_{pb} = i\epsilon_{pb} \cdot \gamma^0 \bar{\epsilon}_{pb} = 2E_p , \quad (44.62)$$

where the value $2E_p$ comes from the fact that, in the rest frame of the spinor or antispinor, the product $\epsilon_{\top b}^\dagger \epsilon_{\top b} = i\bar{\epsilon}_{\top b} \cdot \gamma^0 \epsilon_{\top b} = 1$ or $\bar{\epsilon}_{\top b}^\dagger \bar{\epsilon}_{\top b} = i\epsilon_{\top b} \cdot \gamma^0 \bar{\epsilon}_{\top b} = 1$ is the time component n^0 of the number density n^m , equation (44.20), which becomes $2E_p$ when each spinor is multiplied by $\sqrt{2m}$ and boosted to nonzero spatial momentum in accordance with equations (44.47).

44.4.4 Dirac Hamiltonian operator

The classical Dirac super-Hamiltonian is (44.13). The corresponding Dirac super-Hamiltonian operator, integrated over all space at $t = 0$ like the real Hamiltonian operator (42.73), is

$$\hat{\mathcal{H}} \equiv \frac{1}{2} \sum_b \int m (-\epsilon_{pb} \bar{\epsilon}_{pb} \cdot + \epsilon_{pb} \cdot \bar{\epsilon}_{pb} - \bar{\epsilon}_{pb} \epsilon_{pb} \cdot + \bar{\epsilon}_{pb} \cdot \epsilon_{pb}) \frac{d^3 p}{(2\pi)^3} \Big|_{p^t = E_p}, \quad (44.63)$$

while the conventional Hamiltonian operator is the same with rest mass m replaced by energy E_p ,

$$\hat{\mathcal{H}} \equiv \frac{1}{2} \sum_b \int E_p (-\epsilon_{pb} \bar{\epsilon}_{pb} \cdot + \epsilon_{pb} \cdot \bar{\epsilon}_{pb} - \bar{\epsilon}_{pb} \epsilon_{pb} \cdot + \bar{\epsilon}_{pb} \cdot \epsilon_{pb}) \frac{d^3 p}{(2\pi)^3} \Big|_{p^t = E_p}. \quad (44.64)$$

The Hamiltonian operator $\hat{\mathcal{H}}$ is Hermitian, $\hat{\mathcal{H}}^\dagger = \hat{\mathcal{H}}$, as can be checked using equations (44.59). Whereas the classical super-Hamiltonian (44.7) is an inner product, the operator super-Hamiltonian (44.63) contains both inner and outer product parts, thanks to the transformation from classical spinors to spinors as destruction and creation operators, as discussed in the final paragraph of §44.4.2. The signs in the expressions (44.63) or (44.64) do not follow self-evidently from the classical Hamiltonian (44.7), but are chosen so that an occupied spinor or antispinor state is energy E_p above the vacant state, which requires that the anticommutators of the Hamiltonian operator $\hat{\mathcal{H}}$ with the destruction and creation spinor operators are

$$\{\hat{\mathcal{H}}, \epsilon_{pb}\} = -E_p \epsilon_{pb}, \quad \{\hat{\mathcal{H}}, \epsilon_{pb} \cdot\} = E_p \epsilon_{pb} \cdot, \quad (44.65a)$$

$$\{\hat{\mathcal{H}}, \bar{\epsilon}_{pb}\} = -E_p \bar{\epsilon}_{pb}, \quad \{\hat{\mathcal{H}}, \bar{\epsilon}_{pb} \cdot\} = E_p \bar{\epsilon}_{pb} \cdot. \quad (44.65b)$$

Equations (44.65) differ from the corresponding equations (42.76) for a real scalar field in that anticommutators rather than commutators are involved.

The time derivatives of the fermionic field operators $\hat{\psi}_b(x^\mu)$ and $\hat{\pi}_b(x^\mu)$ defined by equations (44.60) satisfy

$$\partial_t \hat{\psi}_b = i \{\hat{\mathcal{H}}, \hat{\psi}_b\}, \quad (44.66a)$$

$$\partial_t \hat{\pi}_b = i \{\hat{\mathcal{H}}, \hat{\pi}_b\}, \quad (44.66b)$$

which can be compared to the corresponding equations (42.78) for a real scalar field.

The product of two fermionic operators is bosonic, and the time derivative of the products satisfies

$$\partial_t (\hat{\psi} \hat{\pi}) = i [\hat{\mathcal{H}}, \hat{\psi} \hat{\pi}] = i (\{\hat{\mathcal{H}}, \hat{\psi}\} \hat{\pi} - \hat{\psi} \{\hat{\mathcal{H}}, \hat{\pi}\}). \quad (44.67)$$

The Hamiltonian $\hat{\mathcal{H}}$ is itself bosonic, and satisfies

$$\partial_t \hat{\mathcal{H}} = [\hat{\mathcal{H}}, \hat{\mathcal{H}}] = 0, \quad (44.68)$$

consistent with the fact that $\hat{\mathcal{H}}$ is independent of time.

The Hamiltonian operator acting on a 0-particle or 1-particle state gives

$$\hat{\mathcal{H}}|0_{pb}\rangle = -\frac{E_p}{2}|0_{pb}\rangle, \quad \hat{\mathcal{H}}|1_{pb}\rangle = \frac{E_p}{2}|1_{pb}\rangle, \quad (44.69a)$$

$$\hat{\mathcal{H}}|\bar{0}_{pb}\rangle = -\frac{E_p}{2}|\bar{0}_{pb}\rangle, \quad \hat{\mathcal{H}}|\bar{1}_{pb}\rangle = \frac{E_p}{2}|\bar{1}_{pb}\rangle. \quad (44.69b)$$

The conventional Dirac Hamiltonian operator $\hat{\mathcal{H}}$ acting on the vacuum state $|0_b\rangle$ is an integral over all momenta,

$$\hat{\mathcal{H}}|0_b\rangle = - \int \frac{E_p}{2} \frac{d^3p}{(2\pi)^3} \Big|_{p^t=E_p} |0_b\rangle, \quad (44.70)$$

which is nominally infinite like the energy (42.83) of the vacuum of a real scalar field, but with a negative sign. As in the case of the real scalar field, quantum field theory simply discards the infinity, deferring understanding to the future.

44.5 The Dirac propagator

The propagator is the amplitude for a field to propagate from one spacetime point x to another x' . The propagator is defined to be the Green's function solution of the wave equation of the field. For a Dirac field, the propagator $\mathbf{S}(x' - x)$ is the solution of

$$(\mathbf{D}' + m)\mathbf{S}(x' - x) = -i\delta^4(x' - x), \quad (44.71)$$

with appropriate boundary conditions. The Dirac propagator $\mathbf{S}(x' - x)$ is written in boldface because the Dirac propagator is a multivector, and the convention of this book is to write multivectors in boldface. In Minkowski (flat) spacetime, the Dirac propagator equation (44.71) simplifies to

$$(\partial' + m)\mathbf{S}(x' - x) = -i\delta^4(x' - x). \quad (44.72)$$

In terms of the free Klein-Gordon propagator defined by equation (42.84), the free Dirac propagator defined by equation (44.72) is

$$\mathbf{S}(x' - x) = (-\partial' + m)D(x' - x), \quad (44.73)$$

as is proved by applying $\partial' + m$ to both sides. Applying equation (44.73) to the expression (42.89) for the Klein-Gordon propagator as a Bessel function yields

$$\mathbf{S}(x' - x) = \left(-\frac{\mathbf{s}}{s} \frac{\partial}{\partial s} + m\right)D(s) = \frac{m^2}{4\pi^2 s} \left(\frac{\mathbf{s}}{s} K_2(ms) + K_1(ms)\right), \quad (44.74)$$

where s , equation (42.86), is the Lorentz-invariant separation from x to x' , and

$$\mathbf{s} \equiv \boldsymbol{\gamma}_n(x'^n - x^n). \quad (44.75)$$

The Dirac propagator (44.74) is a function of $\mathbf{s}$ as well as s , but is still Lorentz invariant. When the frame

is Lorentz transformed, both γ_n and $x'^n - x^n$ Lorentz transform, but they Lorentz transform in opposite directions, leaving $\mathbf{s}$ as a Lorentz scalar. However, if the sign of $x'^n - x^n$ is flipped (without also flipping the sign of γ_n) then the sign of $\mathbf{s}$ also flips. Thus, unlike the Klein-Gordon propagator, which remains unchanged when $x' - x$ is flipped, the $\mathbf{s}$ part of the Dirac propagator changes sign when $x' - x$ is flipped to $x - x'$,

$$\mathbf{S}(x - x') = \frac{m^2}{4\pi^2 s} \left(-\frac{\mathbf{s}}{s} K_2(ms) + K_1(ms) \right). \quad (44.76)$$

As with the Klein-Gordon propagator, the expression (44.74) for the Dirac propagator holds for both spacelike, $s^2 > 0$, and timelike, $s^2 < 0$, separations. For spacelike separations, s is real, and the Dirac propagator (44.74) is a real multivector. For spacelike separations, a separation $x' - x$ with $t' - t > 0$ can be rotated by a continuous Lorentz transformation into a separation $x' - x$ with $t' - t < 0$. Note that although a spacelike $x' - x$ can be rotated into $x - x'$ by a continuous Lorentz transformation, this is not the same as flipping $x' - x$ (without also flipping γ^n).

At timelike separations $s = i\tau$ where τ is the proper time from x to x' , the Klein-Gordon propagator $D(s)$ is a function of $i|\tau|$, equation (42.90), and the expression (44.74) for the Dirac propagator becomes

$$\mathbf{S}(x' - x) = \frac{m^2}{4\pi^2(i|\tau|)} \left(\frac{\mathbf{s}}{i|\tau|} K_2(im|\tau|) + K_1(im|\tau|) \right). \quad (44.77)$$

If the sign of $x' - x$ is flipped, $i|\tau|$ remains unchanged, but again the sign of $\mathbf{s}$ flips.

44.5.1 Free Dirac propagator by contour integration

The Green's function equation (44.72) for the Dirac propagator in Minkowski (flat) spacetime solves to

$$\mathbf{S}(x' - x) = -i \int \frac{-i\mathbf{p} + m}{p^2 + m^2} e^{ip \cdot (x' - x)} \frac{d^4 p}{(2\pi)^4} \quad (44.78)$$

subject to appropriate boundary conditions, where $\mathbf{p} \equiv \gamma_n p^n$. The energy part p^0 of the integral (44.78) is over all energies from $-\infty$ to $+\infty$. Similarly to the Klein-Gordon propagator, §42.4.1, the integrand of equation (44.78) has poles at $p^0 = \pm E_p$, and different boundary conditions, retarded, advanced, or Feynman, are obtained by perturbing the contour of integration either above or below each of the two poles. Equation (44.78) may be checked using equation (44.73), by applying the operator $-\partial' + m$ to the Feynman-Klein-Gordon propagator (42.85).

The correct boundary conditions for quantum field theory are Feynman boundary conditions, in which the contour goes below the negative pole $p^0 = -E_p$ and above the positive pole $p^0 = E_p$. Feynman's prescription enforces *CPT* symmetry by propagating positive energy waves into the future, and negative energy waves into the past. As with the Klein-Gordon propagator (42.101), a convenient way to implement Feynman's prescription is to perturb the poles of p^0 to $\pm(E_p - i\epsilon)$ where ϵ is a small positive quantity, yielding the Feynman-Dirac propagator

$$\mathbf{S}(x' - x) = -i \int \frac{-i\mathbf{p} + m}{p^2 + m^2 - i\epsilon} e^{ip \cdot (x' - x)} \frac{d^4 p}{(2\pi)^4}. \quad (44.79)$$

Flipping the sign of $x' - x$ in equation (44.79) is equivalent to flipping the sign of p , which leaves p^2 and d^4p unchanged, but flips the sign of $\mathbf{p}$. Similarly to the Feynman-Klein-Gordon propagator (42.102), the Feynman-Dirac propagator (44.79) integrates to

$$\mathcal{S}(x' - x) = \int (\mp i\mathbf{p} + m) e^{\pm i\mathbf{p} \cdot (x' - x)} \Big|_{p^0 = E_p} \frac{d^3p}{2E_p(2\pi)^3}, \quad (44.80)$$

where the upper and lower signs hold for respectively $t' - t > 0$ and $t' - t < 0$. Equation (44.80) may be checked using equation (44.73), by applying the operator $-\partial' + m$ to the Feynman-Klein-Gordon propagator (42.102). Equation (44.80) along with equations (44.48) recovers the earlier claimed expression (44.81) for the Dirac propagator.

It is worth emphasizing again how the paradoxical behaviour of quantum mechanics, that it achieves instantaneous “faster-than-light” propagation without violating causality, is built in to the propagators of quantum field theory. The propagators of quantum field theory are nonvanishing not only at future timelike separations, but also at spacelike and past timelike separations. The propagator for a particle moving backwards in time (along either a timelike or spacelike direction) is indistinguishable from the propagator for an antiparticle moving forwards in time. Thus the future influencing the past is indistinguishable from the past influencing the future.

44.5.2 The Dirac propagator as a two-point function

Equations (44.48) inserted into equation (44.80) yield the following expression for the Feynman-Dirac propagator,

$$\mathcal{S}(x' - x) = \begin{cases} \int \sum_b (\epsilon_{pb} \bar{\epsilon}_{pb} \cdot) e^{i(x' - x) \cdot p} \Big|_{p^0 = E_p} \frac{d^3p}{2E_p(2\pi)^3} & (t' - t \geq 0) \\ - \int \sum_b (\bar{\epsilon}_{pb} \epsilon_{pb} \cdot) e^{-i(x' - x) \cdot p} \Big|_{p^0 = E_p} \frac{d^3p}{2E_p(2\pi)^3} & (t' - t < 0) \end{cases}, \quad (44.81)$$

which can be recast as a product of the field operator $\hat{\psi}(x')$ and its conjugate $\hat{\pi}(x)$, equations (44.57),

$$\mathcal{S}(x' - x) = \begin{cases} \langle 0 | \hat{\psi}(x') \hat{\pi}(x) | 0 \rangle & (t' - t \geq 0) \\ - \langle 0 | \hat{\pi}(x) \hat{\psi}(x') | 0 \rangle & (t' - t < 0) \end{cases}. \quad (44.82)$$

The sandwiching between 0-particle out- and in-states $\langle 0 |$ and $| 0 \rangle$ ensures that only outer products of spinors survive in the product of field operators. Notice the minus sign in going between the $t' - t \geq 0$ and $t' - t < 0$ expressions; this is a general feature of products of fermion fields, that they anticommute, therefore pick up a minus sign whenever their order is swapped. Equation (44.82) can be abbreviated

$$\mathcal{S}(x' - x) = \langle 0 | T \hat{\psi}(x') \hat{\pi}(x) | 0 \rangle, \quad (44.83)$$

where the time-ordering symbol T instructs to time-order the operators in the usual right-to-left physics convention, with the additional instruction to insert a minus sign whenever the order of two fermionic

field is swapped. The Feynman-Dirac propagator (44.83) can be compared to the Feynman-Klein-Gordon propagator (42.105) of a real scalar field.

The way that the Dirac propagator (44.81) encodes antispinors shows how antispinors must be encoded in external legs of a Feynman diagram. Equations (44.48) show that replacing a spinor by an antispinor is accomplished by replacing a spinor ϵ_{pb} by i times its conjugate,

$$\epsilon_{pb} \rightarrow i\bar{\epsilon}_{pb} . \quad (44.84)$$

The factor of the imaginary i derives from the minus sign in equation (44.48b), and cannot be eliminated by adjusting the phases of basis spinors. The rule (44.84) is used in §47.8 to derive amplitudes and rates for the creation and annihilation of spinor-antispinor pairs.

Quantum field theory of the electromagnetic field

The electromagnetic action has been discussed extensively in this book, §§16.5 and 16.6. The coordinates of the electromagnetic field are the components A_ν of the electromagnetic potential, and its conjugate momenta are, in Heaviside units, $\Pi^{\mu\nu} = -F^{\mu\nu}$, equation (16.29), where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, equation (16.23). In a general curved spacetime, the electromagnetic potential A_ν satisfies a wave equation (16.45). The units of the coordinates A_ν are 1/length, while the units of the conjugate momenta $\Pi^{\mu\nu}$ are 1/length², the same as the units of the coordinate φ of a real scalar field and its conjugate momentum π , Concept question 42.2.

As expounded in §27.6, although the electric potential A_ν has 4 components, only the two transverse components in a locally inertial frame aligned with the field are gauge-invariant propagating degrees of freedom, and therefore comprise physical degrees of freedom in quantum field theory. The longitudinal (scalar) component of the electric potential can be removed by a gauge transformation (27.36), and the evolution of the time component of the electric potential is governed by a constraint equation (27.43) whose continued satisfaction is guaranteed by conservation of electric charge. In flat (Minkowski) spacetime, the transverse components A_+ and A_- satisfy a wave equation (27.40), which in Heaviside units is

$$\square A_a = -j_a , \quad (45.1)$$

with index a running over the two circular polarizations $+$ and $-$, or over linear polarizations 1 and 2.

As described in §35.1, photons are described by a gauge-invariant transverse polarization vector, with two circular polarization components γ_+ and γ_- in a frame where the photon direction is along the 3-direction (z -direction), equation (35.1). The two components $\gamma_\pm$ describe photons with spin 1 in a direction respectively aligned ($+$, right-handed) and anti-aligned ($-$, left-handed) with the direction of motion. The two components are conjugates of each other, $\bar{\gamma}_+ = \gamma_-$, equation (41.115). The condition that the spin of the photon is aligned or anti-aligned with its direction of motion is Lorentz invariant, unchanged by any Lorentz transformation. Let γ_{ka} , with spin index a running over $+$ and $-$, denote the photon rotated by some spatial rotor R into a frame where its energy-momentum is in the direction $\hat{k}$, part 3 of Exercise 14.10,

$$\gamma_{ka} = R\gamma_a\tilde{R} . \quad (45.2)$$

Basis vectors γ_a are dimensionless, and so are their Lorentz-transformed versions γ_{ka} . It is possible to allow

R to be a more general Lorentz rotor involving a boost as well as a spatial rotation, but $\gamma_{\pm}$ are unchanged by a Lorentz boost along the direction of motion, so a pure spatial rotation R suffices.

45.1 Quantized electromagnetic field

Just as Dirac basis spinors ϵ_{pb} can be promoted from classical objects to destruction and creation operators, §44.4.2, so also transverse basis vectors γ_{ka} can be promoted from classical objects to destruction and creation operators for spin 1 photons. Classically, the Hermitian conjugate of a basis vector coincides with its conjugate, which in turn coincides with the basis vector with raised index, $\gamma_{ka}^{\dagger} = \bar{\gamma}_{ka} = \gamma_k^a$. In quantum field theory, γ_{ka} becomes a destruction operator, while its conjugate $\gamma_{ka}^{\dagger}$ becomes a creation operator, satisfying

$$\langle 0 | \gamma_{ka}^{\dagger} = 0, \quad \gamma_{ka} | 0 \rangle = 0. \quad (45.3)$$

In accordance with the Feynman rules, γ_{ka} destroys a photon, while $\gamma_{ka}^{\dagger}$ creates an antiphoton, a photon of opposite spin. Equations (45.3) can be compared to the case (42.67) of a real scalar field, or (44.50) of a Dirac spinor field. The state consisting of zero photons in the momentum and spin state ka , without regard to the possible presence of photons in other momentum states, is denoted $|0_{ka}\rangle$. Similarly to the case of a real scalar field, equations (42.70), out- and in-states $\langle n_{ka}|$ and $|n_{ka}\rangle$ of an electromagnetic field containing n photons in momentum and spin state ka , correctly normalized to $\langle n_{ka} | n_{ka} \rangle = 1$, are

$$\langle n_{ka} | = \frac{1}{\sqrt{n!}} \langle 0_{ka} | (\gamma_{ka})^n, \quad \frac{1}{\sqrt{n!}} (\gamma_{ka}^{\dagger})^n | 0_{ka} \rangle = |n_{ka}\rangle, \quad (45.4a)$$

$$\langle 0_{ka} | = \frac{1}{\sqrt{n!}} \langle n_{ka} | (\gamma_{ka}^{\dagger})^n, \quad \frac{1}{\sqrt{n!}} (\gamma_{ka})^n | n_{ka} \rangle = |0_{ka}\rangle. \quad (45.4b)$$

It is not necessary to introduce a separate set of operators for antiphotons, because the antiphoton of a photon is the photon of opposite spin, and the spin index a in equations (45.4) already runs over both spin states $+$ and $-$ of the photon. The commutators of the destruction and creation operators are, similarly to those (42.63) of a scalar field,

$$[\gamma_{k'a'}, \gamma_{ka}^{\dagger}] = (2\pi)^3 \delta^3(k'_{\alpha} - k_{\alpha}) \delta_{a'a}, \quad (45.5a)$$

$$[\gamma_{k'a'}, \gamma_{ka}] = [\gamma_{k'a'}^{\dagger}, \gamma_{ka}^{\dagger}] = 0. \quad (45.5b)$$

At this point one might be inclined to define the electromagnetic field operator $\hat{A}(x^{\mu})$ as a sum (45.7) over spin states $a = +, -$, analogously to the sum (44.57) over spin states of the Dirac spinor field. However, the electromagnetic field operator (45.8a) is an integral over all momentum directions, and the transverse plane varies with the direction of momentum. The difficulty can be fixed by extending the sum (45.7) over a to all 3 spatial directions $+, -, 3$. The third spatial component A_3 can be admitted by relaxing the gauge condition that it vanishes to the Lorenz gauge condition (27.45). Under the Lorenz gauge condition, all 4 components A_m of the electromagnetic potential satisfy the 4-dimensional wave equation (27.48), which in

Heaviside units is

$$\square A_m = -j_m . \quad (45.6)$$

The momentum Π^{ta} conjugate to the coordinate A_a equals minus the time component of the contravariant electromagnetic field, which equals minus the usual electric field, $\Pi^{ta} = -F^{ta} = F_{ta} = -E_a$. An electromagnetic wave travelling freely in Minkowski space sees no external charge, in which case the time component of the electric potential can be gauge-transformed to zero, equation (27.48a), and the momentum field reduces to the time derivative of the coordinate, $\Pi^{ta} = \partial_t A_a$. Thus the electromagnetic field operator $\hat{\mathbf{A}}(x^\mu)$ and its conjugate momentum $\hat{\mathbf{\Pi}}(x^\mu)$ are defined by

$$\hat{\mathbf{A}}(x^\mu) \equiv \sum_{a=+,-,3} \hat{\mathbf{A}}_a(x^\mu) , \quad \hat{\mathbf{\Pi}}(x^\mu) \equiv \sum_{a=+,-,3} \hat{\mathbf{\Pi}}_a(x^\mu) , \quad (45.7)$$

with components

$$\hat{\mathbf{A}}_a(x^\mu) \equiv \int \sqrt{\frac{1}{2\omega_k}} \left(\gamma_{ka} e^{ik_\mu x^\mu} + \gamma_{ka}^\dagger e^{-ik_\mu x^\mu} \right) \frac{d^3k}{(2\pi)^3} \Big|_{k^t=\omega_k} , \quad (45.8a)$$

$$\hat{\mathbf{\Pi}}_a(x^\mu) \equiv -i \int \sqrt{\frac{\omega_k}{2}} \left(\gamma_{ka} e^{ik_\mu x^\mu} - \gamma_{ka}^\dagger e^{-ik_\mu x^\mu} \right) \frac{d^3k}{(2\pi)^3} \Big|_{k^t=\omega_k} = \partial_t \hat{\mathbf{A}}_a(x^\mu) , \quad (45.8b)$$

where, since the photon is massless, the frequency satisfies $\omega_k = \sqrt{k_\mu k^\mu}$. Equations (45.8) can be compared to the corresponding expressions (42.61) and (44.58) for the real scalar and Dirac spinor field operators and their conjugate momenta.

Equal-time commutators of the electromagnetic field $\hat{\mathbf{A}}$ and its conjugate momentum $\hat{\mathbf{\Pi}}$ satisfy

$$[\hat{\mathbf{A}}_{a'}(t, x'^\alpha), \hat{\mathbf{\Pi}}_a(t, x^\alpha)] = i \delta^3(t, x'^\alpha - x^\alpha) \delta_{a'a} , \quad (45.9a)$$

$$[\hat{\mathbf{A}}_{a'}(t, x'^\alpha), \hat{\mathbf{A}}_a(t, x^\alpha)] = [\hat{\mathbf{\Pi}}_{a'}(t, x'^\alpha), \hat{\mathbf{\Pi}}_a(t, x^\alpha)] = 0 , \quad (45.9b)$$

which can be compared to the commutators (42.60) of a real scalar field and its conjugate.

45.1.1 Electromagnetic Hamiltonian operator

Analogous to the conventional Hamiltonian operator (42.75) of a real scalar field or (44.64) of a Dirac spinor field, the conventional Hamiltonian operator of the electromagnetic field, integrated over all space at $t = 0$, is

$$\hat{\mathcal{H}} = \sum_{a=+,-,3} \int \omega_k^2 (\gamma_{ka} \gamma_{ka}^\dagger + \gamma_{ka}^\dagger \gamma_{ka}) \frac{d^3k}{2\omega_k (2\pi)^3} \Big|_{k^t=\omega_k} . \quad (45.10)$$

Results analogous to those obtained in §42.3.6 for the conventional Hamiltonian operator of a real scalar field, equations (42.76)–(42.83), hold also for the electromagnetic field, including the fact that the Hamiltonian generates the time evolution of field operators, equation (42.79), and energy levels of a state ka are spaced like those of a simple harmonic oscillator, equally spaced with the ground state half a unit above zero.

45.2 Electromagnetic propagator

The electromagnetic propagator is the Green's function solution $D_{m'm}(x' - x)$ to the wave equation (45.6),

$$\square' D_{m'm}(x' - x) = i \delta^4(x' - x) \eta_{m'm} , \quad (45.11)$$

with appropriate boundary conditions. The reason for the Minkowski metric $\eta_{m'm}$ on the right hand side is that each individual component A_m of the electromagnetic potential satisfies the same wave equation (45.6), and the propagator must be a Lorentz-covariant tensor.

Actually, the gauge freedom in the electromagnetic potential, which has the consequence that only its transverse components have physical significance, allows a more general choice of Lorentz-covariant tensor than $\eta_{m'm}$, namely $\eta_{m'm}$ with an arbitrary longitudinal tensor subtracted,

$$\eta_{m'm} \rightarrow \eta_{m'm} - \xi(k^2) \frac{k_{m'} k_m}{k^2} , \quad (45.12)$$

where $\xi(k^2)$ is some arbitrary function of k^2 . The choice $\xi = 1$ is called Landau gauge. Amplitudes of electromagnetic processes prove to be independent of the choice of ξ . The nonphysicality of longitudinal modes has the consequence that amplitudes satisfy a transversality condition called the Ward-Takahashi identity (47.148). There is an elaborate but powerful and mathematically deep procedure, the BRST formalism (Becchi, Rouet, and Stora, 1976; Tyutin, 1975), that yields gauge-invariant amplitudes at each loop order of perturbation theory, in any gauge theory, without choosing a particular gauge, by adding to the Lagrangian “ghost” fields that eliminate the unphysical gauge modes. The BRST formalism becomes essential to deal with nonabelian gauge groups such as $SU(n)$, but the simpler approach of choosing a particular gauge proves adequate for quantum electrodynamics.

With the gauge choice $\xi = 0$, the resulting electromagnetic propagator is similar to the Feynman-Klein-Gordon propagator (42.102),

$$\begin{aligned} D_{m'm}(x' - x) &\equiv \eta_{m'm} D(x' - x) \\ &= \eta_{m'm} \int e^{\pm i(x' - x) \cdot k} \Big|_{k^0 = \omega_k} \frac{d^3 k}{2\omega_k (2\pi)^3} \\ &= \eta_{m'm} \int \frac{1}{k^2 - i\epsilon} e^{i(x' - x) \cdot k} \frac{d^4 k}{(2\pi)^4} . \end{aligned} \quad (45.13)$$

The second and third lines of equations (45.13) are from the expressions (42.102) and (42.101) for the Klein-Gordon propagator. The $\pm$ sign in the exponent on the second line is positive for $t' - t > 0$, negative for $t' - t < 0$. Like the Klein-Gordon propagator $D(x' - x)$, equation (42.91), the photon propagator $D_{m'm}(x' - x)$ is invariant under a flip of the direction $x' - x$,

$$D_{m'm}(x - x') = D_{m'm}(x' - x) . \quad (45.14)$$

The invariance of the photon propagator under $x' \leftrightarrow x$ means that a photon is its own antiparticle.

The spatial components $D_{a'a}(x' - x)$ of the electromagnetic propagator (45.13) can be expressed as an

expectation value of time-ordered products of the electromagnetic field operator $\hat{\mathbf{A}}_a(x)$,

$$D_{a'a}(x' - x) = \langle 0 | T \hat{\mathbf{A}}_{a'}(x') \hat{\mathbf{A}}_a(x) | 0 \rangle , \quad (45.15)$$

which can be compared to the corresponding expressions (42.105) and (44.83) for the Feynman-Klein-Gordon and Feynman-Dirac propagators.

Quantum field theory of the gravitational field

As discussed at the beginning of Chapter 42, there is no consensus theory of quantum gravity. The problem is not that the gravitational field cannot be quantized, but rather that the resulting theory is not renormalizable, because the gravitational coupling parameter $8\pi G$ has positive length dimension length^2 . This chapter walks through the quantization of the gravitational field, showing that the quantum of the gravitational field is expected to be a spin 2 graviton, without worrying about renormalizability.

The gravitational action has been discussed extensively in this book, §§16.7, 16.13, and 16.14. The coordinates of the gravitational field are the components of the e^λ of the vierbein, a vector 1-form, and its conjugate momenta are $\pi^\kappa{}_\lambda/(8\pi G)$, a vector 2-form equations (16.115). The momentum can also be treated as a bivector 1-form, equations (16.185) or (16.259).

The vierbein coefficients $e^{n\lambda}$ are dimensionless, while the conjugate momentum coefficients $\pi_n{}^\kappa{}_\lambda$ defined by equation (16.113) have units 1/length, equation (16.114). The coordinates and momenta of a bosonic field should have units 1/length and 1/length² respectively, like those of a real scalar field (Concept question 42.2) or of the electromagnetic field. To ensure that the coordinates and momenta of the gravitational field have the right units for quantization, it is necessary to scale them by $1/\sqrt{8\pi G}$, which has dimension 1/length as required,

$$\frac{e^\lambda}{\sqrt{8\pi G}}, \quad \frac{\pi^\kappa{}_\lambda}{\sqrt{8\pi G}}. \quad (46.1)$$

The momentum conjugate to the coordinate $e^\lambda/\sqrt{8\pi G}$ is indeed $\pi^\kappa{}_\lambda/\sqrt{8\pi G}$. The gravitational field operator $\hat{h}$ and its conjugate momentum operator $\hat{\pi}$ defined by equations (46.12) below should be understood as the quantum operator equivalents of the scaled gravitational coordinates and momenta (46.1), including the scaling by $1/\sqrt{8\pi G}$. The correctness of the normalization (46.1) can be verified by expanding the Hilbert Lagrangian (16.87) to lowest (2nd) order in tensor perturbations h_{ab} , Exercise 46.1.

In general relativistic perturbation theory in Minkowski space, there are 6 coordinate- and gauge-invariant perturbations (27.13) of the vierbein, but only the 2 tensor components h_{ab} propagate as waves (27.27), whereas the 2 scalar and 2 vector components Ψ , Φ , and W_a satisfy instantaneous spatial Einstein equations (27.28) and are not propagating. Only the two propagating tensor components h_{ab} comprise physical degrees of freedom in quantum field theory. The physical degrees of freedom can be taken to be the spin +2

and -2 components h_{++} and h_{--} , equations (27.23). In the absence of a tensor source of energy-momentum, the transverse components h_{++} and h_{--} satisfy the massless wave equation (27.27),

$$\square h_{ab} = 0, \quad (46.2)$$

with index ab running over the two circular polarizations $++$ and $--$. The two polarizations could alternatively be chosen to be the linear polarizations $+$ and $\times$, equations (27.19).

Recall that, as described in §35.1, photons are described by a gauge-invariant transverse polarization vector, with two circular polarization components γ_+ and γ_- in a frame where the photon direction is along the 3-direction (z -direction), equation (35.1). Similarly, gravitons can be described by a gauge-invariant polarization vector γ_{ab} defined to be a symmetric tensor product of two vectors γ_a and γ_b (*not* a multivector product $\gamma_a \gamma_b$),

$$\gamma_{ab} \equiv \gamma_a \otimes \gamma_b. \quad (46.3)$$

The gauge-invariant transverse polarizations are γ_{++} and γ_{--} , which describe gravitons with spin 2 in a direction respectively aligned ($++$, right-handed) and anti-aligned ($--$, left-handed) with the direction of motion. The conjugate of γ_{ab} is $\bar{\gamma}_{ab} = \bar{\gamma}_a \otimes \bar{\gamma}_b$, and in particular the two chiral components are conjugates of each other, $\bar{\gamma}_{++} = \gamma_{--}$. The condition that the spin of the graviton is aligned or anti-aligned with its direction of motion is Lorentz invariant, unchanged by any Lorentz transformation. Let γ_{kab} , with spin index ab running over $++$ and $--$, denote the graviton rotated by some spatial rotor R into a frame where its energy-momentum is in the direction $\hat{k}$, part 3 of Exercise 14.10,

$$\gamma_{kab} = (R\gamma_a \tilde{R}) \otimes (R\gamma_b \tilde{R}). \quad (46.4)$$

Basis vectors γ_{ab} are dimensionless, and so are their Lorentz-transformed versions γ_{kab} . It is possible to allow R to be a more general Lorentz rotor involving a boost as well as a spatial rotation, but $\gamma_{\pm\pm}$ are unchanged by a Lorentz boost along the direction of motion, so a pure spatial rotation R suffices.

Exercise 46.1 Hilbert gravitational Lagrangian in perturbation theory. Show that the lowest (2nd) order contribution R_h to the Ricci scalar R from tensor perturbations h_{ab} defined by equation (27.6d) in Minkowski space is

$$R_h = -\eta^{\mu\nu} \partial_\mu h_{ab} \partial_\nu h^{ab} = \dot{h}_{ab} \dot{h}^{ab} - \nabla_c h_{ab} \nabla^c h^{ab}, \quad (46.5)$$

with $h_{ab} h^{ab}$ a sum over the two polarizations h_{++} and h_{--} ,

$$h_{ab} h^{ab} = 2h_{++} h_{--} = 2(h_{xx}^2 + h_{xy}^2). \quad (46.6)$$

Hence conclude that the lowest order contribution L_h to the Hilbert Lagrangian (16.87) from tensor perturbations is

$$L_h \equiv \frac{1}{16\pi G} R_h = -\frac{1}{16\pi G} \eta^{\mu\nu} \partial_\mu h_{ab} \partial_\nu h^{ab}, \quad (46.7)$$

which looks like the Lagrangian (42.47) of a pair (one for each polarization) of real scalar fields $h_{ab}/\sqrt{8\pi G}$.

46.1 Quantized gravitational field

Just as Dirac basis spinors ϵ_{pb} and transverse electromagnetic basis vectors γ_{ka} can be promoted from classical objects to destruction and creation operators, §44.4.2 and §45.1, so also transverse gravitational basis vectors γ_{kab} can be promoted from classical objects to destruction and creation operators for spin 2 gravitons. Classically, the Hermitian conjugate of γ_{kab} is its conjugate, $\gamma_{kab}^\dagger = \bar{\gamma}_{kab}$. In quantum field theory, γ_{kab} becomes a destruction operator, while its conjugate $\gamma_{kab}^\dagger$ becomes a creation operator, satisfying

$$\langle 0 | \gamma_{kab}^\dagger = 0, \quad \gamma_{kab} | 0 \rangle = 0. \quad (46.8)$$

Equations (46.8)–(46.10) for gravitons essentially duplicate equations (45.3)–(45.5) for photons, with the spin 1 index a replaced by the symmetric spin 2 index ab . In accordance with the Feynman rules, γ_{kab} destroys a graviton, while $\gamma_{kab}^\dagger$ creates an antigraviton, a graviton of opposite spin. Equations (46.8) can be compared to the case (42.67) of a real scalar field, or (44.50) of a Dirac spinor field, or (45.3) of the electromagnetic field. The state consisting of zero gravitons in the momentum and spin state kab , without regard to the possible presence of gravitons in other momentum states, is denoted $|0_{kab}\rangle$. Similarly to the case of a real scalar or electromagnetic field, equations (42.70) or (45.4), out- and in-states $\langle n_{kab}|$ and $|n_{kab}\rangle$ of a gravitational field containing n gravitons in momentum and spin state kab , correctly normalized to $\langle n_{kab} | n_{kab} \rangle = 1$, are

$$\langle n_{kab} | = \frac{1}{\sqrt{n!}} \langle 0_{kab} | (\gamma_{kab})^n, \quad \frac{1}{\sqrt{n!}} (\gamma_{kab}^\dagger)^n | 0_{kab} \rangle = |n_{kab}\rangle, \quad (46.9a)$$

$$\langle 0_{kab} | = \frac{1}{\sqrt{n!}} \langle n_{kab} | (\gamma_{kab}^\dagger)^n, \quad \frac{1}{\sqrt{n!}} (\gamma_{kab})^n | n_{kab} \rangle = |0_{kab}\rangle. \quad (46.9b)$$

Similarly to photons, it is not necessary to introduce a separate set of operators for antigravitons, because the antigraviton of a graviton is the graviton of opposite spin, and the spin index ab in equations (46.9) already runs over both spin states $++$ and $--$ of the graviton. The commutators of the destruction and creation operators are, similarly to those (42.63) of a scalar field or (45.5) of the electromagnetic field,

$$[\gamma_{k'a'b'}, \gamma_{kab}^\dagger] = (2\pi)^3 \delta^3(k'_\alpha - k_\alpha) \delta_{a'a} \delta_{b'b}, \quad (46.10a)$$

$$[\gamma_{k'a'b'}, \gamma_{kab}] = [\gamma_{k'a'b'}^\dagger, \gamma_{kab}^\dagger] = 0. \quad (46.10b)$$

One might be inclined to define the gravitational field operator $\hat{h}(x^\mu)$ as a sum (46.12) over spin states $ab = ++, --$, analogously to the sum (44.57) over spin states of the Dirac spinor field. However, the gravitational field operator (46.12) is an integral over all momentum directions, and the transverse plane varies with the direction of momentum. The difficulty can be fixed by extending the sum (46.12) over ab to all 6 symmetric spatial tensors. The 10 tetrad-gauge invariant (i.e. symmetric) vierbein perturbations $\varphi_{(mn)} \equiv \frac{1}{2}(\varphi_{mn} + \varphi_{nm})$ can be reduced to 6 by imposing the 4 Lorentz-invariant harmonic gauge conditions (27.51). Under the harmonic gauge condition, all 10 components $\varphi_{(mn)}$ of the symmetric perturbation satisfy the 4-dimensional wave equation (27.56), which in the absence of a source R_{mn} of energy-momentum is

$$\square \varphi_{(mn)} = 0. \quad (46.11)$$

In a background Minkowski spacetime, components of perturbed quantities can be raised and lowered with the Minkowski metric. The salient components of the momentum conjugate to a spatial perturbation h_{ab} are, from equation (16.113), $\pi^{a0b} = -\pi_{a0b} = \pi_{0ab}$, which is related to the tetrad-frame connection Γ_{0ab} by equation (16.114), $\pi_{0ab} = \Gamma_{0ab} + \delta_{ab}\Gamma_{0c}^c$. For the transverse traceless tensor components h_{ab} , the conjugate momentum reduces to $\pi_{0ab} = \Gamma_{0ab}$. But according to equation (27.15b), the time derivative of the transverse traceless tensor perturbation is $\dot{h}_{ab} = \Gamma_{0ab}$, which equals π_{0ab} . To summarize, in perturbed Minkowski spacetime, the time derivative of the transverse traceless tensor perturbation equals its conjugate momentum, the same as for a real scalar or electromagnetic field. This motivates *defining* the general components $\hat{\pi}_{ab}$ of the momentum operator conjugate to the gravitational field operator $\hat{h}_{ab}$ to be equal to the time derivative of $\hat{h}_{ab}$.

Define therefore the gravitational field operator $\hat{h}(x^\mu)$ and its conjugate momentum $\hat{\pi}(x^\mu)$ by

$$\hat{h}(x^\mu) \equiv \sum_{ab} \hat{h}_{ab}(x^\mu) , \quad \hat{\pi}(x^\mu) \equiv \sum_{ab} \hat{\pi}_{ab}(x^\mu) , \quad (46.12)$$

in which the sums are over the 6 symmetric pairs of spatial indices ab . The components of the operators are

$$\hat{h}_{ab}(x^\mu) \equiv \int \sqrt{\frac{1}{2\omega_k}} \left(\gamma_{kab} e^{ik_\mu x^\mu} + \gamma_{kab}^\dagger e^{-ik_\mu x^\mu} \right) \frac{d^3k}{(2\pi)^3} \Big|_{k^t=\omega_k} , \quad (46.13a)$$

$$\hat{\pi}_{ab}(x^\mu) \equiv -i \int \sqrt{\frac{\omega_k}{2}} \left(\gamma_{kab} e^{ik_\mu x^\mu} - \gamma_{kab}^\dagger e^{-ik_\mu x^\mu} \right) \frac{d^3k}{(2\pi)^3} \Big|_{k^t=\omega_k} = \partial_t \hat{h}_{ab}(x^\mu) . \quad (46.13b)$$

Equations (46.13) can be compared to the corresponding expressions (42.61), (44.58), and (45.8) for the real scalar, Dirac spinor, and electromagnetic field operators, and their conjugate momenta.

Equal-time commutators of the electromagnetic field $\hat{h}$ and its conjugate momentum $\hat{\pi}$ satisfy

$$[\hat{h}_{a'b'}(t, x'^\alpha), \hat{\pi}_{ab}(t, x^\alpha)] = i \delta^3(t, x'^\alpha - x^\alpha) \delta_{a'a} \delta_{b'b} , \quad (46.14a)$$

$$[\hat{h}_{a'b'}(t, x'^\alpha), \hat{h}_{ab}(t, x^\alpha)] = [\hat{\pi}_{a'b'}(t, x'^\alpha), \hat{\pi}_{ab}(t, x^\alpha)] = 0 , \quad (46.14b)$$

which can be compared to the commutators (42.60) or (45.9) of a real scalar or electromagnetic field and their conjugates.

46.1.1 Gravitational Hamiltonian operator

Analogous to the conventional Hamiltonian operator (42.75) of a real scalar field, (44.64) of a Dirac spinor field, or (45.10) of the electromagnetic field, the conventional Hamiltonian operator of the gravitational field, integrated over all space at $t = 0$, is

$$\hat{\mathcal{H}} = \sum_{ab} \int \omega_k^2 (\gamma_{kab} \gamma_{kab}^\dagger + \gamma_{kab}^\dagger \gamma_{kab}) \frac{d^3k}{2\omega_k (2\pi)^3} \Big|_{k^t=\omega_k} . \quad (46.15)$$

Results analogous to those obtained in §42.3.6 for the conventional Hamiltonian operator of a real scalar field, equations (42.76)–(42.83), hold also for the gravitational field, including the fact that the Hamiltonian

generates the time evolution of field operators, equation (42.79), and energy levels of a state kab are spaced like those of a simple harmonic oscillator, equally spaced with the ground state half a unit above zero.

46.2 Gravitational propagator

The gravitational propagator is the Green's function solution $D_{m'n'mn}(x' - x)$ to the wave equation (46.11),

$$\square' D_{m'n'mn}(x' - x) = i \delta^4(x' - x) P_{m'n'mn} , \quad (46.16)$$

with appropriate boundary conditions. The $P_{m'n'mn}$ in equation (46.16) is a projection tensor, the most commonly used version of which is

$$P_{m'n'mn} = \frac{1}{2}(\eta_{m'm}\eta_{n'n} + \eta_{m'n}\eta_{n'm} - \eta_{m'n'}\eta_{mn}) , \quad (46.17)$$

which has the property that, acting on any tensor φ^{mn} (not necessarily symmetric, or transverse, or traceless) satisfying the harmonic gauge condition (27.51), it yields a tensor that explicitly satisfies the harmonic condition (27.51),

$$P_{m'n'mn}\varphi^{mn} = \frac{1}{2}(\varphi_{m'n'} + \varphi_{n'm'} - \eta_{m'n'}\varphi_k^k) . \quad (46.18)$$

The projection operator (46.17) preserves the harmonic gauge condition from one vertex to another along a line in a Feynman diagram.

It is to be expected that, as in the electromagnetic propagator (45.12), a more general choice of Lorentz-covariant projection tensor than (46.17) would be admissible, but the possibilities are not explored here.

The spatial components $D_{a'b'ab}(x' - x)$ of the gravitational propagator (46.16) can be expressed as an expectation value of time-ordered products of the gravitational field operator $\hat{\mathbf{h}}_{ab}(x)$,

$$D_{a'b'ab}(x' - x) = \langle 0 | T \hat{\mathbf{h}}_{a'b'}(x') \hat{\mathbf{h}}_{ab}(x) | 0 \rangle , \quad (46.19)$$

which can be compared to the corresponding expressions (42.105), (44.83), and (45.15) for the Feynman-Klein-Gordon, Feynman-Dirac and Feynman-electromagnetic propagators.

Quantum electrodynamics

This Chapter gives an account of **quantum electrodynamics** (QED), which is the quantum field theory of charged fermions, such as electrons or muons, interacting with an electromagnetic field.

In today's low-energy world, the unbroken gauge symmetries of the standard model of physics comprise the product $U(1)_Q \times SU(3)_c$ of the electromagnetic symmetry $U(1)_Q$ and the colour symmetry $SU(3)_c$. The electromagnetic coupling parameter $e \equiv \sqrt{4\pi\alpha}$, equation (44.2), is small enough that perturbation theory works well, and scattering amplitudes and cross-sections can be adequately approximated by the lowest order Feynman diagrams. By contrast, the colour coupling strength $g_c \equiv \sqrt{4\pi\alpha_s}$, where α_s is the strong fine-structure parameter, exceeds 1 at low energy (Tanabashi et al., 2018), invalidating a perturbation expansion.

As summarized in §47.5, this Chapter goes through the basic 2-body processes of QED to lowest order, including spinor-photon scattering $e + \gamma \leftrightarrow e + \gamma$, spinor-antispinor creation from and annihilation into photons $\gamma + \gamma \leftrightarrow e + \bar{e}$, and Coulomb scattering $e + \mu \leftrightarrow e + \mu$.

The Chapter touches on how amplitudes and rates are calculated from general Feynman diagrams, §47.4, but it does not offer a formal derivation of the Feynman rules, nor does it cover the processes of renormalization that become essential when higher order diagrams are considered. The reader is referred to standard textbooks on quantum field theory, such as Peskin and Schroeder (1995), for further development.

47.1 Interactions

Consider the term

$$ie \bar{\psi} \cdot \mathbf{A} \psi \quad (47.1)$$

in the Dirac Lagrangian (44.30) that describes the interaction between a spinor ψ of charge e and the electromagnetic field $\mathbf{A} \equiv A^m \boldsymbol{\gamma}_m$. The interaction term (47.1) has a natural interpretation as a quantized interaction, illustrated in Figure 47.1. The interaction term is a product of three factors, a column spinor ψ that has quantum spin $\frac{1}{2}$, an electromagnetic potential $\mathbf{A}$ that is a vector of quantized spin 1, and a row spinor $\bar{\psi}$ that again has quantum spin $\frac{1}{2}$. In the quantum interaction illustrated in Figure 47.1, a column spinor ψ is destroyed at the interaction vertex, adjusted (multiplied) by the electromagnetic interaction $\mathbf{A}$,

and recreated as a row spinor $\bar{\psi}$ emerging from the interaction; the sequence follows the standard physics convention that interactions progress from right to left. Consistent with the discussion in §44.4.2, the column spinor ψ represents the destruction of a spinor, while the row spinor $\bar{\psi}$ represents the creation of another spinor. The electromagnetic factor $\mathbf{A}$ could represent either the destruction (absorption) of a photon, or else the creation (emission) of a photon, depending on whether the photon propagates backwards or forwards in time from the interaction.

The second key thing to notice about the spinor-electromagnetic interaction term $ie\bar{\psi}(x) \cdot \mathbf{A}(x)\psi(x)$ in the Dirac Lagrangian, besides the fact that it has a natural quantum interpretation, is that it is a product of three fields at the same spacetime point $x \equiv \{t, x_a\}$. This means that interactions between spinors and the electromagnetic field occur at spacetime points. If the Dirac action is recast into, for example, the momentum representation by Fourier transforming the spinor and electromagnetic fields, then the Lagrangian is *not* a product $ie\bar{\psi}(p) \cdot \mathbf{A}(p)\psi(p)$ of three fields at the same point p in energy-momentum space (nor, if only spatial components are Fourier transformed, is the Lagrangian a product $ie\bar{\psi}(t, p_a) \cdot \mathbf{A}(t, p_a)\psi(t, p_a)$ of three fields at the same point $\{t, p_a\}$ in time t and spatial momentum p_a). Thus the notion that spinors and the electromagnetic field interact at spacetime points is built into the Dirac Lagrangian. One might well suspect that something more complicated is going on at a deeper level, similarly perhaps to the way that thermodynamics and hydrodynamics deal with quantities defined at spacetime points, such as density, pressure, and temperature, that ultimately prove to be mean properties of collections of discrete objects, particles.

47.2 Single-vertex interaction

The interaction term (47.1) describes an interaction between a spinor ψ and the electromagnetic field $\mathbf{A}$ at a spacetime point x , as illustrated in Figure 47.1. In Minkowski space, quantum field theory interprets the spinor and electromagnetic fields as quanta of definite energy-momentum and spin that propagate as free waves except at the interaction point. With fields replaced by free eigenstates, $\psi \rightarrow \epsilon_{pb}e^{ip \cdot x}$, $\mathbf{A} \rightarrow \gamma_{ka}e^{ik \cdot x}$, and $\bar{\psi} \rightarrow \bar{\epsilon}_{p'b'}e^{-ip' \cdot x}$, as suggested by the expressions (44.58) for the spinor field operator $\hat{\psi}$ and (45.8a) for the electromagnetic field operator $\hat{\mathbf{A}}$, the interaction term (47.1) becomes

$$ie \bar{\epsilon}_{p'b'} \cdot \gamma_{ka} \epsilon_{pb} e^{i(-p'+k+p) \cdot x} . \quad (47.2)$$

The x -dependent exponential factors are free-wave eigenstates expressed as functions of spacetime x , which is necessary because, according to the Dirac Lagrangian, the interaction takes place at a point x in spacetime. Quantum field theory interprets the interaction term (47.2), a spacetime-, momentum-, and spin-dependent complex number, as a quantum mechanical amplitude for the quantized interaction illustrated in Figure 47.1 to occur. The quantized interaction term (47.2) is one of the two basic building blocks of Feynman diagrams, a vertex. It describes a single quantized interaction at a single spacetime point x .

If the spacetime is not Minkowski, then the plane waves in equation (47.2) should be replaced by eigenstates of the wave equations that are plane waves in the immediate vicinity of a locally inertial frame at the spacetime position x .

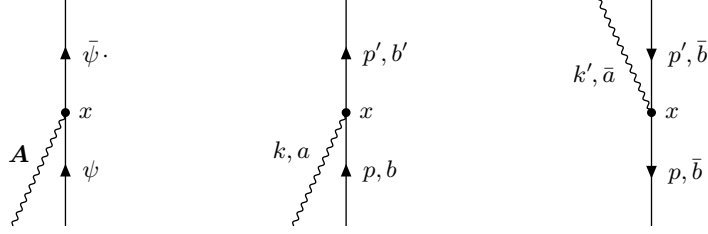


Figure 47.1 These Feynman diagrams illustrate one of the two building blocks of Feynman diagrams, a vertex. The other building block, illustrated in Figure 47.2, is a line joining vertices. (Left) An incoming spinor ψ interacts with the electromagnetic field $\mathbf{A}$ at spacetime position x , producing an outgoing spinor $\bar{\psi}$. (Middle) Same, but with energy-momenta and spins explicitly quantized. An incoming spinor of energy-momentum p and spin b interacts with a photon of energy-momentum k and spin a at spacetime position x , producing an outgoing spinor of energy-momentum p' and spin b' . (Right) The middle diagram flipped vertically describes an antispinor of opposite spin emitting instead of absorbing a photon, also of opposite spin. The middle and right diagrams are CPT conjugates of each other, and their amplitudes are the same.

It might seem that the mapping $\psi \rightarrow \epsilon_{pb} e^{ip \cdot x}$, $\mathbf{A} \rightarrow \boldsymbol{\gamma}_{ka} e^{ik \cdot x}$, and $\bar{\psi} \rightarrow \bar{\epsilon}_{p'b'} e^{-ip' \cdot x}$ has the wrong units: ψ has units $\text{mass}^{3/2}$ whereas ϵ_{pb} has units $\text{mass}^{1/2}$, and $\mathbf{A}$ has units mass whereas $\boldsymbol{\gamma}_{ka}$ is dimensionless. As will be seen in §47.4, the correct units are restored in the end because each pair of spinors ϵ , and each pair of photon vectors $\boldsymbol{\gamma}$, are integrated over $d^3p / (2E_p (2\pi)^3)$, which adds one factor of mass for each spinor or vector.

For spinors and photons of prescribed incoming and outgoing energy-momenta and spins, in the lowest order of perturbation theory the single interaction (47.2) is all that occurs. Since the spacetime location x of the interaction is unspecified, the interaction must be integrated over all spacetime, yielding

$$ie \int \bar{\epsilon}_{p'b'} \cdot \boldsymbol{\gamma}_{ka} \epsilon_{pb} e^{i(-p' + k + p) \cdot x} d^4x = ie \bar{\epsilon}_{p'b'} \cdot \boldsymbol{\gamma}_{ka} \epsilon_{pb} (2\pi)^4 \delta^4(p' - k - p). \quad (47.3)$$

The 4-dimensional delta-function expresses conservation of total energy-momentum by the interaction. In practice this conservation of total energy-momentum cannot be achieved in a single-vertex interaction, where a free spinor absorbs (or emits) just one photon. To see the impossibility of a single-photon interaction, without loss of generality take the outgoing spinor to be in its rest frame, $p' = \{m, 0\}$; then the spatial momenta of the incoming spinor and photon must be equal and opposite, $p_a = -k_a$; and then the energy of the incoming positive-energy spinor and photon must be $E + \omega = \sqrt{|p_a|^2 + m^2} + |p_a|$, which necessarily exceeds the outgoing energy m unless $p_a = 0$, that is, unless the photon energy-momentum is zero. To permit conservation of total energy-momentum in an interaction between a spinor and the electromagnetic field in Minkowski space, at least two interaction vertices are necessary, as considered in the next section 47.3.

The interaction between an antispinor and an electromagnetic field is described by the one-vertex term (47.2) with spinors replaced by antispinors, $\epsilon_{pb} \leftrightarrow i\bar{\epsilon}_{pb}$ in accordance with the rule (44.84). The resulting antispinor-

photon interaction term is

$$ie(-i\epsilon_{p'b'}) \cdot \gamma_{ka} (i\bar{\epsilon}_{pb}) e^{i(-p'+k+p) \cdot x} = ie \bar{\epsilon}_{pb} \cdot \gamma_{ka} \epsilon_{p'b'} e^{i(-p'+k+p) \cdot x}, \quad (47.4)$$

in which an overall $+$ sign follows from a product of a $-$ sign from the antisymmetry of the spinor metric, Table 40.1, and a $-$ sign from anticommuting basis vectors γ_a through the spinor metric, Table 40.2. If the spinor momenta and spins are swapped, $p \leftrightarrow p'$ and $b \leftrightarrow b'$, and each spin replaced by its conjugate (flipped) spin $b \leftrightarrow \bar{b}$, and if furthermore the photon momentum is flipped, $k \rightarrow k' = -k$ (corresponding to emission rather than absorption) and the photon spin replaced by its conjugate (flipped) spin, then the right hand side of equation (47.4) becomes

$$ie \bar{\epsilon}_{p'\bar{b}'} \cdot \gamma_{k'a} \epsilon_{p\bar{b}} e^{i(-p-k+p') \cdot x}, \quad (47.5)$$

whose exponential factor is the complex conjugate of the exponential factor of the original spinor-photon scattering vertex (47.2). Integration over all x imposes energy-momentum conservation $p' - k - p = 0$, which coincides with the energy-momentum conservation relation (47.3) for the original spinor-photon scattering vertex. Thus the spinor-photon and antispinor-photon scattering vertices yield the same amplitude provided that the Feynman diagram is flipped vertically, as in the right panel of Figure 47.1, and all spins are flipped. The equality of amplitudes is an expression of CPT symmetry.

Electromagnetism is parity invariant, so flipping the spins of all particles in an electromagnetic interaction does not change the amplitude. However, if the gauge boson were an electroweak gauge boson, transforming under $SU(2)_L$, then it would be essential to flip spins, since the electroweak symmetry breaks parity maximally.

47.3 Two-vertex interaction

Consider stringing together two interactions of the form (47.2),

$$(ie \bar{\epsilon}_{p'b'} \cdot \gamma_{k'a'} \overbrace{\epsilon_{qc}}^{\text{Wick bracket}}) (ie \bar{\epsilon}_{qc} \cdot \gamma_{ka} \epsilon_{pb}) e^{i(-p'-k'+q) \cdot x'} e^{i(-q+k+p) \cdot x}, \quad (47.6)$$

a process illustrated in the left panel of Figure 47.2. In this process, an incoming spinor ϵ_{pb} and photon γ_{ka} are destroyed at the spacetime point x , where they are recreated as the intermediate spinor $\bar{\epsilon}_{qc}$, which then propagates to the spacetime point x' , where it is destroyed and recreated as the outgoing spinor $\bar{\epsilon}_{p'b'}$ and photon $\gamma_{k'a'}$. The propagation of the intermediate spinor from x to x' is described by the Feynman-Dirac propagator (44.81), the Green's function of the wave equation starting at x and ending at x' . The propagator ensures that the momentum q and spin c of the spinor are the same at both points, as would be expected from conservation of momentum and spin. The Wick bracket $\overline{}$ linking the creation operator $\bar{\epsilon}_{qc}$ to the destruction operator ϵ_{qc} in the expression (47.6) emblemizes that the operators are two ends of the same spinor leg, and that they share the same momentum q and spin c .

The process (47.6) describes the scattering of a spinor by an electromagnetic field. A second two-vertex interaction that contributes to the same process, illustrated in the right panel of Figure 47.2, is one in which the incoming photon interacts with the outgoing spinor instead of the incoming spinor, and likewise the

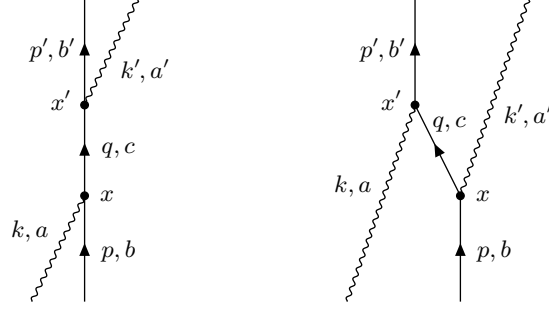


Figure 47.2 These Feynman diagrams illustrate both of the two basic building blocks of Feynman diagrams: spacetime vertices, here x and x' , where quantized interactions occur; and lines joining vertices, here a spinor line joining x and x' that represents a spinor of momentum q and spin c propagating freely between the interaction vertices. The two-vertex spinor-photon interactions are described by equations (47.6) and (47.7) respectively. In both cases the incoming spinor and photon have energy-momenta and spins respectively pa and kb , while the outgoing spinor and photon have energy-momenta and spins respectively $p'a'$ and $k'b'$. The two diagrams differ in which spinor interacts with which photon.

outgoing photon interacts with the incoming spinor instead of the outgoing spinor. This second interaction is described by

$$(ie \bar{\epsilon}_{p'b'} \cdot \overline{\gamma_{ka}} \epsilon_{qc})(ie \overline{\epsilon_{qc}} \cdot \gamma_{k'a'} \epsilon_{pb}) e^{i(-p'+k+q) \cdot x'} e^{i(-q-k+p) \cdot x}, \quad (47.7)$$

which differs from the first interaction (47.6) by the exchange $\gamma_{ka} \leftrightarrow \gamma_{k'a'}$ of the incoming and outgoing photons.

Quantum field theory interprets the two two-vertex interactions (47.6) and (47.7) as complex-valued quantum mechanical amplitudes for the process to occur. Both processes contribute to the amplitude, with the net amplitude of the process being the sum of the two, equation (47.30).

A process indistinguishable from the second interaction (47.7) is one in which the spacetime position x' , instead of being in the future of x ($t' > t$), is to the past of x ($t' < t$), as illustrated in Figure 47.3. According to the Feynman-Dirac propagator (44.81), the Wick-bracketed pair $\overline{\epsilon_{qc}} \epsilon_{qc}$ when $t' < t$ is to be interpreted as the backward-in-time outer product $-\overline{\epsilon_{qc}} \epsilon_{qc}$ in place of the forward-in-time outer product $\epsilon_{qc} \overline{\epsilon_{qc}}$.

Despite appearances, the intermediate momentum q that joins the spacetime vertices x and x' in Figures 47.2 and 47.3 is *not* aligned with the direction $x' - x$. Rather, q is the energy-momentum of a plane wave that covers all of spacetime, and all plane waves, regardless of their energy-momentum, pass through both interaction points x and x' .

In the two-vertex interactions (47.6) and (47.7), the initial and final energy-momenta and spins are all specified, but the energy-momentum and spin q and c of the intermediate spinor ϵ_{qc} are not. To cover all possibilities, the energy-momentum and spin of the intermediate spinor ϵ_{qc} must be integrated and summed over. The integration and summation is built in to the Feynman-Dirac propagator (44.81). This is an example

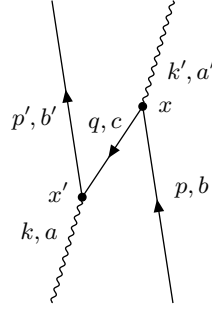


Figure 47.3 Two-vertex spinor-photon interaction like that in the right panel of Figure 47.2, but with the intermediate spinor moving backwards in time from x to x' . The situation is indistinguishable from one where a spinor-antispinor pair is created from an incoming photon at x' , and then the spinor part of the pair continues onward, while its antispinor partner annihilates with the incoming spinor at x , leaving an outgoing photon.

of the Feynman rule that the energy-momentum and spin of each internal line of a Feynman diagram must be integrated and summed over.

The second thing that remains indeterminate in the two-vertex interactions (47.6) and (47.7) is the positions x and x' of each of the interaction points, and each of these must be integrated over all of 4-dimensional spacetime. Integrating the interactions over x and x' enforces conservation of energy-momentum at each of the vertices. This is an example of the Feynman rule that the spacetime position of each interaction vertex of a Feynman diagram must be integrated over.

47.3.1 Energy-momentum conservation at interaction vertices

For the two-vertex interactions (47.6) and (47.7), conservation of energy-momentum at each vertex, enforced by integration over all x and x' , implies that the energy-momentum q of the intermediate spinor is related to the energy-momenta of the incoming and outgoing spinors and photons by, for the two cases,

$$p + k = q = p' + k' , \quad (47.8a)$$

$$p - k' = q = p' - k . \quad (47.8b)$$

The signs in equations (47.8) assume that Figures 47.2 and 47.3 are drawn in the convention of spacetime diagrams, where time goes upwards, and that the time components E, E', ω, ω' of the energy-momenta of each of the incoming and outgoing spinors and photons are positive.

Importantly, the energy-momentum q of the intermediate spinor, equations (47.8), does *not* satisfy the mass condition $q^2 + m^2 = 0$. Since $q^2 + m^2$ is a Lorentz scalar, it can be evaluated in any frame. A convenient frame is the centre-of-mass frame where the total incoming and outgoing spatial momenta are zero, $p_a + k_a = p'_a + k'_a = 0$, illustrated in the left panel of Figure 47.4. The scattering can be characterized by two parameters in the centre-of-mass frame, the rapidity η of the spinor, and the scattering angle θ . If the incoming photon is oriented in the 3-direction (z -direction), and the outgoing photon in the 31-plane

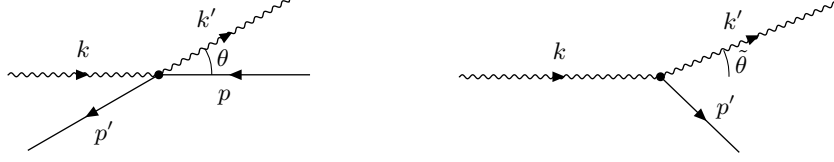


Figure 47.4 Spinor-photon scattering (left) in the centre-of-mass frame, where the total incoming and outgoing spatial momenta are zero, $p_a + k_a = p'_a + k'_a = 0$, and (right) in the rest frame of the incoming spinor, where $p_a = 0$. The lengths and directions of the lines are proportional to the spatial momenta. In the centre-of-mass frame, the spinor and photon are each deflected by angle θ . In the rest frame of the incoming spinor, the photon is deflected by angle $\tilde{\theta}$, equation (47.11). The parameters of the collision in this instance are $\eta = \text{atanh } \frac{7}{25}$ and $\theta = \frac{\pi}{6}$.

(zx -plane), then the energy-momenta of the incoming and outgoing spinors and photons in the centre-of-mass frame are

$$p = m\{\cosh \eta, 0, 0, -\sinh \eta\} , \quad (47.9a)$$

$$p' = m\{\cosh \eta, -\sinh \eta \sin \theta, 0, -\sinh \eta \cos \theta\} , \quad (47.9b)$$

$$k = m \sinh \eta \{1, 0, 0, 1\} , \quad (47.9c)$$

$$k' = m \sinh \eta \{1, \sin \theta, 0, \cos \theta\} . \quad (47.9d)$$

In the limit of zero mass, $m \cosh \eta \rightarrow m \sinh \eta \rightarrow E$. Another convenient frame is the “lab” frame, the rest frame of the incoming spinor, where $p_a = 0$, illustrated in the right panel of Figure 47.4. Lorentz boosting (by rapidity η) the centre-of-mass energy-momenta (47.9) yields energy-momenta in the lab frame,

$$p = m\{1, 0, 0, 0\} , \quad (47.10a)$$

$$p' = m\{\cosh^2 \eta - \sinh^2 \eta \cos \theta, -\sinh \eta \sin \theta, 0, \cosh \eta \sinh \eta (1 - \cos \theta)\} , \quad (47.10b)$$

$$k = m e^\eta \sinh \eta \{1, 0, 0, 1\} , \quad (47.10c)$$

$$k' = m \sinh \eta \{\cosh \eta + \sinh \eta \cos \theta, \sin \theta, 0, \sinh \eta + \cosh \eta \cos \theta\} . \quad (47.10d)$$

The photon scattering angle $\tilde{\theta}$ in the lab frame can be read off from equation (47.10d),

$$\sin \tilde{\theta} = \frac{\sin \theta}{\cosh \eta + \sinh \eta \cos \theta} , \quad \cos \tilde{\theta} = \frac{\sinh \eta + \cosh \eta \cos \theta}{\cosh \eta + \sinh \eta \cos \theta} . \quad (47.11)$$

The intermediate spinor energy-momentum q for the two cases, equations (47.8), then satisfy,

$$q^2 + m^2 = (p + k)^2 + m^2 = 2p \cdot k = -2m\omega , \quad (47.12a)$$

$$q^2 + m^2 = (p - k')^2 + m^2 = -2p \cdot k' = 2m\omega' , \quad (47.12b)$$

where ω and ω' are the initial and final photon energies in the lab frame, from equations (47.10),

$$\omega = m e^\eta \sinh \eta , \quad \omega' = m \sinh \eta (\cosh \eta + \sinh \eta \cos \theta) . \quad (47.13)$$

In the first case (47.12a), $q^2 + m^2$ is strictly negative, while in the second case (47.12b), $q^2 + m^2$ is strictly positive (if a photon is present, as is being assumed, then $\omega > \omega' > 0$),

$$q^2 + m^2 < 0 , \quad (47.14a)$$

$$q^2 + m^2 > 0 . \quad (47.14b)$$

The intermediate energy-momentum q is said to be off-shell, meaning not satisfying the (on-shell) mass condition $q^2 + m^2 = 0$. The fact that the energy-momentum q of the intermediate spinor is off-shell may seem disconcerting, but the situation is especially surprising in the second case (47.12b), where q can actually be spacelike, $q^2 > 0$, if the rapidity η is large enough and the scattering angle θ small enough,

$$q^2 > 0 \quad \text{if} \quad \cos \frac{\theta}{2} > \frac{1}{2e^\eta \sinh \eta} . \quad (47.15)$$

Spacelike q means that the intermediate spinor ϵ_{qc} can apparently move faster than light between the interaction vertices x and x' , an alarming result since it would seem to admit a violation of causality.

The strange faster-than-light behaviour of the intermediate spinor in the second case (47.12b) is a symptom of the usual wave-particle paradox of quantum mechanics. In quantum mechanics, the act of observation locates a wave-particle at just one place, instantaneously transmitting to all other (spacelike- and timelike-separated) places that the particle should not appear anywhere else. Ingeniously, quantum mechanics achieves this instantaneous transmission without actually violating causality. Quantum mechanics cleverly cleans up after itself: before you observe a wave-particle it explores a vast range of possibilities; but after you observe it, it apparently explored only the subset of possibilities consistent with the observation.

47.4 Amplitudes and rates

This section 47.4 discusses how a process such as the spinor-photon scattering illustrated in the Feynman diagrams 47.2 and 47.3 is converted into a quantum mechanical amplitude, and then how the square of that amplitude yields a rate for the process.

As discussed above in §§47.2 and 47.3, the building blocks of Feynman diagrams are vertices, where quantized interactions occur, and lines joining vertices, which represent particles propagating between interactions. Basic quantum field theory takes the background spacetime to be Minkowski, in which case particles propagate as free waves between interactions. Converting a process represented by a Feynman diagram into a quantum mechanical amplitude $\mathcal{M}$ ($\mathcal{M}$ for matrix) for the diagram requires integrating over all spacetime interaction vertices x , and over all momenta q and spins c of particles on internal lines joining vertices. The defining expression for the complex amplitude $\mathcal{M}$ contributed by a Feynman diagram is an integral over a product of vertices and internal lines (momentum and spin indices in the following equation (47.16) are suppressed for brevity; equation (47.18) gives an explicit example with indices restored),

$$\mathcal{M}(2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}}) \equiv \int \left(\prod_{\text{vertices}} ie \bar{\epsilon} \cdot \gamma \epsilon e^{i \sum p \cdot x} d^4x \right) \left(\prod_{\text{internal lines}} \sum_c \frac{d^3q}{2E_q (2\pi)^3} \Big|_{p^0=E_p} \right) . \quad (47.16)$$

The 4-dimensional delta-function $(2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}})$ expresses overall conservation of energy-momentum, the incoming and outgoing energy-momenta P_{in} and P_{out} being sums over energy-momenta of respectively incoming and outgoing particles.

The 3-dimensional integration over momenta q in the amplitude (47.16) derives from the Feynman propagator, and may give the impression that the integral is only over the 3 spatial components q_a of the momentum, the time component q^0 being set on-shell, $q^0 = E_q = \sqrt{q_a q^a + m^2}$, but this is incorrect. The 3-dimensional version of the Feynman propagator comes from integrating the 4-dimensional version over the entire range of energies q^0 from $-\infty$ to $+\infty$, which when treated as an integral over a complex energy q^0 in the complex plane acquires its nonzero values from the poles $q^2 + m^2 = 0$ of the wave equation, as in equation (42.93). In practice it is often more convenient to evaluate amplitudes by integrating over the full 4-dimensional energy-momentum d^4q , as for example in equations (47.30), (47.33), (47.64), and so on, since this removes the need to worry about $\pm$ signs in the 3-dimensional integral d^3q .

For a diagram with v vertices and i internal lines, the dimension of the amplitude $\mathcal{M}$ is

$$\mathcal{M} \sim \text{mass}^{4-3v+2i} = \text{mass}^{4-e}, \quad (47.17)$$

where in the final expression e is the number of external lines (the combined number of incoming and outgoing particles), equation (47.27). If the number of external lines is $e = 4$, as in the Feynman diagrams 47.2, then the amplitude $\mathcal{M}$ is dimensionless, but otherwise $\mathcal{M}$ is dimensionful.

For example, the expression (47.16) for the amplitude $\mathcal{M}$ in the case of the first (47.6) of the two two-vertex interactions is, with indices restored,

$$\begin{aligned} & \mathcal{M}^1 (2\pi)^4 \delta^4(p' + k' - p - k) \\ &= \int \sum_c (ie \bar{\epsilon}_{p'b'} \cdot \gamma_{k'a'} \epsilon_{qc}) (ie \bar{\epsilon}_{qc} \cdot \gamma_{ka} \epsilon_{pb}) e^{i(-p'-k'+q) \cdot x'} e^{i(-q+k+p) \cdot x} \Big|_{q^0=E_q} \frac{d^3q}{2E_q (2\pi)^3} d^4x' d^4x. \end{aligned} \quad (47.18)$$

Equation (47.18) is for x' to the future of x (that is, $t' - t > 0$); if x' is to the past of x , then q should be flipped in sign, and $\epsilon_{qc} \bar{\epsilon}_{qc} \cdot$ replaced by $-\bar{\epsilon}_{qc} \epsilon_{qc} \cdot$, in accordance with equation (44.81) for the Dirac propagator. The second (47.7) of the two two-vertex interactions yields an amplitude $\mathcal{M}^2$ similar to equation (47.18), but with $\gamma_{k'a'} \leftrightarrow \gamma_{ka}$, and $k' \leftrightarrow -k$ in the exponentials. The two amplitudes combine into a net amplitude $\mathcal{M}$,

$$\mathcal{M} = \mathcal{M}^1 + \mathcal{M}^2. \quad (47.19)$$

The amplitude $\mathcal{M}$ is a matrix element between incoming and outgoing spinors and photons of prescribed energy-momenta and spins,

$$\mathcal{M} = \langle (p'b') (k'a') | (ka) (pb) \rangle. \quad (47.20)$$

The integration in equation (47.18) will be carried out in the next section 47.6.

The two two-vertex interactions (47.6) and (47.7) represented by the Feynman diagram 47.2 describe spinor-photon scattering to lowest nonvanishing order of perturbation theory. Higher order interactions involve more vertices, and more internal lines joining them. The net amplitude for the process of spinor-photon scattering with one incoming spinor and photon, and one outgoing spinor and photon, is a sum over

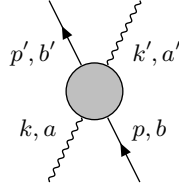


Figure 47.5 The amplitude $\mathcal{M}$ for the spinor-photon scattering process is a sum $\sum_i \mathcal{M}^i$ over all Feynman diagrams that connect an incoming spinor and photon to an outgoing spinor and photon, equation (47.21).

all Feynman diagrams, as illustrated in Figure 47.5,

$$\mathcal{M} = \sum_i \mathcal{M}^i . \quad (47.21)$$

The amplitude of the two-vertex interactions is proportional to charge squared, and the square of the amplitude is proportional to the fourth power of charge, $e^4 = (4\pi\alpha)^2 \approx 0.1$, equation (44.2), which is small enough that the lowest nonvanishing order of perturbation theory yields an adequate approximation to the process of spinor-photon scattering,

$$\mathcal{M} \approx \mathcal{M}^1 + \mathcal{M}^2 . \quad (47.22)$$

Turning a complex amplitude $\mathcal{M}$ into a rate involves taking an absolute square of (47.16), and integrating/summing over external momenta and spins. The square of the overall energy-momentum-conserving delta-function $(2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}})$ requires interpretation. Put the system into a large but finite 4-dimensional periodic box of spatial volume $V = L^3$ and time extent $T = L$. Then each of the 4 components p^m of the energy-momenta p runs over discrete values $p^m = 2\pi n/L$ with integral n , and the orthogonality of Fourier eigenmodes is expressed by

$$\int e^{i(p'-p)\cdot x} d^4x = VT \delta_{p'p}^4 , \quad (47.23)$$

where $\delta_{p'p}^4$ denotes a 4-dimensional Kronecker delta. When the box is enlarged to infinity and the discrete energy-momenta replaced by a continuum, the right hand side of the orthogonality relation (47.23) becomes $(2\pi)^4 \delta^4(p' - p)$. The square of the orthogonality relation (47.23) can therefore be interpreted as $VT(2\pi)^4 \delta^4(p' - p)$ in the limit of a large box. The product VT of spatial volume V and time extent T is Lorentz invariant.

When the defining expression (47.16) for the amplitude $\mathcal{M}$ is squared, one factor of the overall energy-momentum-conserving delta-function $(2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}})$ can be taken over to the left hand side, and the result reinterpreted as yielding a rate, the number of events per unit volume and unit time (again with spin indices suppressed for brevity; equation (47.36) gives an explicit example with spin indices restored),

$$\frac{\text{number of events}}{\text{volume} \cdot \text{time}} = \int \sum_{\text{spins}} \mathcal{M} \mathcal{M}^* (2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}}) \left(\prod_{\text{outgoing}} (1 \pm f) \frac{d^3p}{2E_p (2\pi)^3} \right) \left(\prod_{\text{incoming}} f \frac{d^3p}{2E_p (2\pi)^3} \right) . \quad (47.24)$$

The factors f in equation (47.24) are density matrices, also known as occupation numbers (density matrices yield numbers when diagonalized), with (suppressed) spin indices that contract against the (suppressed) spin indices of the amplitude factors $\mathcal{M}$. The factors $1 \pm f$ in the outgoing particles are boson-stimulation (+) and fermion-blocking (−) factors. The range of integration over incoming and outgoing energy-momenta, and range of summation over incoming and outgoing spins, depends on the system being considered: what incoming energy-momenta and spins are present, and what outgoing energy-momenta and spins are being observed?

The complex conjugate of a spinor-photon vertex term is, equation (41.126),

$$(ie \bar{\epsilon}_{qc} \cdot \gamma_{ka} \epsilon_{pb})^* = (ie \bar{\epsilon}_{pb} \cdot \bar{\gamma}_{ka} \epsilon_{qc}) . \quad (47.25)$$

That is, the complex conjugate of a spinor-photon vertex term may be obtained by conjugating each of the spinors and photons ($\epsilon \leftrightarrow \bar{\epsilon}$ and $\gamma \leftrightarrow \bar{\gamma}$), and reversing their order. Conjugation leaves orthonormal (linearly polarized) photon vectors unchanged, $\bar{\gamma}_a = \gamma_a$ for $a = 1, 2$, equation (41.115), but flips the polarization of chiral (circularly polarized) photon vectors, $\bar{\gamma}_{\pm} = \gamma_{\mp}$, equation (41.116).

The rate (47.24) for a process represented by a Feynman diagram can be written in a suggestive shorthand. In the case of the two-vertex interaction (47.18), for example, the number of events in a hypothetical large box VT enclosing the interaction is

$$\text{number of events in } VT = \left((ie \bar{\epsilon} \cdot \gamma \epsilon) \overbrace{(ie \bar{\epsilon} \cdot \gamma \epsilon)} \right) \left((ie \bar{\epsilon} \cdot \bar{\gamma} \epsilon) \overbrace{(ie \bar{\epsilon} \cdot \bar{\gamma} \epsilon)} \right) . \quad (47.26)$$

Each factor $(ie \epsilon \cdot \gamma \bar{\epsilon})$ in equation (47.26) represents a vertex which must be integrated over spacetime d^4x . Each Wick-bracketed pair involves an integral over momenta $d^3p / (2E_p(2\pi)^3)$, and a possible sum over spins. Each Wick-bracketed pair with brackets above represent an internal line, which must be integrated over all momenta and summed over all spins. Each Wick-bracketed pair with brackets below represents an external line, for which the integration/summation over momenta and spins is over the initial and final distributions of incoming and outgoing particles, as for example in equation (31.40). As motivated around equation (47.23), the integrals over vertices on the right hand side of equation (47.26) produce two overall energy-momentum-conserving delta-functions $(2\pi)^4 \delta^4(P_{\text{out}} - P_{\text{in}})$, and one of those factors can be reinterpreted as the volume times time VT of the hypothetical large box enclosing the interaction.

As remarked in §47.2, quantizing the interaction term $ie\bar{\psi} \cdot \mathbf{A}\psi$ in the Dirac Lagrangian by replacing $\psi \rightarrow \epsilon_{pb} e^{ip \cdot x}$, $\mathbf{A} \rightarrow \gamma_{ka} e^{ik \cdot x}$, and $\bar{\psi} \rightarrow \bar{\epsilon}_{p'b'} e^{-ip' \cdot x}$, equation (47.2), seems to have the wrong units: the spinor field ψ has units mass times ϵ , and likewise the electromagnetic field $\mathbf{A}$ has units mass times γ . The example rate (47.26) illustrates the resolution of the discrepancy: each pair of basis spinors ϵ , and each pair of basis vectors γ , is integrated over $d^3p / (2E_p(2\pi)^3)$, which has units mass², restoring the correct units. Notice that the $\sqrt{2m}$ factors introduced into the definitions (44.47) of boosted spinors ensure not only that spinor sums (44.48) over spins are finite in the massless limit, but also that spinor integration intervals $d^3p / (2E_p(2\pi)^3)$ are finite in the massless limit.

Exercise 47.1 Relation between numbers of vertices, lines, and loops in a Feynman diagram.

Consider a general Feynman diagram in a quantum field theory where each vertex connects n lines. For example, $n = 3$ in QED, as illustrated in Figure 47.1. Show that, in any number N of spacetime dimensions, the number e of external lines is related to the numbers v of vertices and i of internal lines by

$$e = nv - 2i, \quad (47.27)$$

while the number i of internal lines is related to the numbers v of vertices, d of disconnected diagrams, and l of loops by

$$i = v - d + l. \quad (47.28)$$

Solution. To prove the relation (47.27), proceed by induction on the number i of internal lines. Suppose that the diagram contains v vertices and no internal lines, $i = 0$, and no loops, $l = 0$. Then there are v disconnected diagrams each containing n external lines, so $e = nv$ and $d = v$. Now introduce internal lines, one by one, without changing the number of vertices. For each internal line introduced, $i \rightarrow i + 1$, two external lines must be joined, $e \rightarrow e - 2$. The relation (47.27) follows. To prove the relation (47.28), notice that each time an internal line is introduced, $i \rightarrow i + 1$, then either the number of disconnected diagrams is reduced by one and the number of loops remains the same, $d \rightarrow d - 1$ and $l \rightarrow l$, or else the number of disconnected diagrams remains the same and the number of loops increases by one, $d \rightarrow d$ and $l \rightarrow l + 1$. The relation (47.28) follows. More generally, if the number n of lines connected to a vertex varies by vertex, then the relation (47.27) holds with nv equal to the sum of n over vertices.

47.5 Strategy

The strategy for the rest of this Chapter is to go through the various basic 2-body processes of quantum electrodynamics at lowest non-trivial (tree level) order in perturbation theory, discovering on the way how the rules of quantum field theory are implemented.

A summary of the processes covered and the insights gained is as follows.

1. Spinor-photon scattering, for example $e + \gamma \leftrightarrow e + \gamma$, a discussion which has been started already §§47.3–47.4, and which will be completed in §47.6. Brings out how the Dirac propagator (44.81) applies. A fermionic minus sign, equation (44.46b), is associated with whether a spinor is positive or negative mass and therefore whether it is going forwards or backwards in time.
2. Spinor-antispinor creation from and annihilation into photons, for example $\gamma + \gamma \leftrightarrow e + \bar{e}$, §47.8. The Feynman diagrams for this process are related to those of spinor-photon scattering by flipping one each of the external spinor and photon legs. The form of the Dirac propagator (44.81), required by Lorentz invariance, implies the rule (44.84) for replacing a spinor by an antispinor.
3. Coulomb scattering of distinguishable spinors, for example $e + \mu \leftrightarrow e + \mu$, and of indistinguishable spinors, for example $e + e \leftrightarrow e + e$, §47.9. Brings out how the photon propagator (45.13) applies. For

indistinguishable pairs, the spinor-exchanged interaction (47.94) has an extra fermionic minus sign compared to the direct interaction (47.72), a minus sign that ensures that the net scalar amplitude (47.103) is symmetric under exchange of identical spinors, as it should be. By comparison, there is no such minus sign for photons, the net scalar amplitude (47.68) for spinor-antispinor pair creation/annihilation from/into photons being symmetric under exchange of identical photons without any additional minus sign.

4. Spinor-antispinor pair annihilation into a distinguishable pair, for example $e + \bar{e} \leftrightarrow \mu + \bar{\mu}$, or into an indistinguishable pair, for example $e + \bar{e} \leftrightarrow e + \bar{e}$, §47.10. The Feynman diagrams for this process are related to those of Coulomb scattering by flipping one of the external legs of each spinor-antispinor pair.

47.6 Spinor-photon scattering

The two-vertex interactions (47.6) and (47.7) represent the scattering of a charged spinor by an electromagnetic field, as illustrated in Figure 47.2, to lowest nonvanishing order of perturbation theory. If the charged spinor is an electron, for example, then the process represents electron-photon scattering,

$$e + \gamma \rightarrow e + \gamma . \quad (47.29)$$

The general process of scattering of photons by charged particles is also known as Compton scattering. The nonrelativistic limit of Compton scattering is called Thomson scattering, which plays a central role in cosmological recombination, §33.8.

Equation (47.18) gives an integral expression for the quantum-mechanical amplitude $\mathcal{M}^1$ for the first (47.6) of the two two-vertex interactions; the amplitude $\mathcal{M}^2$ of the second (47.7) of the two interactions is similar with $\gamma_{k'a'} \leftrightarrow \gamma_{ka}$, and $k' \leftrightarrow -k$ in the exponentials. The integral of (47.18) over spacetime vertices x' and x imposes conservation of energy-momentum at each vertex, while the sum and integral over the intermediate spinor spin c and momenta q yields the Dirac propagator (44.81). The integrations are most straightforward if the Dirac propagator is written in the form (44.79) of a 4-dimensional integral d^4q over momenta, which removes the need to worry about $\pm$ signs in the 3-dimensional integral d^3q , equation (44.80). The result for the combined amplitude of the two two-vertex interactions is

$$\mathcal{M} = \mathcal{M}^1 + \mathcal{M}^2 = (ie)^2 \left(\frac{\bar{\epsilon}_{p'b'} \cdot \gamma_{k'a'} (-i\mathbf{q} + m) \gamma_{ka} \epsilon_{pb}}{q^2 + m^2} \right)_{\substack{q = p + k \\ = p' + k'}} + \frac{\bar{\epsilon}_{p'b'} \cdot \gamma_{ka} (-i\mathbf{q} + m) \gamma_{k'a'} \epsilon_{pb}}{q^2 + m^2} \bigg|_{\substack{q = p - k' \\ = p' - k}} , \quad (47.30)$$

in which the integration over d^4q has imposed energy-momentum conservation, $q = p + k = p' + k'$ for $\mathcal{M}^1$ and $q = p - k' = p' - k$ for $\mathcal{M}^2$.

The amplitude (47.30) can be evaluated using the supergeometric algebra, Chapter 40. Since the amplitude $\mathcal{M}$ is a scalar, it can be evaluated in any convenient frame of reference. A convenient frame in the present case is the centre-of-mass frame, illustrated in the left panel of Figure 47.4, with the incoming photon directed along the 3-direction (z -direction), and the outgoing photon lying in the 31-plane (zx -plane). The energy-momenta of the incoming and outgoing spinors and photons in this centre-of-mass frame are given

by equations (47.9). Spinors and photons of given spin and energy-momentum are obtained by Lorentz-transforming them out of a canonical frame, as in equations (44.47) for spinors, or (45.2) for photons. The Lorentz rotors relevant in the present case are a Lorentz rotor R_η that Lorentz boosts along the 3-direction, and a spatial rotor R_θ that rotates about the 31-direction (a right-handed rotation about the 2-direction):

$$R_\eta \equiv \cosh \frac{\eta}{2} + \gamma_{03} \sinh \frac{\eta}{2} , \quad (47.31a)$$

$$R_\theta \equiv \cos \frac{\theta}{2} - \gamma_{31} \sin \frac{\theta}{2} . \quad (47.31b)$$

In the centre-of-mass frame, the ingoing and outgoing spinors are both boosted by rapidity η out of their rest frames. Relative to the direction of the incoming photon momentum k , the outgoing photon momentum k' is rotated by θ , while the incoming and outgoing spinor momenta p and p' are rotated by respectively π and $\theta + \pi$. In the centre-of-mass frame, the energy-momenta $\overset{1}{q}$ and $\overset{2}{q}$ of the intermediate spinor in the first and second cases are, equations (47.8),

$$\overset{1}{q} = p + k = m\{e^\eta, 0, 0, 0\} , \quad (47.32a)$$

$$\overset{2}{q} = p - k' = m\{e^{-\eta}, -\sinh \eta \sin \theta, 0, -\sinh \eta(1 + \cos \theta)\} . \quad (47.32b)$$

The amplitudes in terms of Lorentz-transformed basis spinors and vectors are then

$$\mathcal{M}^1 = (ie)^2 \frac{(2m)(\bar{\epsilon}_{\top b'} \cdot \tilde{R}_\eta \tilde{R}_{\theta+\pi})(R_\theta \gamma_{a'} \tilde{R}_\theta)(-i\mathbf{q} + m)(\gamma_a)(R_\pi R_\eta \epsilon_{\top b})}{q^2 + m^2} \Big|_{q=\overset{1}{q}} , \quad (47.33a)$$

$$\mathcal{M}^2 = (ie)^2 \frac{(2m)(\bar{\epsilon}_{\top b'} \cdot \tilde{R}_\eta \tilde{R}_{\theta+\pi})(\gamma_a)(-i\mathbf{q} + m)(R_\theta \gamma_{a'} \tilde{R}_\theta)(R_\pi R_\eta \epsilon_{\top b})}{q^2 + m^2} \Big|_{q=\overset{2}{q}} , \quad (47.33b)$$

where the intermediate energy-momenta q in the two cases are given by equations (47.32). To guide the eye, each Lorentz-transformed spinor and photon in equations (47.33) is enclosed in parentheses; the parentheses are not necessary, since the expressions satisfy the associative law.

Any basis of spin eigenvectors can be chosen. For Dirac spinors, a natural choice of spin eigenvectors is up ($\uparrow$) and down ($\downarrow$) in a frame aligned with the direction of motion of the spinors in the centre-of-mass frame. For photons, a natural choice of spin eigenvectors is right-handed (+) and left-handed (−) spins (right- and left-handed circular polarizations), again in a frame aligned with the direction of motion of the photons. Since the incoming spinor and photon in the centre-of-mass frame are moving in opposite directions, as illustrated in the left panel of Figure 47.4, the adopted spin convention means that, for example, the spins of an incident + photon and an incident $\uparrow$ spinor point in opposite directions. For this choice of spin eigenvectors,

the amplitudes (47.33) evaluate to, for the various outgoing and incoming spins,

$$\left. \begin{array}{l} \langle \uparrow - | + \uparrow \rangle = \langle \downarrow + | - \downarrow \rangle \\ \langle \downarrow + | + \uparrow \rangle = -\langle \uparrow - | - \downarrow \rangle \end{array} \right\}_1 = (ie)^2 2 \left\{ \begin{array}{l} \cos \frac{\theta}{2}, \\ e^{-\eta} \sin \frac{\theta}{2}, \end{array} \right. \quad (47.34a)$$

$$\left. \begin{array}{l} \langle \uparrow - | + \uparrow \rangle = \langle \downarrow + | - \downarrow \rangle \\ \langle \uparrow + | - \uparrow \rangle = \langle \downarrow - | + \downarrow \rangle \\ \langle \uparrow + | + \uparrow \rangle = \langle \uparrow - | - \uparrow \rangle = \langle \downarrow + | + \downarrow \rangle = \langle \downarrow - | - \downarrow \rangle \\ \langle \downarrow + | + \uparrow \rangle = -\langle \uparrow - | - \downarrow \rangle \\ \langle \downarrow - | + \uparrow \rangle = \langle \downarrow + | - \uparrow \rangle = -\langle \uparrow + | - \downarrow \rangle = -\langle \uparrow - | + \downarrow \rangle \\ \langle \downarrow - | - \uparrow \rangle = -\langle \uparrow + | + \downarrow \rangle \end{array} \right\}_2 = (ie)^2 \frac{2}{\cosh \eta + \sinh \eta \cos \theta} \left\{ \begin{array}{l} (e^\eta - 2e^{-\eta}) \cos \frac{\theta}{2} \sin^2 \frac{\theta}{2}, \\ e^\eta \cos^3 \frac{\theta}{2}, \\ -e^{-\eta} \cos \frac{\theta}{2} \sin^2 \frac{\theta}{2}, \\ -\cos^2 \frac{\theta}{2} \sin \frac{\theta}{2}, \\ -\cos^2 \frac{\theta}{2} \sin \frac{\theta}{2}, \\ \sin^3 \frac{\theta}{2}. \end{array} \right. \quad (47.34b)$$

Equations (47.34a) and (47.34b) give the components $\mathcal{M}_{b'a'ab} \equiv \langle b'a' | ab \rangle$ of the matrix elements $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. The top row of $\mathcal{M}_1$, equations (47.34a), and top three rows of $\mathcal{M}_2$, equations (47.34b), are interactions in which the spinor does not flip spin, while the bottom row of $\mathcal{M}_1$ and bottom three rows of $\mathcal{M}_2$ are interactions in which the spinor flips spin (the fourth and fifth rows of $\mathcal{M}_2$ in equation (47.34b) appear to be the same, but they become different when added to $\mathcal{M}_1$, equation (47.35)). As is evident from the expressions (47.34), for most spin choices only $\mathcal{M}_2$ is nonzero. The exceptions, where both $\mathcal{M}_1$ and $\mathcal{M}_2$ contribute, are (the top row of $\mathcal{M}_1$ with the top row of $\mathcal{M}_2$, and the second row of $\mathcal{M}_1$ with the fourth row of $\mathcal{M}_2$)

$$\left. \begin{array}{l} \langle \uparrow - | + \uparrow \rangle = \langle \downarrow + | - \downarrow \rangle \\ \langle \downarrow + | + \uparrow \rangle = -\langle \uparrow - | - \downarrow \rangle \end{array} \right\}_{1+2} = (ie)^2 \frac{2}{\cosh \eta + \sinh \eta \cos \theta} \left\{ \begin{array}{l} (e^\eta - e^{-\eta} \sin^2 \frac{\theta}{2}) \cos \frac{\theta}{2}, \\ e^{-2\eta} \sin^3 \frac{\theta}{2}. \end{array} \right. \quad (47.35)$$

As described in §35.2, the occupation numbers f of an ensemble of spin states are described by a complex Hermitian density matrix of rank equal to the number of spin eigenstates. Since photons have two spin eigenstates, and likewise spinors have two spin eigenstates, in both cases the density matrices f_c^a (photons) and f_d^b (spinors) are 2×2 matrices. The scattering rate, the number of scatterings per unit time, is obtained by summing/integrating the amplitude squared over spins and energy-momenta, appropriately weighted by density matrices, equation (47.24),

$$\frac{\text{number of events}}{\text{volume} \cdot \text{time}} = \int \mathcal{M}_{b'a'ab} \mathcal{M}^{d'cc'd'} (\delta_{d'}^{b'} - f_{d'}^{b'}) (\delta_{c'}^{a'} + f_{c'}^{a'}) f_c^a f_d^b \\ (2\pi)^4 \delta^4(p' + k' - p - k) \frac{d^3 p'}{2E'(2\pi)^3} \frac{d^3 k'}{2\omega'(2\pi)^3} \frac{d^3 k}{2\omega(2\pi)^3} \frac{d^3 p}{2E(2\pi)^3}, \quad (47.36)$$

with implicit summation over paired indices $b'a'ab d'c'c'd'$. The factors $\delta_{d'}^{b'} - f_{d'}^{b'}$ and $\delta_{c'}^{a'} + f_{c'}^{a'}$ are fermion-blocking and boson-stimulation factors. For fermions, scattering into a final state can occur only if the final state is empty; a fermion already in the final state blocks the reaction from occurring. For bosons, bosons already in the final state enhance the reaction rate. The factors f_c^a and f_d^b on the incoming spinor and photon states count the number of particles in those incoming states. If incoming and outgoing particles occupy a region of spacetime over which spin distributions are approximately uniform, as is often true in astrophysical

contexts, then the density matrices of incoming and outgoing particles of the same species are the same. If incoming and outgoing particles come from and go to different places, as is often true in particle accelerators, then the density matrices of incoming and outgoing particles of the same species may differ.

The raised indices on the amplitude $\mathcal{M}^{d'cc'd'}$ in equation (47.36) essentially signifies complex conjugation, a notation justified as follows. The complex conjugate of a spinor-vertex is given by equation (47.25), so the complex conjugate $\mathcal{M}^*$ of an amplitude may be written

$$\mathcal{M}_{d'c'cd}^* \equiv \mathcal{M}_{(p'd')(k'c')(kc)(pd)}^* = \mathcal{M}_{(\bar{p}\bar{d})(\bar{k}\bar{c})(\bar{k}'\bar{c}')(\bar{p}'\bar{d}')} , \quad (47.37)$$

which has conjugated momenta and spins in reverse order. In relativity, indices are lowered and raised with the metric and its inverse. In the present case, transverse photon indices are lowered and raised with the Newman-Penrose metric for circular polarizations, or the Minkowski metric for linear polarizations. In both cases lowering and raising flips the component to its conjugate, $f_{\bar{c}} = f^c$; thus $f_{\mp} = f^{\pm}$ for circular polarizations, or $f_c = f^c$ with $c = 1$ or 2 for linear polarizations. Similarly, spinor indices are lowered and raised with the spinor metric and its inverse, §41.6.4, which likewise flips components to their conjugates, modulo a sign, $f_d = \pm f^{\bar{d}}$. The antisymmetry of the spinor metric implies that there is a minus sign for each contraction of conjugated versus unconjugated spinors, $f_{\bar{b}}g^{\bar{b}} = -f^bg_b$, but in the present case the overall sign is positive because charge conservation implies that there is always an even number of combined incoming and outgoing spinors, therefore an even number of spinor contractions. The result is

$$\mathcal{M}_{d'c'cd}^* = \mathcal{M}_{(\bar{p}\bar{d})(\bar{k}\bar{c})(\bar{k}'\bar{c}')(\bar{p}'\bar{d}')} = \mathcal{M}^{(pd)(kc)(k'c')(p'd')} \equiv \mathcal{M}^{d'cc'd'} , \quad (47.38)$$

which is the conjugate matrix element in the integrand of equation (47.36). In §35.2 the photon density matrix was written $f^{b\bar{d}}$ with second index conjugated; the same density matrix is being written here as f_d^b .

A common situation is where there are equal numbers of particles in each of the 2 incoming spinor spins b ($\uparrow$ and $\downarrow$) and each of the 2 incoming photon spins a ($+$ and $-$). For example, this is true in thermodynamic equilibrium, where there are equal numbers of particles in orthogonal spin eigenstates, and the density matrix is the unit matrix. The density matrix in this case is equal to the unit matrix with respect to any choice of orthogonal spin eigenstates. The scattering amplitude squared averaged over incoming and summed over outgoing spins is the Klein and Nishina (1929) formula for Compton scattering,

$$\begin{aligned} \frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} &= \frac{1}{4} \sum_{b'a'ab} |\mathcal{M}_{b'a'ab}|^2 \\ &= 2e^4 \left[e^{-\eta} (\cosh \eta + \sinh \eta \cos \theta) + \frac{e^{\eta}}{\cosh \eta + \sinh \eta \cos \theta} - \frac{\sin^2 \theta}{(\cosh \eta + \sinh \eta \cos \theta)^2} \right] \\ &= 2e^4 \left(\frac{\omega'}{\omega} + \frac{\omega}{\omega'} - \sin^2 \tilde{\theta} \right) , \end{aligned} \quad (47.39)$$

where ω and ω' are the initial and final photon energies in the lab frame, equations (47.13), and $\tilde{\theta}$ is the photon scattering angle in the lab frame, equation (47.11). The factor of $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$ comes from averaging over incoming states, a factor of $\frac{1}{2}$ for each of the incoming spinor and photon. The averaging is conventional: the averaged amplitude squared represents a rate per incoming spinor and photon. In the end, the net rate of

scattering events must be summed over all incoming spinors and photons. If there are equal numbers of spins in each state, as here, then the rate of scatterings must be multiplied by the number $4 = 2 \times 2$ of incoming spin states, cancelling the averaging factor of $\frac{1}{4}$, before integrating momenta over occupation numbers.

In accordance with equation (47.54), the Klein-Nishina amplitude squared (47.39) is converted into a differential cross-section by multiplying by $1/(8\pi E_{\text{com}})^2$ where $E_{\text{com}} = me^\eta$ is the total centre-of-mass energy,

$$\frac{d\sigma}{d\omega} = 2 \left(\frac{e^2}{8\pi E_{\text{com}}} \right)^2 \left(\frac{\omega'}{\omega} + \frac{\omega}{\omega'} - \sin^2 \tilde{\theta} \right). \quad (47.40)$$

In the nonrelativistic limit $\eta \rightarrow 0$, the Klein-Nishina formula (47.39) reduces to the Thomson scattering formula

$$\frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = 2e^4 (1 + \cos^2 \theta), \quad (47.41)$$

which reproduces equation (33.55) when the Heaviside charge e is eliminated in favour of the fine-structure constant, $e^2 = 4\pi\alpha$. Note that in the nonrelativistic limit the photon scattering angle $\tilde{\theta}$ in the lab frame coincides with the scattering angle θ in the centre-of-mass frame. The corresponding Thomson differential cross-section is

$$\frac{d\sigma}{d\omega} = 2 \left(\frac{\alpha}{2m} \right)^2 (1 + \cos^2 \theta), \quad (47.42)$$

in agreement with equation (33.56).

47.6.1 Polarized spinor-photon scattering

Also relevant to cosmology, §35.10, is the photon-polarized cross-section for electron-photon scattering, illustrated in Figure 35.1. Collisions between charged particles drive charged particles into thermodynamic equilibrium, so it is a good assumption to take the density matrix for electrons to be the unit matrix. The matrix of photon-polarized amplitudes squared takes its simplest form with respect to a linearly polarized basis of transverse photon vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, with (as usual) the 1-direction (x -direction) lying in the scattering plane, and the 2-direction (y -direction) orthogonal to the scattering plane. Averaged over incoming and summed over outgoing electron spins b and b' , the matrix of squared amplitudes for Compton scattering between ingoing and outgoing photon polarizations is (the averaging factor here is $\frac{1}{2}$ because the spins only of incoming electrons, not of incoming photons, are being averaged over)

$$\frac{1}{2} \mathcal{M}_{b'a'ab} \mathcal{M}^{bcc'b'} = \frac{e^4}{\omega'\omega} \begin{pmatrix} (\omega' + \omega)^2 - 4\omega'\omega \sin^2 \tilde{\theta} & 0 & 0 & (\omega' - \omega)^2 \\ 0 & (\omega' + \omega)^2 \cos \tilde{\theta} & -(\omega' - \omega)^2 \cos \tilde{\theta} & 0 \\ 0 & -(\omega' - \omega)^2 \cos \tilde{\theta} & (\omega' + \omega)^2 \cos \tilde{\theta} & 0 \\ (\omega' - \omega)^2 & 0 & 0 & (\omega' + \omega)^2 \end{pmatrix}, \quad (47.43)$$

where again ω and ω' are the initial and final photon energies in the lab frame, equations (47.13), and $\tilde{\theta}$ is the photon scattering angle in the lab frame, equation (47.11). The ordering of rows ac and columns $a'c'$ here is 11, 12, 21, 22.

In the nonrelativistic limit, scattering leaves the photon energy unchanged, $\omega' = \omega$, and the matrix of photon-polarized Compton scattering amplitudes squared (47.43) reduces to the polarized Thomson scattering matrix

$$\frac{1}{2} \mathcal{M}_{b'a'ab} \mathcal{M}^{bcc'b'} = 4e^4 \begin{pmatrix} \cos^2 \theta & 0 & 0 & 0 \\ 0 & \cos \theta & 0 & 0 \\ 0 & 0 & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (47.44)$$

The relatively simple form of the scattering matrix (47.44) allows the polarized Thomson scattering amplitude squared between an incoming photon of polarization $\mathbf{a} = \cos \chi \boldsymbol{\gamma}_1 + e^{-i\phi} \sin \chi \boldsymbol{\gamma}_2$ and an outgoing photon of polarization $\mathbf{a}' = \cos \chi' \boldsymbol{\gamma}_1 + e^{-i\phi'} \sin \chi' \boldsymbol{\gamma}_2$ to be abbreviated to (no sum over $\mathbf{a}', \mathbf{a}$)

$$\frac{1}{2} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = 4e^4 |\bar{\mathbf{a}}' \cdot \mathbf{a}|^2 = 4e^4 |\cos \chi' \cos \chi + e^{-i(\phi'+\phi)} \sin \chi' \sin \chi|^2, \quad (47.45)$$

which reproduces equation (35.54).

47.7 Cross-section

The rate of collisions between two particles of species p and q with incoming energy-momenta p and q (the label serving to identify both the species and its energy-momentum), such as the Compton scattering rate (47.36), is commonly translated into a cross-section σ defined by

$$\frac{\text{number of events}}{\text{volume} \cdot \text{time}} = n_p n_q \sigma v, \quad (47.46)$$

where n_p and n_q are the total number densities of the incoming particle species, and v is the relative velocity of the colliding particles. The cross-section σ has units of area. The number density n of each species is the time component $n = n^0$ of a number current n^k , equation (10.120),

$$n_p = f_a^a \frac{d^3 p}{(2\pi)^3}, \quad n_q = f_b^b \frac{d^3 q}{(2\pi)^3}, \quad (47.47)$$

where f_a^a and f_b^b , the trace of the density matrix for each species, counts the total occupation number of spin states. For example, if there are equal numbers of particles in each of the two spin states, then $f_a^a = 2f$ where f is the occupation number of each state. The incoming number densities n , equations (47.47), should strictly be integrated over incoming momenta; but here for simplicity the momenta are being assumed to be distributed narrowly about some particular value. When there is a range of incoming momenta, the cross-section is defined to be an average over incoming momenta.

For a 2-body collision that yields 2 outgoing particles, such as Compton scattering described by the rate (47.36), the definition (47.46) of the cross-section σ implies that

$$d\sigma = \frac{1}{f_a^a f_b^b} \frac{1}{4E_q E_q v} \mathcal{M}_{b'a'ab} \mathcal{M}^{dcc'd'} \delta_{d'}^{b'} \delta_{c'}^{a'} f_c^a f_d^b (2\pi)^4 \delta^4(p' + q' - p - q) \frac{d^3 p'}{2E'_p (2\pi)^3} \frac{d^3 q'}{2E'_q (2\pi)^3}. \quad (47.48)$$

Equation (47.48) gives a differential cross-section $d\sigma$ into an interval do of angular directions, because there are 6 integrals over final momenta, of which 4 are absorbed by the 4-dimensional energy-momentum-conserving delta-function, leaving a 2-dimensional interval over angles, as will be seen in equation (47.52). The $1/(f_a^a f_b^b)$ factor is the conventional average-over-incoming-spins factor.

The only non-Lorentz-invariant factor in the expression (47.48) for the cross-section is $E_p E_q v$. However, this factor is Lorentz invariant with respect to boosts along, and rotations about, the direction of collision, the 3-direction (z -direction), as is evident from the fact that $E_p E_q v$ is the 03 component of the bivector $[\mathbf{p}, \mathbf{q}]$,

$$E_p E_q v = p^0 q^3 - p^3 q^0 = [\mathbf{p}, \mathbf{q}]^{03} , \quad (47.49)$$

and the bivector γ_{03} commutes with the generators γ_{03} and γ_{12} of Lorentz boosts along and spatial rotations about the 3-direction. Thus the cross-section σ is Lorentz invariant with respect to Lorentz boosts along and spatial rotations about the direction of the collision. Since $E_p E_q v$ is Lorentz invariant under boosts along the collision direction, it can be evaluated in any convenient frame. A convenient frame is the centre-of-mass frame, with respect to which

$$E_p E_q v = E_{\text{com}} p_{\text{com}} , \quad (47.50)$$

where p_{com} is the magnitude of the momentum of each particle in the centre-of-mass frame, and E_{com} is the total energy in the centre-of-mass frame,

$$p_{\text{com}} \equiv |p_a| = |q_a| , \quad E_{\text{com}} \equiv E_p + E_q \quad \text{in centre-of-mass frame} . \quad (47.51)$$

The delta-function and the integration elements over final momenta in equation (47.48) are Lorentz invariant, so they may be evaluated in any convenient frame. A convenient frame is again the centre-of-mass frame, where the total incoming and outgoing spatial momenta are zero, $p_a + q_a = p'_a + q'_a = 0$, and the magnitudes of each of the incoming and outgoing spatial momenta are all equal, $|p_a| = |q_a| = |p'_a| = |q'_a| = p_{\text{com}}$. The integral over final momenta then gives

$$\begin{aligned} \int (2\pi)^4 \delta^4(p' + q' - p - q) \frac{d^3 p'}{2E'_p (2\pi)^3} \frac{d^3 q'}{2E'_q (2\pi)^3} &= \frac{1}{16\pi^2} \int \delta(E'_p + E'_q - E_{\text{com}}) \frac{d^3 p'}{E'_p E'_q} \\ &= \frac{1}{16\pi^2} \frac{p'^2 dp' do}{E'_p E'_q d(E'_p + E'_q)} \Big|_{E'_p + E'_q = E_{\text{com}}} = \frac{1}{16\pi^2} \frac{p_{\text{com}}}{E_{\text{com}}} do . \end{aligned} \quad (47.52)$$

The final step of equations (47.52) follows because the centre-of-mass energies satisfy $dE'_p/dp' = p'/E'_p$ and $dE'_q/dp' = p'/E'_q$.

Inserting equations (47.50) and (47.52) into equation (47.48) gives the differential cross-section $d\sigma/do$

$$\frac{d\sigma}{do} = \frac{1}{f_a^a f_b^b} \frac{1}{(8\pi E_{\text{com}})^2} \mathcal{M}_{b'a'ab} \mathcal{M}^{dca'b'} f_c^a f_d^b , \quad (47.53)$$

where again the $1/(f_a^a f_b^b)$ factor is the conventional average-over-incoming-spins factor. Equation (47.53) says that for a 2-body collision with 2 incoming particles and 2 outgoing particles, the differential cross-section $d\sigma/do$ is obtained from the (appropriately weighted) amplitude squared by multiplying by $1/(8\pi E_{\text{com}})^2$,

where E_{com} is the total centre-of-mass collision energy, equation (47.51). If there are equal numbers of particles in each incoming state, then the differential cross-section reduces to

$$\frac{d\sigma}{do} = \frac{1}{f_a^a f_b^b} \frac{1}{(8\pi E_{\text{com}})^2} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} . \quad (47.54)$$

47.8 Spinor-antispinor creation from and annihilation into photons

A process closely related to spinor-photon scattering is the creation of a spinor and antispinor from the collision of energetic photons, or the inverse process of the annihilation of a spinor and an antispinor into photons. Spinor-antispinor creation and annihilation, to lowest nonvanishing order of perturbation theory, is illustrated in Figure 47.6. The diagrams 47.6 are the same as those for spinor-photon scattering, Figure 47.2, with flips in the direction of time for one each of the spinors and photons.

Spinor-antispinor creation and annihilation, illustrated in Figure 47.6, is described by expressions which look the same as the spinor-photon scattering interactions (47.6) and (47.7), except that $k \rightarrow -k$ and $p' \rightarrow -p'$, and the antispinor is accommodated by the rule (44.84), which in the present case requires the replacement $\bar{\epsilon}_{p'b'} \rightarrow -i\epsilon_{p'b'}$,

$$(ie(-i\epsilon_{p'b'}) \cdot \overline{\gamma_{k'a'} \epsilon_{qc}})(ie \bar{\epsilon}_{qc} \cdot \gamma_{ka} \epsilon_{pb}) e^{i(p'-k'+q) \cdot x'} e^{i(-q-k+p) \cdot x} , \quad (47.55a)$$

$$(ie(-i\epsilon_{p'b'}) \cdot \overline{\gamma_{ka} \epsilon_{qc}})(ie \bar{\epsilon}_{qc} \cdot \gamma_{k'a'} \epsilon_{pb}) e^{i(p'-k+q) \cdot x'} e^{i(-q-k'+p) \cdot x} . \quad (47.55b)$$

The same expressions hold for either creation or annihilation because, as remarked in the penultimate paragraph of §47.2, spinor-photon vertices are invariant under reordering, and an expression with all vertices reordered can be interpreted as the same process going backwards in time.

Integrating the two two-vertex interactions (47.55) over vertices x and x' enforces conservation of energy-momentum at each of the two vertices. The energy-momentum q of the intermediate spinor is then related to the energy-momenta of the incoming and outgoing spinors and photons by, for the two cases,

$$p - k = q = -p' + k' , \quad (47.56a)$$

$$p - k' = q = -p' + k , \quad (47.56b)$$

which differ from the spinor-photon scattering equations (47.8) by $k \rightarrow -k$ and $p' \rightarrow -p'$. As with equations (47.8), the sign of the energy-momentum q of the intermediate spinor in equations (47.56) assumes that x' is to the future of x (that is, $t' - t \geq 0$). If instead x' is to the past of x , then the sign of q must be flipped. When the Wick-bracketed pair in the interactions (47.55) are appropriately summed and integrated over the intermediate spin c and momentum q , the result is, as in the case of spinor-photon scattering, the Dirac propagator (44.81).

The energy-momenta of the colliding particles in their centre-of-mass frame, as illustrated in the left panels

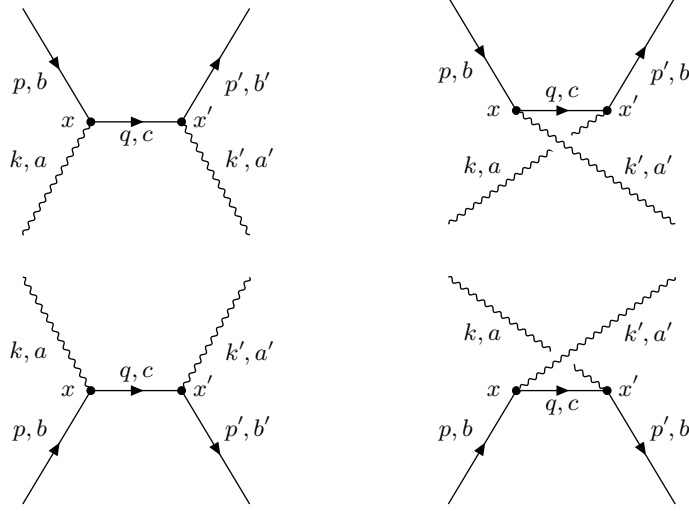


Figure 47.6 (Top) Two photons collide to create a spinor-antispinor pair. (Bottom) Spinor-antispinor annihilation into two photons. The bottom diagrams are the same as the top diagrams flipped vertically. The arrows on the spinor lines represent the direction of charge.

of Figure 47.7, are

$$p = m\{\cosh \eta, 0, 0, -\sinh \eta\} , \quad (47.57a)$$

$$p' = m\{\cosh \eta, 0, 0, \sinh \eta\} , \quad (47.57b)$$

$$k = m \cosh \eta \{1, \sin \theta, 0, \cos \theta\} , \quad (47.57c)$$

$$k' = m \cosh \eta \{1, -\sin \theta, 0, -\cos \theta\} . \quad (47.57d)$$

Boosted into the rest frame of the spinor ϵ_{pb} , as illustrated in the right panels of Figure 47.7, the spinor and photon energy-momenta are

$$p = m\{1, 0, 0, 0\} , \quad (47.58a)$$

$$p' = m\{\cosh 2\eta, 0, 0, \sinh 2\eta\} , \quad (47.58b)$$

$$k = m \cosh \eta \{\cosh \eta + \sinh \eta \cos \theta, \sin \theta, 0, \sinh \eta + \cosh \eta \cos \theta\} , \quad (47.58c)$$

$$k' = m \cosh \eta \{\cosh \eta - \sinh \eta \cos \theta, -\sin \theta, 0, \sinh \eta - \cosh \eta \cos \theta\} . \quad (47.58d)$$

The angle $\tilde{\theta}$ between the photon directions in the lab frame, right panels of Figure 47.7, satisfies, from equations (47.58c) and (47.58d),

$$\sin \tilde{\theta} = \frac{2 \sinh \eta \sin \theta}{\cosh^2 \eta - \sinh^2 \eta \cos^2 \theta} , \quad \cos \tilde{\theta} = \frac{1 - \sinh^2 \eta \sin^2 \theta}{\cosh^2 \eta - \sinh^2 \eta \cos^2 \theta} . \quad (47.59)$$

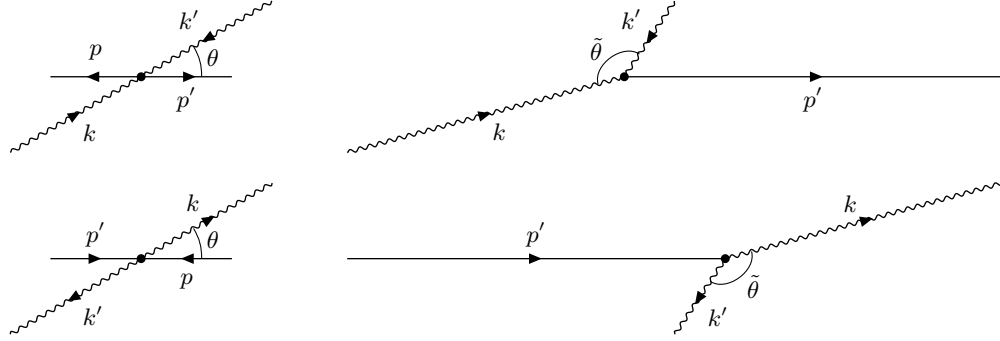


Figure 47.7 Spinor-antispinor (top) creation, and (bottom) annihilation, (left) in the centre-of-mass frame, where $p_a + p'_a = k_a + k'_a = 0$, and (right) in the rest frame of the spinor, where $p_a = 0$. In the centre-of-mass frame, the angle between the incoming and outgoing pairs is θ . In the rest frame of the incoming p spinor, the photon is deflected by angle $\tilde{\theta}$, equation (47.59). The lengths and directions of the lines are proportional to the spatial momenta. The parameters of the collision in this instance are $\eta = \text{atanh } \frac{3}{5}$ and $\theta = \frac{\pi}{6}$.

The intermediate spinor energy-momentum q for the two cases, equations (47.56), then satisfy,

$$q^2 + m^2 = (p - k)^2 + m^2 = -2p \cdot k = 2m\omega, \quad (47.60a)$$

$$q^2 + m^2 = (p - k')^2 + m^2 = -2p \cdot k' = 2m\omega', \quad (47.60b)$$

where ω and ω' are the initial and final photon energies in the lab frame, from equations (47.58),

$$\omega = m \cosh \eta (\cosh \eta + \sinh \eta \cos \theta), \quad \omega' = m \cosh \eta (\cosh \eta - \sinh \eta \cos \theta). \quad (47.61)$$

The amplitude $\mathcal{M}$ is a scalar, so can be evaluated in any convenient frame of reference. A convenient frame is the centre-of-mass frame, illustrated in the left panel of Figure 47.7, with the outgoing (creation) or incoming (annihilation) spinors directed along the 3-direction (z -direction), and the incoming or outgoing photons lying in the 31-plane (zx -plane). The energy-momenta of the incoming and outgoing spinors and photons in this centre-of-mass frame are given by equations (47.57). In the centre-of-mass frame, the outgoing or ingoing spinors are all boosted by rapidity η out of their rest frames. Relative to the direction of the outgoing (creation) or incoming (annihilation) spinor momentum p , the spinor momentum p' is rotated by π , while the photon momenta k and k' rotated by respectively $\theta + \pi$ and θ . In the centre-of-mass frame, the energy-momenta $\overset{1}{q}$ and $\overset{2}{q}$ of the intermediate spinor in the first and second cases are, equations (47.56),

$$\overset{1}{q} = p - k = -m\{0, \cosh \eta \sin \theta, 0, \sinh \eta + \cosh \eta \cos \theta\}, \quad (47.62a)$$

$$\overset{2}{q} = p - k' = m\{0, \cosh \eta \sin \theta, 0, -\sinh \eta + \cosh \eta \cos \theta\}. \quad (47.62b)$$

The energy-momenta $\overset{1}{q}$ and $\overset{2}{q}$ are related by

$$\overset{2}{q}(\theta) = \overset{1}{q}(\theta + \pi) = \overset{1}{q}(\theta - \pi). \quad (47.63)$$

The amplitudes in terms of Lorentz-transformed basis spinors and vectors are then

$$\mathcal{M}^1 = (ie)^2 \frac{(2m)(-i\epsilon_{\top b'} \cdot \tilde{R}_\eta)(R_{\theta+\pi}\gamma_{a'}\tilde{R}_{\theta+\pi})(i\mathbf{q}+m)(R_\theta\gamma_a\tilde{R}_\theta)(R_\pi R_\eta\epsilon_{\top b})}{q^2+m^2} \Big|_{q=\frac{1}{q}}, \quad (47.64a)$$

$$\mathcal{M}^2 = (ie)^2 \frac{(2m)(-i\epsilon_{\top b'} \cdot \tilde{R}_\eta)(R_\theta\gamma_a\tilde{R}_\theta)(i\mathbf{q}+m)(R_{\theta+\pi}\gamma_{a'}\tilde{R}_{\theta+\pi})(R_\pi R_\eta\epsilon_{\top b})}{q^2+m^2} \Big|_{q=\frac{2}{q}}, \quad (47.64b)$$

where q in the two cases are given by equations (47.62), and the boost and spatial rotation rotors R_η and R_θ are defined by equations (47.31). The amplitudes $\mathcal{M}_{b'a'ab}^1$ and $\mathcal{M}_{b'a'ab}^2$ given by equations (47.64) are related by $\theta \rightarrow \theta + \pi$ (or $\theta \rightarrow \theta - \pi$), and an exchange $a' \leftrightarrow a$ of photon spins,

$$\mathcal{M}_{b'a'ab}^2(\theta) = \mathcal{M}_{b'aa'b}^1(\theta + \pi). \quad (47.65)$$

The amplitudes (47.64) evaluate to, for the various outgoing and incoming spins,

$$\left. \begin{aligned} \langle \uparrow + | + \uparrow \rangle &= -\langle \downarrow - | - \downarrow \rangle \\ \langle \uparrow - | - \uparrow \rangle &= -\langle \downarrow + | + \downarrow \rangle \\ \langle \uparrow - | + \uparrow \rangle &= \langle \uparrow + | - \uparrow \rangle = -\langle \downarrow + | - \downarrow \rangle = -\langle \downarrow - | + \downarrow \rangle \\ \langle \downarrow - | + \uparrow \rangle &= -\langle \uparrow + | - \downarrow \rangle \\ \langle \downarrow + | - \uparrow \rangle &= -\langle \uparrow - | + \downarrow \rangle \\ \langle \downarrow + | + \uparrow \rangle &= \langle \downarrow - | - \uparrow \rangle = -\langle \uparrow - | - \downarrow \rangle = -\langle \uparrow + | + \downarrow \rangle \end{aligned} \right\}_1 = (ie)^2 \frac{1}{\kappa_+} \left\{ \begin{aligned} &\tanh \eta \sin^2 \theta - 2 \cos^2 \frac{\theta}{2}, \\ &\tanh \eta \sin^2 \theta + 2 \sin^2 \frac{\theta}{2}, \\ &\tanh \eta \sin^2 \theta, \\ &-2 \sinh \eta \sin \theta \cos^2 \frac{\theta}{2}, \\ &2 \sinh \eta \sin \theta \sin^2 \frac{\theta}{2}, \\ &-\kappa_+ \sin \theta, \end{aligned} \right. \quad (47.66a)$$

$$\left. \begin{aligned} \langle \uparrow + | + \uparrow \rangle &= -\langle \uparrow - | - \downarrow \rangle \\ \langle \uparrow - | - \uparrow \rangle &= -\langle \uparrow + | + \downarrow \rangle \\ \langle \uparrow - | + \uparrow \rangle &= \langle \uparrow + | - \uparrow \rangle = -\langle \uparrow + | - \downarrow \rangle = -\langle \uparrow - | + \downarrow \rangle \\ \langle \uparrow + | - \uparrow \rangle &= \langle \uparrow - | + \downarrow \rangle \\ \langle \uparrow + | + \uparrow \rangle &= \langle \uparrow - | - \uparrow \rangle = \langle \uparrow - | - \downarrow \rangle = \langle \uparrow + | + \downarrow \rangle \end{aligned} \right\}_2 = (ie)^2 \frac{1}{\kappa_-} \left\{ \begin{aligned} &\tanh \eta \sin^2 \theta - 2 \sin^2 \frac{\theta}{2}, \\ &\tanh \eta \sin^2 \theta + 2 \cos^2 \frac{\theta}{2}, \\ &\tanh \eta \sin^2 \theta, \\ &-2 \sinh \eta \sin \theta \cos^2 \frac{\theta}{2}, \\ &2 \sinh \eta \sin \theta \sin^2 \frac{\theta}{2}, \\ &\kappa_- \sin \theta, \end{aligned} \right. \quad (47.66b)$$

where $\kappa_\pm$ are defined by

$$\kappa_\pm \equiv \cosh \eta \pm \sinh \eta \cos \theta. \quad (47.67)$$

Equations (47.66a) and (47.66b) give the components $\mathcal{M}_{b'a'ab} \equiv \langle b'a' | ab \rangle$ of the matrix elements $\mathcal{M}^1$ and $\mathcal{M}^2$ respectively. The top three rows of each of $\mathcal{M}^1$ and $\mathcal{M}^2$ are interactions in which the spinor does not flip spin, while the bottom three rows of each of $\mathcal{M}^1$ and $\mathcal{M}^2$ are interactions in which the spinor flips spin.

The amplitude $\mathcal{M}$ and its photon-exchanged partner amplitude $\mathcal{M}$, equations (47.66) sum to the combined amplitude $\mathcal{M}$,

$$\mathcal{M}_{b'a'ab}(\theta) = \mathcal{M}_{b'a'ab}^1(\theta) + \mathcal{M}_{b'a'ab}^2(\theta) = \mathcal{M}_{b'a'ab}^1(\theta) + \mathcal{M}_{b'aa'b}^1(\theta + \pi), \quad (47.68)$$

which is symmetric under exchange of photon momenta, that is $\theta \rightarrow \theta + \pi$, and exchange $a' \leftrightarrow a$ of photon spins. Equation (47.68) expresses the indistinguishability of the incoming (creation) or outgoing (annihilation) photons: the amplitude is invariant under exchange of the two photons. For the various choices of incoming

and outgoing spins, the combined amplitude $\langle b'a'|ab \rangle \equiv \mathcal{M}_{b'a'ab}$ is

$$\left. \begin{aligned} \langle \uparrow + | \uparrow + \rangle &= -\langle \downarrow - | \downarrow - \rangle \\ \langle \uparrow - | \uparrow - \rangle &= -\langle \downarrow + | \downarrow + \rangle \\ \langle \uparrow - | \uparrow + \rangle &= \langle \uparrow + | \uparrow - \rangle = -\langle \downarrow + | \downarrow - \rangle = -\langle \downarrow - | \downarrow + \rangle \\ \langle \downarrow - | \uparrow + \rangle &= \langle \uparrow + | \downarrow - \rangle \\ \langle \downarrow + | \uparrow - \rangle &= \langle \uparrow - | \downarrow + \rangle \\ \langle \downarrow + | \uparrow + \rangle &= \langle \downarrow - | \uparrow - \rangle = \langle \uparrow - | \downarrow - \rangle = \langle \uparrow + | \downarrow + \rangle \end{aligned} \right\}_{1+2} = (ie)^2 \frac{1}{\kappa_+ \kappa_-} \begin{cases} -2e^{-\eta}, \\ 2e^{\eta}, \\ 2 \sinh \eta \sin^2 \theta, \\ -2 \sinh 2\eta \sin \theta \cos^2 \frac{\theta}{2}, \\ 2 \sinh 2\eta \sin \theta \sin^2 \frac{\theta}{2}, \\ 0. \end{cases} \quad (47.69)$$

Squared amplitudes for each choice of incoming and outgoing spins follow from squaring the amplitudes (47.69). In the situation where there are equal numbers of particles in each of the two incoming spins states, the amplitude squared for spinor-antispinor creation or annihilation averaged over initial spins and summed over final spins is

$$\begin{aligned} \frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} &= 4e^4 \left[\frac{\cosh^2 \eta + \sinh^2 \eta \cos^2 \theta}{\cosh^2 \eta - \sinh^2 \eta \cos^2 \theta} + 2 \left(\frac{\sinh \eta \sin \theta}{\cosh^2 \eta - \sinh^2 \eta \cos^2 \theta} \right)^2 \right] \\ &= 2e^4 \left(\frac{\omega'}{\omega} + \frac{\omega}{\omega'} + \sin^2 \tilde{\theta} \right), \end{aligned} \quad (47.70)$$

where ω and ω' are the initial and final photon energies in the lab frame, equations (47.61), and $\tilde{\theta}$ is the angle between the incoming (creation) or outgoing (annihilation) photons in the lab frame, equations (47.59), as illustrated in Figure 47.7. In accordance with equation (47.54), the amplitude squared (47.70) is converted into a differential cross-section by multiplying by $1/(8\pi E_{\text{com}})^2$ where $E_{\text{com}} = 2m \cosh \eta$ is the total centre-of-mass energy,

$$\frac{d\sigma}{d\Omega} = 2 \left(\frac{e^2}{8\pi E_{\text{com}}} \right)^2 \left(\frac{\omega'}{\omega} + \frac{\omega}{\omega'} - \sin^2 \tilde{\theta} \right). \quad (47.71)$$

47.9 Coulomb scattering

The scattering of electrically charged particles by each other is called Coulomb scattering. Coulomb scattering is mediated by the electromagnetic field. To lowest order in perturbation theory, Coulomb scattering of spinors is a two-vertex process described by the Feynman diagrams 47.8. Section 47.9.1 considers the scattering of distinguishable spinors, for example electrons scattering off muons, in which case only the Feynman diagram on the left contributes. Section 47.9.2 then deals with the case of indistinguishable spinors, for example electrons scattering off electrons, for which both diagrams in Figure 47.8 contribute.

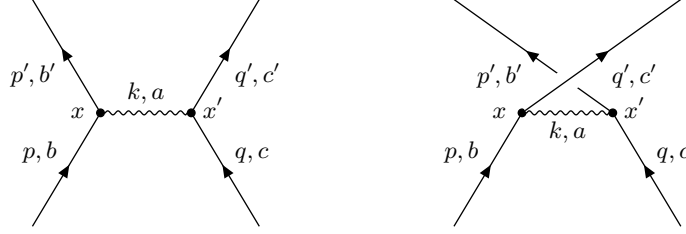


Figure 47.8 Feynman diagrams for Coulomb scattering of spinors. If the particles are distinguishable, for example electrons scattering off muons, then only the diagram on the left contributes. If on the other hand the particles are indistinguishable, for example electrons scattering off electrons, then both diagrams contribute.

47.9.1 Coulomb scattering of distinguishable particles

For distinguishable spinors p and q of charges e_p and e_q and masses m_p and m_q , the interaction described by the left Feynman diagram of Figure 47.8 is

$$(ie_q \bar{\epsilon}_{q'c'} \cdot \overbrace{\gamma_{ka} \epsilon_{qc}})(ie_p \bar{\epsilon}_{p'b'} \cdot \gamma_{ka} \epsilon_{pb}) e^{i(-q'+k+q) \cdot x'} e^{i(-p'-k+p) \cdot x} . \quad (47.72)$$

The interaction (47.72) must be integrated over all unspecified quantities, namely over the two vertices x and x' at which the intermediate photon interacts with respectively the p and q spinors, and over the momentum k and spin a of the intermediate photon.

Integration over vertices x and x' imposes conservation of energy-momentum at each of the two vertices, implying that the energy-momentum k of the intermediate photon is

$$p - p' = k = -q + q' . \quad (47.73)$$

Equation (47.73) assumes that x' is to the future of x . If instead x' is to the past of x , then the sign of k in equation (47.73) should be flipped.

The sum over spins a of the intermediate photon in the interaction (47.72) must be accomplished in a Lorentz-invariant fashion. In the case of spinor-photon scattering, the photons were on-shell, external legs of the Feynman diagrams 47.2, and the photon spins were transverse. For Coulomb scattering on the other hand, the photons are off-shell, internal lines of the Feynman diagrams 47.8, and the photon spins are not necessarily transverse. The photon vectors γ_{ka} in the interaction (47.72) occur sandwiched between spinors in separate vertex terms. Each vertex term defines an electric current, a current j_p^m for the p spinor, and a current j_q^m for the q spinor,

$$j_p^m \equiv ie_p \bar{\epsilon}_{p'b'} \cdot \gamma^m \epsilon_{pb} , \quad (47.74a)$$

$$j_q^m \equiv ie_q \bar{\epsilon}_{q'c'} \cdot \gamma^m \epsilon_{qc} . \quad (47.74b)$$

As discussed in §45.1, a Lorentz-invariant treatment of the electromagnetic field requires extending the spin index a to run over all 3 spatial indices, not just the 2 transverse indices; and as discussed in §45.2, Lorentz

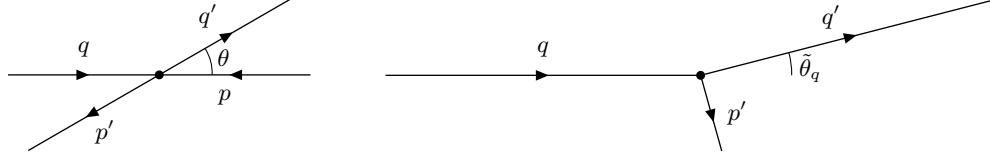


Figure 47.9 Coulomb scattering of spinors (left) in the centre-of-mass frame, and (right) in the rest frame of the incoming p spinor. The lengths and directions of the lines are proportional to the spatial momenta. In the centre-of-mass frame, the two spinors are each deflected by angle θ . In the rest frame of the incoming p spinor, the q spinor is deflected by angle $\tilde{\theta}_q$, equation (47.79). The spinors in this instance have been taken to have the same mass, $m_q = m_p$, and the parameters of the collision are $\eta = \text{atanh } \frac{7}{25}$ and $\theta = \frac{\pi}{6}$.

invariance requires the electromagnetic propagator to be proportional to the Minkowski metric (45.11), possibly modified by the subtraction of an arbitrary longitudinal tensor, equation (45.12). In the present case the Minkowski metric $\eta_{m'm}$ suffices to link the currents j_q^m and j_p^m in a Lorentz-invariant fashion, yielding the sum $\eta_{m'm} j_q^{m'} j_p^m$.

The photon momentum index k has been dropped from the currents (47.74) because it is the same for both currents, and is determined by energy-momentum conservation (47.73), and the only remaining dependence on k in the interaction (47.72) is on an exponential factor $e^{i(x'-x) \cdot k}$. The exponential factor depends only on the energy-momentum k of the intermediate photon, not on the spins or energy-momenta of the q and p spinors, nor on the spin of the photon. Integrating the exponential factor over photon momenta d^3k in a Lorentz-invariant fashion yields the Klein-Gordon propagator $D(x' - x)$, equation (42.102), in the massless limit since photons are massless. When this propagator is multiplied by the Minkowski metric $\eta_{m'm}$ that links the photon vectors $\gamma^{m'}$ and γ^m , the result is the photon propagator $D_{m'm}(x' - x)$, equation (45.13).

Integrating the interaction (47.72) over vertices x and x' , and summing/integrating over the energy-momentum k and spin of the intermediate photon, yields the amplitude for Coulomb scattering of distinguishable spinors,

$$\mathcal{M} = \int \frac{\eta_{m'm} j_q^{m'} j_p^m}{k^2 - i\epsilon} e^{i(-q'+k+q) \cdot x'} e^{i(-p'-k+p) \cdot x} d^4x' d^4x \frac{d^4k}{(2\pi)^4}. \quad (47.75)$$

The Coulomb amplitude (47.75) is Lorentz invariant, so can be evaluated in any convenient frame of reference. A convenient frame is the centre-of-mass frame of the colliding spinors, illustrated in the left panel of Figure 47.9. In the centre-of-mass frame, the energy-momenta p and p' of the incoming and outgoing p spinor of mass m_p , and q and q' of the incoming and outgoing q spinor of mass m_q , are

$$p = m_p \{ \cosh \eta_p, 0, 0, -\sinh \eta_p \}, \quad (47.76a)$$

$$p' = m_p \{ \cosh \eta_p, -\sinh \eta_p \sin \theta, 0, -\sinh \eta_p \cos \theta \}, \quad (47.76b)$$

$$q = m_q \{ \cosh \eta_q, 0, 0, \sinh \eta_q \}, \quad (47.76c)$$

$$q' = m_q \{ \cosh \eta_q, \sinh \eta_q \sin \theta, 0, \sinh \eta_q \cos \theta \}, \quad (47.76d)$$

with rapidities η_p and η_q related by

$$m_p \sinh \eta_p = m_q \sinh \eta_q = p_{\text{com}} , \quad (47.77)$$

where p_{com} is the magnitude of the momentum of each particle in the centre-of-mass frame, equation (47.51). Lorentz boosting the centre-of-mass energy-momenta (47.76) by rapidity η_p yields energy-momenta in the rest frame of the incoming p spinor,

$$p = m_p \{1, 0, 0, 0\} , \quad (47.78a)$$

$$p' = m_p \{ \cosh^2 \eta_p - \sinh^2 \eta_p \cos \theta, -\sinh \eta_p \sin \theta, 0, \cosh \eta_p \sinh \eta_p (1 - \cos \theta) \} , \quad (47.78b)$$

$$q = m_q \{ \cosh(\eta_p + \eta_q), 0, 0, \sinh(\eta_p + \eta_q) \} , \quad (47.78c)$$

$$q' = m_q \{ \cosh \eta_p \cosh \eta_q + \sinh \eta_p \sinh \eta_q \cos \theta, \sinh \eta_q \sin \theta, 0, \sinh \eta_p \cosh \eta_q + \cosh \eta_p \sinh \eta_q \cos \theta \} . \quad (47.78d)$$

The scattering angle $\tilde{\theta}_q$ of the q spinor in the rest frame of the incoming p spinor, illustrated in Figure 47.9, is, from equation (47.78d),

$$\tan \tilde{\theta}_q = \frac{\sinh \eta_q \sin \theta}{\sinh \eta_p \cosh \eta_q + \cosh \eta_p \sinh \eta_q \cos \theta} . \quad (47.79)$$

The energy-momentum k of the intermediate photon in the centre-of-mass frame is

$$k = p - p' = 2p_{\text{com}} \sin \frac{\theta}{2} \{0, \cos \frac{\theta}{2}, 0, -\sin \frac{\theta}{2}\} , \quad (47.80)$$

which is spacelike.

In the centre-of-mass frame, the ingoing and outgoing spinors p and q are boosted respectively by rapidity η_p and η_q out of their rest frames. Relative to the direction of the incoming spinor momentum q (chosen along the positive 3-direction), the outgoing photon momentum q' is rotated by θ , while the incoming and outgoing spinor momenta p and p' are rotated by respectively π and $\theta + \pi$. The electric currents j_p^m and j_q^m , equations (47.74), in spinors p and q are then

$$j_p^m = ie_p (2m_p) (\bar{\epsilon}_{\uparrow b'} \cdot \tilde{R}_{\eta_p} \tilde{R}_{\theta+\pi}) (\gamma^m) (R_\pi R_{\eta_p} \epsilon_{\uparrow b}) , \quad (47.81a)$$

$$j_q^m = ie_q (2m_q) (\bar{\epsilon}_{\uparrow c'} \cdot \tilde{R}_{\eta_q} \tilde{R}_\theta) (\gamma^m) (R_{\eta_q} \epsilon_{\uparrow c}) , \quad (47.81b)$$

where the boost and spatial rotation rotors R_η and R_θ are defined by equations (47.31). As previously with spinor-photon scattering, equations (47.34), a natural choice of spin eigenvectors of spinors is up ($\uparrow$) and down ($\downarrow$) in a frame aligned with the direction of motion of the spinors. Since the momenta of the q and p spinors point in opposite directions, their spin-up directions $\uparrow$ point in opposite directions, in the centre-of-mass frame. And since the momenta of the outgoing spinors are rotated by angle θ from the incoming spinors, the spin directions of the outgoing spinors are similarly rotated by angle θ from the incoming spinors. For each set of spinor spins, the spinor currents $\langle b'|b \rangle \equiv j_p^m$ and $\langle c'|c \rangle \equiv j_q^m$, equations (47.81), evaluate to

$$\langle \uparrow | \uparrow \rangle^m = \langle \downarrow | \downarrow \rangle^m = 2em \{ \cosh \eta \cos \frac{\theta}{2}, \begin{pmatrix} +q \\ -p \end{pmatrix} \sinh \eta \sin \frac{\theta}{2}, \begin{pmatrix} +\uparrow \\ -\downarrow \end{pmatrix} i \sinh \eta \sin \frac{\theta}{2}, \begin{pmatrix} +q \\ -p \end{pmatrix} \sinh \eta \cos \frac{\theta}{2} \} , \quad (47.82a)$$

$$\langle \downarrow | \uparrow \rangle^m = \langle \uparrow | \downarrow \rangle^m = 2em \{ \begin{pmatrix} +q \\ -p \end{pmatrix} \sin \frac{\theta}{2}, 0, 0, 0 \} , \quad (47.82b)$$

where e , m , and η denote e_p , m_p , and η_p for spinor p , and e_q , m_q and η_q for spinor q , and $\left(\begin{smallmatrix} + \\ - \end{smallmatrix} q\right)$ is $+$ for spinor q and $-$ for spinor p , and $\left(\begin{smallmatrix} + \\ - \end{smallmatrix} \uparrow\right)$ is $+$ for $\uparrow$ and $-$ for $\downarrow$. Each current (47.82) satisfies the Ward-Takahashi identities (47.148),

$$k_m j_p^m = k_m j_q^m = 0 , \quad (47.83)$$

where the photon energy-momentum k is given by equation (47.80). Inserting the currents (47.82) into the amplitude (47.75), and integrating over vertices x and x' and the intermediate photon momentum k , yields the amplitudes for Coulomb scattering of distinguishable spinors for the various choices of incoming c, b and outgoing c', b' spins,

$$\langle c'b'|cb \rangle \equiv \mathcal{M}_{c'cb'b} = \frac{\eta_{m'm} j_q^{m'} j_p^m}{k^2} = \frac{\eta_{m'm} \langle q'c'|qc \rangle^{m'} \langle p'b'|pb \rangle^m}{k^2} . \quad (47.84)$$

The amplitudes $\langle c'b'|cb \rangle$ for each set of incoming and outgoing spins are

$$\left. \begin{aligned} \langle \uparrow\uparrow|\uparrow\uparrow \rangle &= \langle \downarrow\downarrow|\downarrow\downarrow \rangle \\ \langle \downarrow\uparrow|\downarrow\uparrow \rangle &= \langle \uparrow\downarrow|\uparrow\downarrow \rangle \\ -\langle \downarrow\uparrow|\uparrow\uparrow \rangle &= \langle \uparrow\downarrow|\downarrow\downarrow \rangle = -\langle \downarrow\downarrow|\uparrow\uparrow \rangle = \langle \uparrow\uparrow|\downarrow\downarrow \rangle \\ -\langle \uparrow\downarrow|\uparrow\uparrow \rangle &= -\langle \downarrow\downarrow|\downarrow\downarrow \rangle = \langle \uparrow\uparrow|\uparrow\downarrow \rangle = \langle \downarrow\downarrow|\uparrow\downarrow \rangle \\ \langle \downarrow\downarrow|\uparrow\uparrow \rangle &= -\langle \uparrow\downarrow|\downarrow\downarrow \rangle = -\langle \downarrow\uparrow|\uparrow\downarrow \rangle = \langle \uparrow\uparrow|\downarrow\downarrow \rangle \end{aligned} \right\} = \frac{(ie_p)(ie_q)}{\kappa^2} \left\{ \begin{aligned} &\cosh(\eta_q + \eta_p) \cos^2 \frac{\theta}{2} + 2\kappa^2 , \\ &\cosh(\eta_q + \eta_p) \cos^2 \frac{\theta}{2} , \\ &\frac{1}{2} \cosh \eta_p \sin \theta , \\ &\frac{1}{2} \cosh \eta_q \sin \theta , \\ &\sin^2 \frac{\theta}{2} , \end{aligned} \right. \quad (47.85)$$

where

$$\kappa^2 \equiv \frac{k^2}{4m_q m_p} = \sinh \eta_q \sinh \eta_p \sin^2 \frac{\theta}{2} . \quad (47.86)$$

The top two lines of equations (47.85) are for scatterings that do not flip spin: in the top line, all spins are aligned (aligned here means p and q both $\uparrow$ or both $\downarrow$, but recall that spin-up and spin-down are with respect to the spinor direction of motion, so because the p and q spinors point in opposite directions, left panel of Figure 47.9, “aligned” actually means the p and q spins point in opposite directions), while in the second line the q and p spins are opposed. The next two lines of equations (47.85) are for scatterings where one of the q or p spins flips: in the third line the q spin flips, while in the fourth line the p spin flips. The last, fifth line of equations (47.85) is for scatterings where both q and p spins flip.

Squared amplitudes for each choice of incoming and outgoing spins follow from squaring the amplitudes (47.85). For equal numbers of spinors in each of the incoming spin states, the amplitude squared for Coulomb scattering of distinguishable spinors, averaged over initial spins and summed over final spins, is

$$\frac{1}{4} \mathcal{M}_{c'cb'b} \mathcal{M}^{bb'cc'} = e_p^2 e_q^2 \left(\frac{1 + \sinh^2(\eta_p + \eta_q) \cos^2 \frac{\theta}{2}}{\sinh^2 \eta_p \sinh^2 \eta_q \sin^4 \frac{\theta}{2}} + 2 \right) . \quad (47.87)$$

The differential cross-section is obtained by dividing by $(8\pi E_{\text{com}})^2$, equation (47.54). In the present case the centre-of-mass energy E_{com} satisfies

$$E_{\text{com}} \sinh \eta_p \sinh \eta_q = p_{\text{com}} \sinh(\eta_p + \eta_q) . \quad (47.88)$$

The resulting Coulomb scattering cross-section between spinors of charges $e_p = Z_p e$ and $e_q = Z_q e$ is

$$\frac{d\sigma}{do} = \left(\frac{Z_p Z_q \alpha}{2p_{\text{com}} \sinh(\eta_p + \eta_q) \sin^2 \frac{\theta}{2}} \right)^2 (1 + \sinh^2(\eta_p + \eta_q) \cos^2 \frac{\theta}{2} + 2 \sinh^2 \eta_p \sinh^2 \eta_q \sin^4 \frac{\theta}{2}) . \quad (47.89)$$

In the nonrelativistic limit, the rapidities of the p and q spinors approximate their velocities, $\eta_p \rightarrow v_p$ and $\eta_q \rightarrow v_q$, and the spin-averaged/summed amplitude squared (47.87) reduces to

$$\frac{1}{4} \mathcal{M}_{c'cb'b} \mathcal{M}^{bb'cc'} \rightarrow \left(\frac{e_p e_q}{\kappa^2} \right)^2 \rightarrow \left(\frac{e_p e_q}{v_p v_q \sin^2 \frac{\theta}{2}} \right)^2 , \quad (47.90)$$

where κ^2 is given by equation (47.86). The cross-section is obtained by dividing by $(8\pi E_{\text{com}})^2$, which equals $(8\pi(m_p + m_q))^2$ in the nonrelativistic limit. It follows from $m_p v_p = m_q v_q$ that

$$(m_p + m_q) v_p v_q = \mu v^2 , \quad (47.91)$$

where $\mu \equiv m_p m_q / (m_p + m_q)$ is the reduced mass, and $v = v_p + v_q$ is the relative velocity of the particles. The result is the Rutherford cross-section for nonrelativistic Coulomb scattering of charges $e_p = Z_p e$ and $e_q = Z_q e$,

$$\frac{d\sigma}{do} = \left(\frac{Z_p Z_q \alpha}{2\mu v^2 \sin^2 \frac{\theta}{2}} \right)^2 . \quad (47.92)$$

In the limit that one of the spinors p or q has infinite mass, so that $\eta_p \rightarrow 0$ or $\eta_q \rightarrow 0$, then the cross-section (47.89) goes over to the Mott cross-section,

$$\frac{d\sigma}{do} = \left(\frac{Z_p Z_q \alpha}{2p_{\text{com}} v \sin^2 \frac{\theta}{2}} \right)^2 (1 - v^2 \sin^2 \frac{\theta}{2}) . \quad (47.93)$$

47.9.2 Coulomb scattering of indistinguishable spinors

The previous section 47.9.1 dealt with Coulomb scattering of distinguishable spinors. For indistinguishable spinors, in addition to the direct interaction (47.72) (with $e_q = e_p = e$ because the spinors are identical) illustrated in the left Feynman diagram 47.8, there is an exchange interaction, illustrated in the right Feynman diagram 47.8, that differs from (47.72) by swapping $\bar{\epsilon}_{q'c'} \leftrightarrow \bar{\epsilon}_{p'b'}$,

$$-(ie \bar{\epsilon}_{p'b'} \cdot \overbrace{\gamma_{ka} \epsilon_{qc}})(ie \bar{\epsilon}_{q'c'} \cdot \gamma_{ka} \epsilon_{pb}) e^{i(-p'+k+q) \cdot x'} e^{i(-q'-k+p) \cdot x} . \quad (47.94)$$

Quantum field theory attributes the overall minus sign to fermion statistics: there is a minus sign whenever two fermions are exchanged. Below, equation (47.103), it will be found that the minus sign ensures that the net amplitude is symmetric under exchange of the fermions. Recall that the amplitude is a scalar (not for example a spinor), so the indistinguishability of fermions requires the amplitude to be symmetric under fermion exchange.

Integration over the interaction vertices x and x' imposes conservation of energy-momentum at each

of the two vertices, implying that the energy-momentum $\overset{2}{k}$ of the intermediate photon for the exchange interaction (47.94) is

$$p - q' = \overset{2}{k} = -q + p' , \quad (47.95)$$

which differs from the photon energy-momentum (47.73) for the direct interaction (47.72) by $q' \leftrightarrow p'$. From equations (47.76), the photon energy-momentum $\overset{2}{k}$ for the exchange interaction (47.94) in the centre-of-mass frame is

$$\overset{2}{k} = p - q' = -2p_{\text{com}} \cos \frac{\theta}{2} \{0, \cos \frac{\theta}{2}, 0, \sin \frac{\theta}{2}\} , \quad (47.96)$$

which is related to the photon momentum $\overset{1}{k}$, equation (47.80), for the direct interaction (47.72) by

$$\overset{2}{k}(\theta) = \overset{1}{k}(\theta + \pi) = \overset{1}{k}(\theta - \pi) . \quad (47.97)$$

The electric currents $\overset{2}{j}_p^m$ and $\overset{2}{j}_q^m$ for the exchange interaction (47.94), with p and q now labelling the incoming energy-momenta of the two spinors but not their identity (since they are identical), are

$$\overset{2}{j}_p^m \equiv ie \bar{\epsilon}_{q'c'} \cdot \gamma^m \epsilon_{pb} = ie (2m) (\bar{\epsilon}_{\top c'} \cdot \tilde{R}_\eta \tilde{R}_\theta) (\gamma^m) (R_\pi R_\eta \epsilon_{\top b}) , \quad (47.98a)$$

$$\overset{2}{j}_q^m \equiv ie \bar{\epsilon}_{p'b'} \cdot \gamma^m \epsilon_{qc} = ie (2m) (\bar{\epsilon}_{\top b'} \cdot \tilde{R}_\eta \tilde{R}_{\theta+\pi}) (\gamma^m) (R_\eta \epsilon_{\top c}) , \quad (47.98b)$$

which are related to the currents $\overset{1}{j}_p^m$ and $\overset{1}{j}_q^m$ for the direct interaction, equations (47.81) with $\eta_q = \eta_p = \eta$ because the spinors p and q are now identical, by

$$\overset{2}{j}_p^m(\theta) = \overset{1}{j}_p^m(\theta - \pi) = -\overset{1}{j}_p^m(\theta + \pi) , \quad (47.99a)$$

$$\overset{2}{j}_q^m(\theta) = \overset{1}{j}_q^m(\theta + \pi) = -\overset{1}{j}_q^m(\theta - \pi) . \quad (47.99b)$$

Notice the minus signs in the rightmost expressions of equations (47.99), a sign that originates from the fact that the angle θ rotates the outgoing spinor in the currents (47.98), and rotating a spinor by 2π flips its sign. For each set of spinor spins, the exchange spinor currents $\langle c'|b \rangle \equiv \overset{2}{j}_p^m$ and $\langle b'|c \rangle \equiv \overset{2}{j}_q^m$, equations (47.98), are

$$\langle \uparrow \uparrow \rangle^m = \langle \downarrow \downarrow \rangle^m = 2em \{ -(\overset{+q}{-p}) \cosh \eta \sin \frac{\theta}{2}, \sinh \eta \cos \frac{\theta}{2}, (\overset{+q}{-p}) (\overset{+1}{-\downarrow}) i \sinh \eta \cos \frac{\theta}{2}, -\sinh \eta \sin \frac{\theta}{2} \} , \quad (47.100a)$$

$$\langle \downarrow \uparrow \rangle^m = \langle \uparrow \downarrow \rangle^m = 2em \{ (\overset{+q}{-p}) \cos \frac{\theta}{2}, 0, 0, 0 \} , \quad (47.100b)$$

which differ from the direct currents (47.82) in accordance with equations (47.99). Each current (47.100) satisfies the Ward-Takahashi identities (47.83) with $k = \overset{2}{k}$ given by equation (47.96). The amplitudes associated with the currents (47.100) are

$$\langle c'b'|cb \rangle \equiv \overset{2}{\mathcal{M}}_{c'cb'b} = -\frac{\eta_{m'm} \overset{2}{j}_q^{m'} \overset{2}{j}_p^m}{\overset{2}{k}^2} = -\frac{\eta_{m'm} \langle p'b'|qc \rangle^{m'} \langle q'c'|pb \rangle^m}{\overset{2}{k}^2} , \quad (47.101)$$

the minus sign coming from the fermionic minus sign postulated for the exchange interaction (47.94). The exchange amplitude $\overset{2}{\mathcal{M}}_{c'cb'b}$, equation (47.101), is related to the direct amplitude $\overset{1}{\mathcal{M}}_{c'cb'b}$, equation (47.84) with $\eta_q = \eta_p = \eta$, by $\theta \rightarrow \theta + \pi$ (or $\theta \rightarrow \theta - \pi$), and an exchange $c' \leftrightarrow b'$ of outgoing spin indices,

$$\overset{2}{\mathcal{M}}_{c'cb'b}(\theta) = \overset{1}{\mathcal{M}}_{b'cc'b}(\theta + \pi) . \quad (47.102)$$

The replacement $\theta \rightarrow \theta + \pi$ arises from the exchange $q' \leftrightarrow p'$ of outgoing spinor momenta, and the fact that the outgoing spinor momenta point in opposite directions in the centre-of-mass frame, Figure 47.9. The postulated minus sign in the exchange interaction (47.94) cancels the minus signs in the rightmost expressions of the exchange currents (47.99).

Combining the amplitude $\mathcal{M}$ from equation (47.84) and its spinor-exchanged partner $\mathcal{M}$ from equation (47.101) yields

$$\mathcal{M}_{c'cb'b}(\theta) = \mathcal{M}_{c'cb'b}^1(\theta) + \mathcal{M}_{c'cb'b}^2(\theta) = \mathcal{M}_{c'cb'b}^1(\theta) + \mathcal{M}_{b'cc'b}^1(\theta + \pi) , \quad (47.103)$$

which is symmetric under $\theta \rightarrow \theta + \pi$ (or $\theta \rightarrow \theta - \pi$) and exchange $c' \leftrightarrow b'$ of outgoing spin indices. The symmetry (47.103) says that the net amplitude is symmetric under exchange of the identical outgoing spinors. The net amplitudes $\langle c'b'|cb \rangle \equiv \mathcal{M}_{c'cb'b}$ for Coulomb scattering of indistinguishable spinors for each choice of spins are

$$\left. \begin{aligned} \langle \uparrow\uparrow|\uparrow\uparrow \rangle &= \langle \downarrow\downarrow|\downarrow\downarrow \rangle \\ \langle \downarrow\uparrow|\downarrow\uparrow \rangle &= \langle \uparrow\downarrow|\uparrow\downarrow \rangle \\ -\langle \downarrow\uparrow|\uparrow\uparrow \rangle &= \langle \uparrow\uparrow|\downarrow\uparrow \rangle = -\langle \downarrow\downarrow|\uparrow\downarrow \rangle = \langle \uparrow\downarrow|\downarrow\downarrow \rangle \\ = -\langle \uparrow\downarrow|\uparrow\uparrow \rangle &= -\langle \downarrow\downarrow|\downarrow\uparrow \rangle = \langle \uparrow\uparrow|\uparrow\downarrow \rangle = \langle \uparrow\downarrow|\uparrow\downarrow \rangle \\ \langle \uparrow\downarrow|\downarrow\uparrow \rangle &= \langle \downarrow\uparrow|\uparrow\downarrow \rangle \\ \langle \downarrow\downarrow|\uparrow\uparrow \rangle &= \langle \uparrow\uparrow|\downarrow\downarrow \rangle \end{aligned} \right\} = \frac{(ie)^2}{\kappa^2} \left\{ \begin{aligned} &2(\cosh 2\eta - \sin^2 \theta) , \\ &4(\cosh 2\eta \cos^2 \frac{\theta}{2} - \sin^2 \frac{\theta}{2}) \cos^2 \frac{\theta}{2} , \\ &\cosh \eta \sin 2\theta , \\ &4(\cosh 2\eta \sin^2 \frac{\theta}{2} - \cos^2 \frac{\theta}{2}) \sin^2 \frac{\theta}{2} , \\ &2\sin^2 \theta , \end{aligned} \right. \quad (47.104)$$

where

$$\frac{1}{\kappa^2} \equiv \frac{m^2}{k^2} + \frac{m^2}{k'^2} = \frac{1}{\sinh^2 \eta \sin^2 \theta} , \quad (47.105)$$

with k^2 and k'^2 the energy-momentum squared of the intermediate photon in the direct and exchange cases, equations (47.80) and (47.96), the left and right of the two Feynman diagrams 47.8. The top two lines of equations (47.104) are for scatterings that do not flip spin: in the top line, all spins are aligned, while in the second line the q and p spins are opposed. The next two lines of equations (47.104) are for scatterings where one of the q or p spins flips: in the third line the q spin flips, while in the fourth line the p spin flips. The last two lines of equations (47.104) are for scatterings where both q and p spins flip: in the fifth line the q and p spins are opposed, while in the sixth line the q and p spins are aligned.

Squared amplitudes for each choice of incoming and outgoing spins follow from squaring the amplitudes (47.104). For equal numbers of spinors in each of the incoming spin states, the amplitude squared for Coulomb scattering of indistinguishable spinors, averaged over initial spins and summed over final spins, is a sum of direct, interference, and exchange terms

$$\frac{1}{4} \mathcal{M}_{c'cb'b} \mathcal{M}^{bb'cc'} = 4e^4 \left(\left(\frac{\sinh^2 2\eta \cos^2 \frac{\theta}{2} + 1}{\sinh^4 \eta \sin^4 \frac{\theta}{2}} + 2 \right) + \left(\frac{4 \sinh^4 \eta - 1}{\sinh^4 \eta \cos^2 \frac{\theta}{2} \sin^2 \frac{\theta}{2}} \right) + \left(\frac{\sinh^2 2\eta \sin^2 \frac{\theta}{2} + 1}{\sinh^4 \eta \cos^4 \frac{\theta}{2}} + 2 \right) \right) , \quad (47.106)$$

which can be compared to the amplitude squared (47.87) for Coulomb scattering of distinguishable spinors. The differential cross-section is obtained by dividing by $(8\pi E_{\text{com}})^2$ with $E_{\text{com}} = 2p_{\text{com}} \coth \eta$ from equation (47.88).

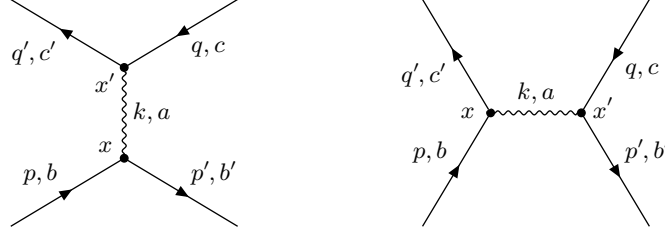


Figure 47.10 Feynman diagrams for the annihilation of a spinor-antispinor pair into another spinor-antispinor pair. If the particles are distinguishable, for example an electron-positron pair annihilating into a muon-antimuon pair, then only the left diagram contributes. If on the other hand the particles are indistinguishable, for example an electron-positron pair annihilating into another electron-positron pair, then both diagrams contribute. The arrows on the spinor lines represent the direction of charge.

47.10 Annihilation/creation of spinor-antispinor pairs

Related to the process of Coulomb scattering of spinors is the process of annihilation of a spinor-antispinor pair into another spinor-antispinor pair. To lowest order in perturbation theory, the process of pair annihilation into another pair is a two-vertex process described by the Feynman diagrams 47.10. Section 47.10.1 considers the annihilation/creation of distinguishable spinors, such as an electron-positron pair annihilating and creating a muon-antimuon pair, $e + \bar{e} \rightarrow \mu + \bar{\mu}$, in which case only the Feynman diagram on the left contributes. Section 47.10.2 then deals with the case of indistinguishable spinors, for which both diagrams in Figure 47.10 contribute.

47.10.1 Annihilation/creation of distinguishable spinors

Let p and q denote distinguishable spinors of charges e_p and e_q and masses m_p and m_q . A spinor-antispinor pair $p\bar{p}$ annihilating into a spinor-antispinor pair $q\bar{q}$, illustrated in the left Feynman diagram 47.10, is described by an interaction which looks the same as the Coulomb interaction (47.72) but with $p' \rightarrow -p'$ and $q' \rightarrow -q'$, and with the outgoing p spinor replaced by an incoming antispinor, and the incoming q spinor replaced by an outgoing antispinor,

$$(ie_q \bar{\epsilon}_{q'c'} \cdot \overline{\gamma_{ka}(i\bar{\epsilon}_{qc})}) (ie_p (-i\epsilon_{p'b'}) \cdot \gamma_{ka}\epsilon_{pb}) e^{i(-q'+k-q)\cdot x'} e^{i(p'-k+p)\cdot x} . \quad (47.107)$$

Here the rule (44.84) has been used to replace the backward-in-time outgoing spinor $\bar{\epsilon}_{p'b'}$ by a forward-in-time ingoing antispinor $-i\epsilon_{p'b'}$, and a backward-in-time incoming q spinor ϵ_{qc} by a forward-in-time outgoing antispinor $i\bar{\epsilon}_{qc}$.

Integration over x and x' vertices imposes energy-momentum conservation at each vertex,

$$p + p' = k = q + q' . \quad (47.108)$$

Each vertex term in the interaction (47.107) defines an electric current, a current j_p^m for the p spinor, and

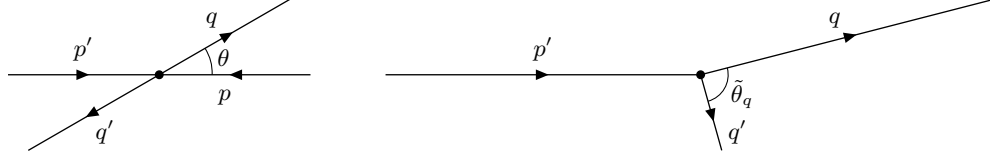


Figure 47.11 Annihilation of a spinor-antispinor pair into another spinor-antispinor pair (left) in the centre-of-mass frame, and (right) in the rest frame of the incoming p spinor. This figure is similar to Figure 47.9 but with $q \rightarrow p' \rightarrow q' \rightarrow q$. The lengths and directions of the lines are proportional to the spatial momenta. In the centre-of-mass frame, the outgoing $q\bar{q}$ pair are deflected by angle θ from the incoming $p\bar{p}$ pair. In the rest frame of the incoming p spinor, the angle between the outgoing $q\bar{q}$ pair is $\tilde{\theta}_q$, equation (47.114). If the p and q spinors are identical, then the situation $p + \bar{p} \rightarrow p + \bar{p}$ where a spinor-antispinor pair annihilate into another pair is indistinguishable from Coulomb scattering of a spinor and antispinor.

a current j_q^m for the q spinor,

$$j_p^m \equiv ie_p (-i\epsilon_{p'b'}) \cdot \gamma^m \epsilon_{pb} , \quad (47.109a)$$

$$j_q^m \equiv ie_q \bar{\epsilon}_{q'c'} \cdot \gamma^m (i\bar{\epsilon}_{qc}) . \quad (47.109b)$$

Integrating the interaction (47.107) over vertices x and x' , and summing/integrating over the energy-momentum k and spins of the intermediate photon, yields the amplitude for spinor-antispinor pair annihilation into a distinct spinor-antispinor pair,

$$\mathcal{M} = \int \frac{\eta_{m'm} j_q^{m'} j_p^m}{k^2 - i\epsilon} e^{i(-q'+k-q)\cdot x'} e^{i(p'-k+p)\cdot x} d^4x' d^4x \frac{d^4k}{(2\pi)^4} . \quad (47.110)$$

The energy-momenta of the colliding particles in their centre-of-mass frame, as illustrated in the left panel of Figure 47.11, are

$$p = m_p \{ \cosh \eta_p, 0, 0, -\sinh \eta_p \} , \quad (47.111a)$$

$$p' = m_p \{ \cosh \eta_p, 0, 0, \sinh \eta_p \} , \quad (47.111b)$$

$$q = m_q \{ \cosh \eta_q, \sinh \eta_q \sin \theta, 0, \sinh \eta_q \cos \theta \} , \quad (47.111c)$$

$$q' = m_q \{ \cosh \eta_q, -\sinh \eta_q \sin \theta, 0, -\sinh \eta_q \cos \theta \} , \quad (47.111d)$$

with rapidities η_p and η_q related by

$$m_p \cosh \eta_p = m_q \cosh \eta_q = \frac{1}{2} E_{\text{com}} , \quad (47.112)$$

with E_{com} the total energy in the centre-of-mass frame, equation (47.51). Boosted by rapidity η_p into the rest frame of the incoming p spinor, as illustrated in the right panel of Figure 47.11, the spinor and photon

energy-momenta are

$$p = m_p \{1, 0, 0, 0\} , \quad (47.113a)$$

$$p' = m_p \{\cosh 2\eta_p, 0, 0, \sinh 2\eta_p\} , \quad (47.113b)$$

$$q = m_q \{\cosh \eta_p \cosh \eta_q + \sinh \eta_p \sinh \eta_q \cos \theta, \sinh \eta_q \sin \theta, 0, \sinh \eta_p \cosh \eta_q + \cosh \eta_p \sinh \eta_q \cos \theta\} , \quad (47.113c)$$

$$q' = m_q \{\cosh \eta_p \cosh \eta_q - \sinh \eta_p \sinh \eta_q \cos \theta, -\sinh \eta_q \sin \theta, 0, \sinh \eta_p \cosh \eta_q - \cosh \eta_p \sinh \eta_q \cos \theta\} . \quad (47.113d)$$

The angle $\tilde{\theta}$ between the directions of the outgoing $q\bar{q}$ pair in the lab frame, right panel of Figure 47.7, satisfies, from equations (47.113c) and (47.113d),

$$\tan \tilde{\theta} = \frac{\sinh \eta_p \sinh 2\eta_q \sin \theta}{\sinh^2 \eta_p \cosh^2 \eta_q - \cosh^2 \eta_p \sinh^2 \eta_q \cos^2 \theta + \sinh^2 \eta_q \sin^2 \theta} . \quad (47.114)$$

The energy-momentum k of the intermediate photon in the centre-of-mass frame is

$$k = p + p' = E_{\text{com}} \{1, 0, 0, 0\} , \quad (47.115)$$

with E_{com} the total energy in the centre-of-mass frame, equation (47.112).

The electric currents j_p^m and j_q^m in spinors p and q are, equations (47.109),

$$j_p^m = ie_p (2m_p) (-i\epsilon_{\top b'} \cdot \tilde{R}_{\eta_p}) (\gamma^m) (R_\pi R_{\eta_p} \epsilon_{\top b}) , \quad (47.116a)$$

$$j_q^m = ie_q (2m_q) (\bar{\epsilon}_{\top c'} \cdot \tilde{R}_{\eta_q} \tilde{R}_{\theta+\pi}) (\gamma^m) (iR_\theta R_{\eta_q} \bar{\epsilon}_{\top c}) , \quad (47.116b)$$

where the boost and spatial rotation rotors R_η and R_θ are defined by equations (47.31). For each set of spinor spins, the spinor currents $\langle b'|b \rangle_p^m \equiv j_p^m$ and $\langle c'|c \rangle_q^m \equiv j_q^m$, equations (47.116), evaluate to

$$-\langle \uparrow \uparrow \rangle_p^m = \langle \downarrow \downarrow \rangle_p^m = 2e_p m_p \{0, 0, 0, 1\} , \quad (47.117a)$$

$$\langle \downarrow \uparrow \rangle_p^m = \langle \uparrow \downarrow \rangle_p^m = 2e_p m_p \{0, -\cosh \eta_p, \begin{pmatrix} + \\ - \end{pmatrix} i \cosh \eta_p, 0\} , \quad (47.117b)$$

$$-\langle \uparrow \uparrow \rangle_q^m = \langle \downarrow \downarrow \rangle_q^m = 2e_q m_q \{0, \sin \theta, 0, 0, \cos \theta\} , \quad (47.117c)$$

$$\langle \downarrow \uparrow \rangle_q^m = \langle \uparrow \downarrow \rangle_q^m = 2e_q m_q \{0, -\cosh \eta_q \cos \theta, \begin{pmatrix} + \\ - \end{pmatrix} i \cosh \eta_q, 0, \cosh \eta_q \sin \theta\} , \quad (47.117d)$$

where $\begin{pmatrix} + \\ - \end{pmatrix}$ is $+$ for $|\uparrow\rangle$ and $-$ for $|\downarrow\rangle$. Each current (47.117) satisfies the Ward-Takahashi identities (47.148)

$$k_m j_p^m = k_m j_q^m = 0 , \quad (47.118)$$

where the photon energy-momentum k is given by equation (47.115). Inserting the currents (47.117) into the amplitude (47.110), and integrating over vertices x and x' and the intermediate photon momentum k , yields the amplitudes for the various choices of incoming c, b and outgoing c', b' spins,

$$\langle c'b' | cb \rangle \equiv \mathcal{M}_{c'cb'b} = \frac{\eta_{m'm} j_q^{m'} j_p^m}{k^2} = \frac{\eta_{m'm} \langle q'c' | qc \rangle^{m'} \langle p'b' | pb \rangle^m}{k^2} . \quad (47.119)$$

The amplitudes $\langle c'b'|cb\rangle$ for each set of incoming and outgoing spins are

$$\left. \begin{aligned} \langle \uparrow\uparrow\uparrow\uparrow \rangle &= \langle \downarrow\downarrow\downarrow\downarrow \rangle \\ \langle \downarrow\uparrow\downarrow\uparrow \rangle &= \langle \uparrow\downarrow\uparrow\downarrow \rangle \\ -\langle \downarrow\uparrow\uparrow\uparrow \rangle &= -\langle \uparrow\uparrow\downarrow\downarrow \rangle = \langle \downarrow\downarrow\uparrow\downarrow \rangle = \langle \uparrow\downarrow\downarrow\downarrow \rangle \\ \langle \uparrow\downarrow\uparrow\uparrow \rangle &= -\langle \downarrow\downarrow\downarrow\uparrow \rangle = \langle \uparrow\uparrow\downarrow\uparrow \rangle = -\langle \downarrow\uparrow\downarrow\downarrow \rangle \\ \langle \downarrow\downarrow\uparrow\uparrow \rangle &= \langle \uparrow\uparrow\downarrow\downarrow \rangle \\ \langle \uparrow\downarrow\downarrow\uparrow \rangle &= \langle \downarrow\uparrow\uparrow\downarrow \rangle \end{aligned} \right\} = \frac{(ie_p)(ie_q)}{\kappa^2} \begin{cases} \cos \theta , \\ -\cos \theta , \\ \cosh \eta_q \sin \theta , \\ \cosh \eta_p \sin \theta , \\ -2 \cosh \eta_p \cosh \eta_q \sin^2 \frac{\theta}{2} , \\ 2 \cosh \eta_p \cosh \eta_q \cos^2 \frac{\theta}{2} , \end{cases} \quad (47.120)$$

where

$$\kappa^2 \equiv -\frac{k^2}{4m_q m_p} = \cosh \eta_q \cosh \eta_p . \quad (47.121)$$

The top two lines of equations (47.120) are for scatterings that do not flip spin: in the top line, all spins are aligned, while in the second line the q and p spins are opposed. The next two lines of equations (47.120) are for scatterings where one of the q or p spins flips: in the third line the q spin flips, while in the fourth line the p spin flips. The last two lines of equations (47.120) are for scatterings where both q and p spins flip: in the fifth line the q and p spin are aligned while in the final line the q and p spins are opposed.

Squared amplitudes for each choice of incoming and outgoing spins follow from squaring the amplitudes (47.120). For equal numbers of spinors in each of the incoming spin states, the amplitude squared for spinor-antispinor pair annihilation into a distinct spinor-antispinor pair, averaged over initial spins and summed over final spins, is

$$\frac{1}{4} \mathcal{M}_{c'cb'b} \mathcal{M}^{bb'cc'} = e_p^2 e_q^2 \left(\frac{1}{\cosh \eta_p^2} + \frac{1}{\cosh \eta_q^2} + \frac{\sinh \eta_p \sinh \eta_q \cos^2 \theta}{\cosh \eta_p \cosh \eta_q} + 1 \right) . \quad (47.122)$$

The differential cross-section is obtained by dividing by $(8\pi E_{\text{com}})^2$ with E_{com} from equation (47.88).

47.10.2 Annihilation/creation of indistinguishable spinors

The previous section 47.10.1 dealt with annihilation of a spinor-antispinor pair into a distinct spinor-antispinor pair. There is also the possibility that a spinor-antispinor pair could annihilate into an identical pair. In the case of electron-positron pairs, the process $e + \bar{e} \rightarrow e + \bar{e}$ is called Bhabha scattering, after Bhabha (1936) who first pointed out the importance of exchange and interference contributions to the process.

If a spinor-antispinor pair annihilates into an identical pair, then there is an exchange interaction, illustrated in the right Feynman diagram 47.10, that differs from the direct interaction (47.107) (with $e_q = -e_p = -e$ because q is the antiparticle of p) by the replacement $p' \rightarrow -q'$ and $\bar{\epsilon}_{q'c'} \leftrightarrow -i\epsilon_{p'c'}$,

$$-(i(-e)(-i\epsilon_{p'b'}) \cdot \overline{\gamma_{ka}(i\bar{\epsilon}_{qc})})(ie\bar{\epsilon}_{q'c'} \cdot \gamma_{ka}\epsilon_{pb})e^{i(p'+k-q)\cdot x'}e^{i(-q'-k+p)\cdot x} . \quad (47.123)$$

The overall minus sign has the same provenance as that in the exchange interaction (47.94) for Coulomb scattering.

Conservation of energy-momentum at each vertex implies that the energy-momentum $\vec{k}$ of the intermediate photon for the exchange interaction (47.123) is

$$p - q' = \vec{k} = q - p' , \quad (47.124)$$

which differs from the photon energy-momentum (47.108) for the direct interaction (47.107) by $q' \leftrightarrow -p'$. From equations (47.111), the photon energy-momentum $\vec{k}$ for the exchange interaction (47.123) in the centre-of-mass frame is

$$\vec{k} = p - q' = 2p_{\text{com}} \sin \frac{\theta}{2} \{0, \cos \frac{\theta}{2}, 0, -\sin \frac{\theta}{2}\} , \quad (47.125)$$

where $p_{\text{com}} = m \sinh \eta$ is the magnitude of the spatial momentum of each particle in the centre-of-mass frame.

The electric currents j_p^m and j_q^m for the exchange interaction (47.123) are

$$j_p^m \equiv ie (2m) (\bar{\epsilon}_{\top c'} \cdot \tilde{R}_\eta \tilde{R}_{\theta+\pi}) (\gamma^m) (R_\pi R_\eta \epsilon_{\top b}) , \quad (47.126a)$$

$$j_q^m \equiv i(-e) (2m) (-i \epsilon_{\top b'} \cdot \tilde{R}_\eta) (\gamma^m) (i R_\theta R_\eta \bar{\epsilon}_{\top c}) . \quad (47.126b)$$

For each set of spinor spins, the exchange currents $\langle c'|b \rangle^m \equiv j_p^m$ and $\langle b'|c \rangle^m \equiv j_q^m$ evaluate to

$$\langle \uparrow \uparrow \rangle^m = \langle \downarrow \downarrow \rangle^m = 2(\pm e) m \left\{ -\cosh \eta \cos \frac{\theta}{2}, \begin{pmatrix} +q \\ -p \end{pmatrix} \sinh \eta \sin \frac{\theta}{2}, \begin{pmatrix} +\uparrow \\ -\downarrow \end{pmatrix} i \sinh \eta \sin \frac{\theta}{2}, \begin{pmatrix} +q \\ -p \end{pmatrix} \sinh \eta \cos \frac{\theta}{2} \right\} , \quad (47.127a)$$

$$\langle \downarrow \uparrow \rangle^m = \langle \uparrow \downarrow \rangle^m = 2(\pm e) m \left\{ \begin{pmatrix} +q \\ -p \end{pmatrix} \sin \frac{\theta}{2}, 0, 0, 0 \right\} , \quad (47.127b)$$

where $\pm e$ denotes e for the spinor p and $-e$ for the antispinor q , the factor $\begin{pmatrix} +q \\ -p \end{pmatrix}$ is $+$ for q and $-$ for p , and $\begin{pmatrix} +\uparrow \\ -\downarrow \end{pmatrix}$ is $+$ for $\uparrow$ and $-$ for $\downarrow$. The exchange currents (47.127) can be compared to the direct currents (47.117). Each current (47.127) satisfies the Ward-Takahashi identities (47.118) with $k = \vec{k}$ given by equation (47.125). The exchange amplitude $\mathcal{M}$ associated with the exchange currents (47.126) is

$$\mathcal{M}_{c'cb'b}^2 = - \frac{\eta_{m'm} j_q^{m'} j_p^m}{\vec{k}^2} = - \frac{\eta_{m'm} \langle p'b'|qc \rangle^{m'} \langle q'c'|pb \rangle^m}{\vec{k}^2} , \quad (47.128)$$

the minus sign coming from the fermionic minus sign in the exchange interaction (47.123). Combining the direct amplitude (47.119) with the exchange amplitude (47.128) yields the net amplitude

$$\langle c'b'|cb \rangle \equiv \mathcal{M}_{c'cb'b}^1 + \mathcal{M}_{c'cb'b}^2 . \quad (47.129)$$

The net amplitudes $\langle c'b'|cb \rangle$ for each set of incoming and outgoing spins are

$$\left. \begin{aligned} \langle \uparrow \uparrow \uparrow \uparrow \rangle &= \langle \downarrow \downarrow \downarrow \downarrow \rangle \\ \langle \downarrow \uparrow \downarrow \uparrow \rangle &= \langle \uparrow \downarrow \uparrow \downarrow \rangle \\ -\langle \downarrow \uparrow \uparrow \uparrow \rangle &= -\langle \uparrow \uparrow \downarrow \downarrow \rangle = \langle \downarrow \downarrow \uparrow \downarrow \rangle = \langle \uparrow \downarrow \downarrow \downarrow \rangle \\ = \langle \uparrow \downarrow \uparrow \uparrow \rangle &= -\langle \downarrow \downarrow \downarrow \uparrow \rangle = \langle \uparrow \uparrow \uparrow \downarrow \rangle = -\langle \downarrow \uparrow \downarrow \downarrow \rangle \\ \langle \downarrow \downarrow \uparrow \uparrow \rangle &= \langle \uparrow \uparrow \downarrow \downarrow \rangle \\ \langle \uparrow \downarrow \downarrow \uparrow \rangle &= \langle \downarrow \uparrow \uparrow \downarrow \rangle \end{aligned} \right\} = \frac{(ie)^2}{\kappa^4} \left\{ \begin{aligned} &\frac{1}{2} (\cosh^2 \eta (4 \sinh^2 \eta + 1) + \sinh^2 \eta \cos^2 \theta + \cos \theta) , \\ &(1 + 2 \sinh^2 \eta \cos^2 \frac{\theta}{2}) \sin^2 \frac{\theta}{2} , \\ &-\frac{1}{2} \cosh \eta (1 + \sinh^2 \eta \cos \theta) \sin \theta , \\ &-\cosh^2 \eta (1 - 2 \sinh^2 \eta \sin^2 \frac{\theta}{2}) \sin^2 \frac{\theta}{2} , \\ &\cosh^2 \eta (1 + 2 \sinh^2 \eta \cos^2 \frac{\theta}{2}) \cos^2 \frac{\theta}{2} , \end{aligned} \right. \quad (47.130)$$

where

$$\kappa^4 \equiv -\frac{\hat{k}^2 \hat{k}^2}{(4m^2)^2} = \cosh^2 \eta \sinh^2 \eta \sin^2 \frac{\theta}{2}, \quad (47.131)$$

with $\hat{k}^2$ and $\hat{k}^2$ the energy-momentum squared of the intermediate photon in the direct and exchange cases, equations (47.115) and (47.125), the left and right of the two Feynman diagrams 47.10.

Squared amplitudes for each choice of incoming and outgoing spins follow from squaring the amplitudes (47.130). For equal numbers of spinors in each of the incoming spin states, the amplitude squared for spinor-antispinor pair annihilation into an identical spinor-antispinor pair, averaged over initial spins and summed over final spins, is a sum of direct, interference, and exchange terms,

$$\begin{aligned} & \frac{1}{4} \mathcal{M}_{c'cb'b} \mathcal{M}^{bb'cc'} = \\ & e^4 \left(\left(\frac{1 - \sinh^4 \eta \sin^2 \theta}{\cosh^4 \eta} + 2 \right) + \left(\frac{1 - 4 \left(1 + \sinh^2 \eta \cos^2 \frac{\theta}{2} \right)^2}{\cosh^2 \eta \sinh^2 \eta \sin^2 \frac{\theta}{2}} \right) + \left(\frac{1 + \sinh^2 2\eta \cos^2 \frac{\theta}{2}}{\sinh^4 \eta \sin^4 \frac{\theta}{2}} + 2 \right) \right). \end{aligned} \quad (47.132)$$

The differential cross-section is obtained by dividing by $(8\pi E_{\text{com}})^2$ with $E_{\text{com}} = p_{\text{com}} \sinh 2\eta / \sinh^2 \eta$ from equation (47.88).

47.11 Crossing symmetry and Mandelstam variables

Feynman diagrams that are related to each other by a flip in the direction of time of one or more external legs are said to be related by crossing symmetry. Flipping the direction of time of a leg is equivalent to flipping the direction of its energy-momentum, and replacing a particle by an antiparticle. For spinors, the replacement by an antispinor is accomplished by replacing a spinor ϵ_{pb} by i times its conjugate $\bar{\epsilon}_{pb}$ in accordance with the rule (44.84).

For example, the Feynman diagrams for spinor-photon scattering, Figure 47.2, are related by crossing symmetry to those for spinor-antispinor creation or annihilation from or into photons, Figure 47.6. Similarly, the Feynman diagrams for Coulomb scattering, Figure 47.8, are related by crossing symmetry to those for spinor-antispinor annihilation into another spinor-antispinor pair, Figure 47.10.

Squared amplitudes of processes related by crossing symmetry are analytically related to each other. The analytic relation between spin-summed amplitudes squared of related processes is most transparent if the squared amplitudes are expressed as scalar functions of the momenta of the incoming and outgoing particles. In scattering involving a total of 4 incoming and outgoing particles, a conventional choice of scalar functions of momenta is so-called Mandelstam variables. Denote the energy-momenta of the 4 particles by p_i with $i = 1, 2, 3, 4$, with signs chosen so positive is incoming. The three Mandelstam variables s , t , and u are

defined by

$$s \equiv (p_1 + p_2)^2 = (p_3 + p_4)^2 , \quad (47.133a)$$

$$t \equiv (p_1 + p_3)^2 = (p_2 + p_4)^2 , \quad (47.133b)$$

$$u \equiv (p_1 + p_4)^2 = (p_2 + p_3)^2 . \quad (47.133c)$$

The equalities of the right hand sides of equations (47.133) follow from conservation of total energy-momentum, $p_1 + p_2 + p_3 + p_4 = 0$. The vanishing of the square of the total energy-momentum implies

$$\frac{1}{2}(p_1 + p_2 + p_3 + p_4)^2 = s + t + u + m_1^2 + m_2^2 + m_3^2 + m_4^2 = 0 , \quad (47.134)$$

where m_i are the rest masses of the particles. Equation (47.134) implies that only two of the three Mandelstam variables are independent, their sum being constrained to equal minus the sum of the squares of their rest masses. The independence of just two of the three Mandelstam variables reflects the fact that a collision between 2 particles is characterized by just 2 independent variables, which can be taken to be for example the centre-of-mass energy E_{com} and the scattering angle θ .

In a 2-body collision, two particles are incoming and two are outgoing, so there is a minus sign between energy-momenta in the Mandelstam variables involving one ingoing and one outgoing particle. It is conventional to choose the s variable to be the sum squared of momenta of the two incoming particles, the t variable to be the difference squared of momenta of the most similar incoming and outgoing particles, and the u variable to be the difference squared of momenta of the most dissimilar incoming and outgoing particles.

If, as is conventional, the Mandelstam variable s is taken to be the sum squared of momenta of the two incoming particles, then the square of the total centre-of-mass energy E_{com} equals $-s$,

$$E_{\text{com}}^2 = -s . \quad (47.135)$$

A cross-section is obtained from an amplitude squared by multiplying by $1/(8\pi E_{\text{com}})^2$, equation (47.53), so an equivalent procedure is to multiply by $1/(64\pi^2|s|)$.

Consider spinor-photon scattering and the related process of spinor-antispinor creation/annihilation into or from photons. The processes are related by flipping the direction of energy-momentum of one of the photons, $k \rightarrow -k$, and of one of the external spinors, $p' \rightarrow -p'$. The Mandelstam variables for spinor-photon scattering, Feynman diagrams 47.2, are

$$s \equiv (p + k)^2 = (p' + k')^2 = -m^2 e^{2\eta} , \quad (47.136a)$$

$$t \equiv (p - p')^2 = (k - k')^2 = 4m^2 \sinh^2 \eta \sin^2 \frac{\theta}{2} , \quad (47.136b)$$

$$u \equiv (p - k')^2 = (p' - k)^2 = m^2 (\sinh 2\eta + 2 \sinh^2 \eta \cos \theta - 1) , \quad (47.136c)$$

while the Mandelstam variables for spinor-antispinor creation/annihilation, Feynman diagrams 47.6, are

$$s \equiv (p + p')^2 = (k + k')^2 = -4m^2 \cosh^2 \eta , \quad (47.137a)$$

$$t \equiv (p - k)^2 = (p' - k')^2 = m^2 (\cosh 2\eta + \sinh 2\eta \cos \theta) , \quad (47.137b)$$

$$u \equiv (p - k')^2 = (p' - k)^2 = m^2 (\cosh 2\eta - \sinh 2\eta \cos \theta) . \quad (47.137c)$$

The flips $k \rightarrow -k$ and $p' \rightarrow -p'$ transform $s \leftrightarrow t$, so the spin-summed squared amplitudes for the two processes should be related by transforming $s \leftrightarrow t$. For spinor-photon scattering, the average-over-incoming- and sum-over-outgoing-spins amplitude squared, equation (47.39), in terms of Mandelstam variables (47.136) is

$$\frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = -2e^4 \left(\frac{u+m^2}{s+m^2} + \frac{s+m^2}{u+m^2} + \frac{4m^2}{s+m^2} + \frac{4m^2}{u+m^2} - \left(\frac{2m^2}{s+m^2} + \frac{2m^2}{u+m^2} \right)^2 \right). \quad (47.138)$$

For spinor-antispinor creation/annihilation from/into photons, the spin-summed amplitude squared, equation (47.70), in terms of Mandelstam variables (47.137) is

$$\frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = 2e^4 \left(\frac{u+m^2}{t+m^2} + \frac{t+m^2}{u+m^2} + \frac{4m^2}{t+m^2} + \frac{4m^2}{u+m^2} - \left(\frac{2m^2}{t+m^2} + \frac{2m^2}{u+m^2} \right)^2 \right). \quad (47.139)$$

The spin-summed amplitudes squared for the two processes are indeed related to each other by $s \leftrightarrow t$, albeit with an overall minus sign. The minus sign comes squaring the imaginary i in the rule (44.84) for replacing a spinor by an antispinor. The general rule in symmetry crossing is that there is a minus sign for each external spinor (fermion) whose time direction is flipped, turning it into an antispinor. In the present case just one spinor has its time direction flipped, for a single overall minus sign.

The spin-summed spinor-photon scattering amplitude squared (47.138) can be resolved into direct, interference, and exchange terms in which the direct and exchange terms satisfy crossing symmetry, although this resolution is not gauge-invariant because the Ward-Takahashi identities (47.148) apply only to the net amplitude for a process, not to the separate contributions. Nevertheless, the analytic relation between separate contributions to the amplitude is preserved provided that the sum over photon polarizations a and a' is extended over all 4 components 0, 1, 2, 3, not just the transverse components 1, 2. With the understanding that the sum over a and a' is over 0, 1, 2, 3, the spinor-photon scattering amplitude squared (47.138) resolves into direct, interference, and exchange terms,

$$\begin{aligned} \frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = & -2e^4 \left(\left(\frac{u+m^2}{s+m^2} + \frac{2m^2}{s+m^2} - \left(\frac{2m^2}{s+m^2} \right)^2 \right) \right. \\ & \left. + \left(\frac{2m^2}{s+m^2} + \frac{2m^2}{u+m^2} - \frac{8m^4}{(s+m^2)(u+m^2)} \right) + \left(\frac{s+m^2}{u+m^2} + \frac{2m^2}{u+m^2} - \left(\frac{2m^2}{u+m^2} \right)^2 \right) \right). \end{aligned} \quad (47.140)$$

The direct and exchange terms are related by $k \leftrightarrow -k'$, hence $s \leftrightarrow u$. The net amplitude is symmetric under $s \leftrightarrow u$. Similarly, the spin-summed amplitude squared (47.139) for spinor-antispinor creation/annihilation from/into photons resolves into direct, interference, and exchange terms, again with the understanding that

the sum over a and a' is over $0, 1, 2, 3$,

$$\begin{aligned} \frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} &= 2e^4 \left(\left(\frac{u+m^2}{t+m^2} + \frac{2m^2}{t+m^2} - \left(\frac{2m^2}{t+m^2} \right)^2 \right) \right. \\ &\quad \left. + \left(\frac{2m^2}{t+m^2} + \frac{2m^2}{u+m^2} - \frac{8m^4}{(t+m^2)(u+m^2)} \right) + \left(\frac{t+m^2}{u+m^2} + \frac{2m^2}{u+m^2} - \left(\frac{2m^2}{u+m^2} \right)^2 \right) \right). \end{aligned} \quad (47.141)$$

The direct and exchange terms are related by $k \leftrightarrow k'$, hence $t \leftrightarrow u$. The net squared amplitude is symmetric under $t \leftrightarrow u$. The squared amplitudes (47.140) and (47.141) are related by $s \leftrightarrow t$.

Now consider Coulomb scattering and the related process of spinor-antispinor pair annihilation into another spinor-antispinor pair. The processes are related by flipping the direction of energy-momentum of one of each pair, $q \rightarrow -q$ and $p' \rightarrow -p'$. The Mandelstam variables for Coulomb scattering, Feynman diagrams 47.8, are

$$s \equiv (p+q)^2 = (p'+q')^2 = -(m_p \cosh \eta_p + m_q \cosh \eta_q)^2, \quad (47.142a)$$

$$t \equiv (p-p')^2 = (q-q')^2 = 4m_p m_q \sinh \eta_p \sinh \eta_q \sin^2 \frac{\theta}{2}, \quad (47.142b)$$

$$u \equiv (p-q')^2 = (p'-q)^2 = 2m_p m_q (\cosh \eta_p \cosh \eta_q + \sinh \eta_p \sinh \eta_q \cos \theta) - m_p^2 - m_q^2, \quad (47.142c)$$

while the Mandelstam variables for annihilation of a spinor-antispinor pair into another pair, Feynman diagrams 47.10, are

$$s \equiv (p+p')^2 = (q+q')^2 = -(m_p \cosh \eta_p + m_q \cosh \eta_q)^2, \quad (47.143a)$$

$$t \equiv (p-q')^2 = (p'-q)^2 = 2m_p m_q (\cosh \eta_p \cosh \eta_q - \sinh \eta_p \sinh \eta_q \cos \theta) - m_p^2 - m_q^2, \quad (47.143b)$$

$$u \equiv (p-q)^2 = (p'-q')^2 = 2m_p m_q (\cosh \eta_p \cosh \eta_q + \sinh \eta_p \sinh \eta_q \cos \theta) - m_p^2 - m_q^2. \quad (47.143c)$$

The flips $q \rightarrow -q$ and $p' \rightarrow -p'$ transform $s \rightarrow u \rightarrow t \rightarrow s$. For Coulomb scattering of distinguishable spinors, the average-over-initial sum-over-final amplitude squared, equation (47.87), in terms of Mandelstam variables (47.142) is

$$\frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = 4e_p^2 e_q^2 \left(\frac{-su + (m_p^2 + m_q^2)^2}{t^2} + \frac{1}{2} \right), \quad (47.144)$$

while for annihilation of a spinor-antispinor pair into a distinct pair, the spin-summed amplitude squared, equation (47.87), in terms of Mandelstam variables (47.143) is

$$\frac{1}{4} \mathcal{M}_{b'a'ab} \mathcal{M}^{baa'b'} = 4e_p^2 e_q^2 \left(\frac{-tu + (m_p^2 + m_q^2)^2}{s^2} + \frac{1}{2} \right). \quad (47.145)$$

Indeed the amplitudes squared of the two processes are related by $s \rightarrow u \rightarrow t \rightarrow s$. Here there is no extra overall minus sign, because two spinors have their energy-momentum flipped, and two minuses make a plus.

For Coulomb scattering of indistinguishable spinors (with $m_p = m_q = m$ because they are indistinguishable), the spin-summed amplitude squared (47.106) in terms of Mandelstam variables (47.142) is a sum of

direct, interference, and exchange terms,

$$\frac{1}{4}\mathcal{M}_{b'a'ab}\mathcal{M}^{baa'b'} = 4e^4 \left(\left(\frac{-su + 4m^4}{t^2} + \frac{1}{2} \right) + \left(\frac{(s + 4m^2)^2 - 4m^4}{tu} \right) + \left(\frac{-st + 4m^4}{u^2} + \frac{1}{2} \right) \right). \quad (47.146)$$

The direct and exchange contributions are related by $p' \leftrightarrow q'$, hence by $t \leftrightarrow u$. The net spin-summed amplitude squared (47.146) is symmetric under $t \leftrightarrow u$. For the annihilation of a spinor-antispinor pair into an indistinguishable pair, the spin-summed amplitude squared, equation (47.132), in terms of Mandelstam variables (47.143) is a sum of direct, interference, and exchange terms,

$$\frac{1}{4}\mathcal{M}_{b'a'ab}\mathcal{M}^{baa'b'} = 4e^4 \left(\left(\frac{-tu + 4m^4}{s^2} + \frac{1}{2} \right) + \left(\frac{(u + 4m^2)^2 - 4m^4}{st} \right) + \left(\frac{-su + 4m^4}{t^2} + \frac{1}{2} \right) \right). \quad (47.147)$$

The direct and exchange contributions are related by $p' \leftrightarrow -q'$, hence by $s \leftrightarrow t$. The net spin-summed amplitude squared (47.147) is symmetric under $s \leftrightarrow t$. Again, the amplitudes squared of Coulomb scattering of indistinguishable spinors, and of spinor-antispinor pair annihilation into an indistinguishable pair, equations (47.146) and (47.147), are indeed related by $s \rightarrow u \rightarrow t \rightarrow s$.

47.12 Ward-Takahashi identity

External photons, those that lie on external legs of a Feynman diagram, are transverse, that is, they have polarization vectors γ_a that are orthogonal to the photon's energy-momentum vector, §35.2. The amplitude $\mathcal{M}_{..a..}$ for such a process has photon polarization indices a that run only over transverse components, $a = +, -$ in a circularly polarized basis, or $a = 1, 2$ in a linearly polarized basis. However, vectors γ_m have indices that run over all 4 components, $m = 0, 1, 2, 3$, and the amplitude $\mathcal{M}_{..m..}$ can be considered to run over all 4 indices.

There is a general result, called the Ward-Takahashi identity, that states that $\mathcal{M}_{..m..}$ contracted over the corresponding photon 4-momentum k vanishes,

$$k^m \mathcal{M}_{..m..} = 0. \quad (47.148)$$

A proof of the identity is beyond the scope of this book, but can be found for example in Peskin and Schroeder (1995, p. 238). The identity is a consequence of electric current conservation, $\partial_m j^m = 0$, equation (44.36), which is in turn a consequence of electromagnetic gauge invariance.

An application of the Ward-Takahashi identity (47.148) is that amplitudes summed over external photon spins can be re-written as contractions over all 4 indices,

$$\mathcal{M}_{..a..}\mathcal{M}^{..a..} = \mathcal{M}_{..m..}\mathcal{M}^{..m..}. \quad (47.149)$$

The photon indices on the spinor-photon scattering amplitudes $\mathcal{M}$, equation (47.30), were defined in a frame aligned with the photon direction, which in a Newman-Penrose basis is the γ_v direction, equations (41.1). Thus the Ward-Takahashi identity (47.148) implies that the spinor-photon scattering amplitudes

satisfy

$$\mathcal{M}_{b'm'mb} = 0 \quad \text{if} \quad m' = v \text{ or } m = v . \quad (47.150)$$

47.13 Loops

Exercise 47.2 asks you to show that the number l of loops in a Feynman diagram is related to the numbers i of internal lines, v of vertices, and c of connected components by

$$l = i - v + c . \quad (47.151)$$

Feynman diagrams that contain no loops are called tree diagrams. For tree diagrams, the energy-momenta along internal lines are determined in terms of external energy-momenta by conservation of energy-momenta at the vertices. If on the other hand a Feynman diagram contains loops, then all but one of the internal energy-momenta around a loop are determined by conservation of energy-momenta at the vertices of the loop (see the solution to Exercise 47.2). If the energy-momenta around a loop are denoted p_i and the ingoing energy-momentum at each vertex $k_i = p_i - p_{i-1}$, as in Figure 47.12, and if the undetermined loop energy-momentum is taken to be p_1 , then after energy-momentum constraints are imposed there remains a propagator integral

$$\int \frac{d^4 p_1}{\prod_i (p_i^2 + m^2)} , \quad (47.152)$$

with $p_i = p_1 + \sum_{j=2}^i k_j$. This integral has logarithmic divergences at the poles of the integrand. Eliminating the divergences requires renormalization, a problem beyond the scope of the present exposition.

Exercise 47.2 Numbers of loops in a Feynman diagram. By considering the energy-momenta of each line, and energy-momentum conservation at each vertex, prove equation (47.151) for the number l of loops in a Feynman diagram in terms of the numbers of vertices and lines, and the number c of connected components.

Solution. There are $e + i$ external plus internal lines, and an energy-momentum attached to each. These $e + i$ energy-momenta are subject to a conservation law at each vertex, leaving $e + i - v$ energy-momenta

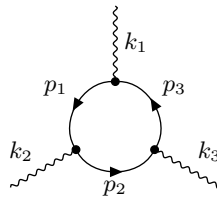


Figure 47.12 A loop in a Feynman diagram.

to be specified. Each of the c connected components of the Feynman diagram must satisfy a conservation of overall energy-momentum among its external energy-momenta, so $e - c$ of the e energy-momenta on external lines can be specified. The difference between the number of energy-momenta to be specified and the number that can be specified on external lines is $(e + i - v) - (e - c) = i - v + c$. This difference equals the number of loops, equation (47.151). That this is indeed the number of loops can be seen by considering a single loop, such as illustrated in Figure 47.12. A loop with n vertices has n internal edges and n other edges connecting to other parts of the diagram. The n vertices impose n constraints on the n internal energy-momenta. In Figure 47.12, the n energy-momentum constraints are $p_i = p_{i-1} + k_i$ in a convention where positive energy k_i points inward. One of these constraints serves to impose overall energy-momentum conservation, $\sum_i k_i = 0$, so one of the n internal energy-momenta, say p_1 , is unspecified. The remaining internal energy-momenta p_i are determined as $p_i = p_1 + \sum_{j=2}^i k_j$.

The Standard Model of Physics

The Standard Model of Physics (SM) is built on remarkably simple ingredients. All matter (leptons, quarks) is made of spinors. And all 4 forces, the 3 forces of the SM together with the force of gravity, arise from symmetries that rotate spinors. The final ingredient, the Higgs field that breaks symmetries, is an essential but poorly understood part of the SM.

The SM is based on the gauge group $U(1)_Y \times SU(2)_L \times SU(3)_c$, the product of hypercharge $U(1)_Y$, left-handed weak $SU(2)_L$, and colour $SU(3)_c$ groups. The SM group breaks to the electromagnetic group $U(1)_Q$ at energies below the electroweak unification scale ~ 200 GeV. The Lorentz group $\text{Spin}(3,1)$ (the covering group of $SO(3,1)$) is also an essential part of the SM, but is commonly treated separately since it is part of the spacetime group of general relativistic symmetries. The symmetries of general relativity comprise the Lorentz group and the group of general coordinate transformations.

The history of the development of the SM is reviewed by Weinberg (2004). The SM became established in the mid 1970s with two major successes. The first success was experimental support for the Weinberg (1967) theory of electroweak symmetry breaking, first from the discovery (Hasert et al., 1974) of a neutral Z gauge boson mediating weak interactions, and second from the discovery (Aubert et al., 1974; Augustin et al., 1974) of the charm quark, the electroweak partner of the strange quark. The second major success was the theoretical discovery (Gross and Wilczek, 1973b; Politzer, 1973) that the $SU(3)_c$ colour group was asymptotically free, meaning that its strength, measured by its coupling parameter g_c , became smaller at higher energies and larger at lower energies, in contrast to the electromagnetic force, whose coupling parameter e grows stronger at higher energies. Asymptotic freedom explained why all free states of coloured quarks seen in Nature are colourless, notably mesons comprising quark-antiquark pairs, and baryons comprising three quarks in a colourless combination. Quarks have fractional electric charges, Table 48.1, and such fractional charges were never seen in the lab, in high-energy collision experiments or otherwise, or in cosmic rays.

Weinberg (1967)'s electroweak theory gained further support from subsequent measurement of the mass of the Z boson (Arnison et al., 1983) consistent with the prediction that the ratio of Z to W boson masses should be related, equation (50.20), to the coupling parameters g_Y and g_L of the $U(1)_Y \times SU(2)_L$ electroweak group. The electroweak theory invoked the Higgs mechanism (Englert and Brout, 1964; Higgs, 1964) to accomplish symmetry breaking; the electroweak Higgs boson itself, with a mass 125 GeV, was at last detected at the Large Hadron Collider (LHC) (Aad et al., 2012; Chatrchyan et al., 2012).

For 50 years the SM has continued to be spectacularly successful in explaining physics seen in earthly experiments (Particle Data Group, 2024). Yet it is widely expected that there must be physics beyond the SM, because the SM does not account for various features of the Universe, including:

1. Quantum gravity at the singular surfaces of black holes;
2. Divergences in SM forces at high energies, a problem alleviated but not cured by renormalization;
3. Cosmological inflation may have been driven by a scalar field associated with grand symmetry breaking;
4. The baryon-antibaryon asymmetry of the Universe;
5. Dark matter;
6. Dark energy;
7. Higgs fields and their potential energy curves;
8. The pattern of masses of fermions;
9. That neutrinos have mass;
10. Why there are 3 generations of fermions.

48.1 Patterns of the Standard Model

This section describes the fermions and symmetries of the SM, emphasizing how the pattern of charges of fermions points to the SM group being a broken remnant of a larger group. See Baez and Huerta (2010) for an expository review.

The SM has 4 conserved charges consisting of hypercharge Y , weak isospin I_L (commonly abbreviated isospin¹), and 2 colours. Colour conservation is commonly described in terms of 3 colours, suggestively called red, green, and blue, which satisfy the condition that the sum of the 3 colours is colourless, or white, $(R + G + B) \pmod{3} = 0$. The fermions of the SM have charges listed in Table 48.1. Table 48.1 omits antifermions, which have charges opposite to their fermion partners. Antifermions are conventionally denoted with a bar; for example, an antineutrino is $\bar{\nu}$ (the bar here signifies a fermion with all opposite charges; in §49.1.2 it will be seen that the bar also signifies the charge conjugate). Each quark has a colour of R or G or B . Antiquarks have opposing colours; for example antired is $-R = G + B$. Actually, Table 48.1 lists only the fermions of the first generation, the electron generation. Altogether there are three generations, the electron, muon, and tauon generations, whose charges duplicate those in Table 48.1. The fermions of the three generations are distinguished by having very different masses, §48.4.

The charges in Table 48.1 show some intriguing patterns that suggest that the SM group is a broken remnant of some larger group. The three kinds of charge—hypercharge, weak isospin, and colour—each add to zero when summed over all right-handed particles (or all left-handed antiparticles), or over all left-handed particles (or all right-handed antiparticles).

The values of the hypercharge Y in Table 48.1 seem random, but they satisfy

$$3Y - 6I_L + 2(R + G + B) = 6N, \quad (48.1)$$

¹ Weak isospin is often denoted I_3 , the 3 signifying the 3rd of the 3 Pauli matrices that generate $SU(2)_L$; but I prefer the designation I_L , to emphasize that weak isospin is nonzero only for left-handed fermions (and right-handed antifermions).

Table 48.1: Conserved charges in the Standard Model

Species	symbol	$U(1)_Q$ charge $Q = \frac{1}{2}Y + I_L$	$U(1)_Y$ hypercharge Y	$SU(2)_L$ weak isospin I_L	$SU(3)_c$ colour C
Left-handed leptons	$\begin{pmatrix} \nu_L \\ e_L \end{pmatrix}$	$\begin{matrix} 0 \\ -1 \end{matrix}$	-1	$\pm \frac{1}{2}$	White
Left-handed quarks	$\begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$\begin{matrix} \frac{2}{3} \\ -\frac{1}{3} \end{matrix}$	$\frac{1}{3}$	$\pm \frac{1}{2}$	R, G, B
Right-handed neutrino	ν_R	0	0	0	White
Right-handed electron	e_R	-1	-2	0	White
Right-handed up quark	u_R	$\frac{2}{3}$	$\frac{4}{3}$	0	R, G, B
Right-handed down quark	d_R	$-\frac{1}{3}$	$-\frac{2}{3}$	0	R, G, B

where N is an integer. As prettily described by Baez and Huerta (2010), the relation (48.1) is precisely such as to allow the SM group $U(1)_Y \times SU(2)_L \times SU(3)_c$, modulo the discrete group $\mathbb{Z}_6$, to be embedded as a subgroup of $SU(5)$,

$$U(1)_Y \times SU(2)_L \times SU(3)_c / \mathbb{Z}_6 = S(U(2)_L \times U(3)_c) \subset SU(5), \quad (48.2)$$

suggesting that the SM could be a broken remnant of a larger Grand Unified Theory (GUT) group $SU(5)$, a possibility first pointed out by Georgi and Glashow (1974). The embedding is

$$\begin{aligned} U(1)_Y \times SU(2)_L \times SU(3)_c / \mathbb{Z}_6 &\rightarrow S(U(2)_L \times U(3)_c) \subset SU(5) \\ \{\alpha, g, h\} &\rightarrow \begin{pmatrix} \alpha^3 g & 0 \\ 0 & \alpha^{-2} h \end{pmatrix}, \end{aligned} \quad (48.3)$$

in which the hypercharge phase α arises as a relative phase between elements of $U(2)_L$ and $U(3)_c$. The choice of powers of α in the mapping (48.3) into $S(U(2)_L \times U(3)_c)$ is consistent with the requirement that the determinant on the right hand side be one, $(\alpha^3)^2(\alpha^{-2})^3 = 1$ (don't forget that g and h are respectively 2×2 and 3×3 matrices each of unit determinant, so the determinants of $\alpha^3 g$ and $\alpha^{-2} h$ are α^6 and α^{-6}). The map (48.3) is modded by $\mathbb{Z}_6$ because if z is any sixth root of unity, then the element $\{z, \text{diag } z^{-3}, \text{diag } z^2\} \in U(1)_Y \times SU(2)_L \times SU(3)_c$ (both $\text{diag } z^{-3} \in SU(2)_L$ and $\text{diag } z^2 \in SU(3)_c$ have unit determinant) maps to the same unit element $\{1, 1\}$ of $S(U(2)_L \times U(3)_c)$. The mapping (48.3) is viable only if the kernel $\mathbb{Z}_6$ acts trivially on all fermions of the SM. But the relation (48.1) ensures precisely this. The action of the sixth root of unity z on a fermion ψ of hypercharge Y , weak isospin I_L , and colour R or G or B is

$$\{z, z^{-3}, z^2\} : \psi \rightarrow (z)^{3Y} (z^{-3})^{2I_L} (z^2)^{R+G+B} \psi = z^{3Y-6I_L+2(R+G+B)} \psi = \psi. \quad (48.4)$$

The factors of 3 in $3Y$ and 2 in $2I_L$ in the exponents arise because hypercharge and weak isospin are

quantized in units of respectively $\frac{1}{3}$ and $\frac{1}{2}$; the choice of exponents ensures that a unit phase factor $z = e^{2\pi i}$ acts trivially on all fermions for each of the $U(1)_Y \times SU(2)_L \times SU(3)_c$ factors individually.

But there are other patterns among SM particles that $SU(5)$ does not explain: right-handed particles look like they should group into $SU(2)_R$ doublets like their left-handed counterparts; and neutrinos and electrons look like they could be another species of up and down quark with a 4th colour. As it happens, as first pointed out by Pati and Salam (1974), the SM group, modulo the discrete group $\mathbb{Z}_3$, extends as a subgroup along precisely these lines,

$$U(1)_Y \times SU(2)_L \times SU(3)_c / \mathbb{Z}_3 \subset SU(2)_R \times SU(2)_L \times SU(4)_c . \quad (48.5)$$

Consider treating the right-handed leptons and quarks as $SU(2)_R$ doublets labelled by right-handed isospin I_R , similar to their left-handed counterparts. Consider also treating White as a 4th colour W . The SM particles in Table 48.1 satisfy

$$3Y - 6I_R + 3W - (R + G + B) = 0 . \quad (48.6)$$

The pattern suggests an embedding

$$\begin{aligned} U(1)_Y \times SU(3)_c / \mathbb{Z}_3 &\rightarrow SU(2)_R \times SU(4)_c \\ \{\alpha, h\} &\rightarrow \left\{ \begin{pmatrix} \alpha^3 & 0 \\ 0 & \alpha^{-3} \end{pmatrix}, \begin{pmatrix} \alpha^{-3} & 0 \\ 0 & \alpha h \end{pmatrix} \right\} . \end{aligned} \quad (48.7)$$

The map (48.7) implies that for example left-handed leptons and quarks (which transform trivially under $SU(2)_R$) transform under $U(1)_Y$ respectively as α^{-3} and α , implying hypercharges -1 and $\frac{1}{3}$, in agreement with Table 48.1; similarly, right-handed up leptons and quarks transform as α^0 and α^4 , while right-handed down leptons and quarks transform as α^{-6} and α^{-2} , implying hypercharges 0 , $\frac{4}{3}$, -2 , and $-\frac{2}{3}$, again in agreement with Table 48.1. The map (48.7) is into only if $U(1)_Y \times SU(3)_c$ is modded by $\mathbb{Z}_3$, because if z is any third root of unity then $\{z, \text{diag } z^{-1}\} \in U(1)_Y \times SU(3)_c$ maps to the same element $\{1, 1\}$ of $SU(2)_R \times SU(4)_c$.

Exercises 39.2 and 39.3 show that $SU(2) \times SU(2)$ is isomorphic to $\text{Spin}(4)$, while $SU(4)$ is isomorphic to $\text{Spin}(6)$. Consequently the Pati-Salam group on the right hand side of the embedding (48.5) is isomorphic to $\text{Spin}(4) \times \text{Spin}(6)$,

$$SU(2)_L \times SU(2)_R \times SU(4)_c \cong \text{Spin}(4)_w \times \text{Spin}(6)_c . \quad (48.8)$$

48.2 Spin(10)

Both the $SU(5)$ embedding (48.3) and the Pati-Salam embedding (48.5) can be accommodated consistently within an even grander group $\text{Spin}(10)$, as originally proposed by Georgi (1975) and Fritzsch and Minkowski (1975). The group $\text{Spin}(4)_w \times \text{Spin}(6)_c$ embeds naturally in $\text{Spin}(10)$:

$$\text{Spin}(4)_w \times \text{Spin}(6)_c / \mathbb{Z}_2 \rightarrow \text{Spin}(10) . \quad (48.9)$$

The mapping is mod $\mathbb{Z}_2$ because flipping the signs of both $\text{Spin}(4)_w$ and $\text{Spin}(6)_c$ rotors leaves the $\text{Spin}(10)$ rotor unchanged. Besides unifying each generation of fermions in a single spinor multiplet, $\text{Spin}(10)$ predicts a right-handed neutrino which, through the see-saw mechanism, §48.4.1, allows the left-handed neutrino to acquire a small mass.

48.2.1 Spin(10) charges

As emphasized by Hamilton (2023) and discussed in Exercise 40.3, spinors in N dimensions are linear combinations of $2^{[N/2]}$ basis spinors ϵ_a labelled by a bitcode $a = a_1 \dots a_{[N/2]}$ of $[N/2]$ bits, with each of a_i being up $\uparrow$ or down $\downarrow$, equation (40.83).

The fact that the charges of $\text{Spin}(10)$ can be reorganized as a bitcode with $10/2 = 5$ bits was first pointed out by Wilczek (1998) (thanks to John Baez for this reference). The 5 bits, here denoted $yzrgb$ (§48.2.3 remarks on the naming of bits), consist of two weak bits yz and three colour bits rgb . Each bit (charge) takes the value either $+\frac{1}{2}$ ($\uparrow$) or $-\frac{1}{2}$ ($\downarrow$) for each of the $2^5 = 32$ fundamental fermions of a generation. The relation between the usual standard model charges Y , I_L , and R , G , and B , Table 48.1, is

$$Y = z + y - \frac{2}{3}(r + g + b) , \quad (48.10a)$$

$$I_L = \frac{1}{2}(z - y) , \quad (48.10b)$$

$$C = c + \frac{1}{2} \quad (c = r, g, b) . \quad (48.10c)$$

The $\text{Spin}(10)$ $yzrgb$ charges on the right hand sides of equations (48.10) are to be interpreted as $+\frac{1}{2}$ if the corresponding bit is up $\uparrow$ and $-\frac{1}{2}$ if down $\downarrow$. For example, equations (48.10) imply that the all-bit-up and all-bit-down fermions ν_R ($\uparrow\uparrow\uparrow\uparrow$) and $\bar{\nu}_L$ ($\downarrow\downarrow\downarrow\downarrow$) have SM electroweak charges $Y = I_L = 0$, and SM colour charges respectively RGB (white) and 0 (black).

Traditionally a quark has colour charge consisting of one unit of either R , G , or B . $\text{Spin}(10)$ on the other hand says that an r quark (for example) has rgb bits $\uparrow\downarrow\downarrow$, meaning that its r charge is $+\frac{1}{2}$ while its g and b charges are $-\frac{1}{2}$. In the $\text{Spin}(10)$ picture, when an r quark turns into a g quark through interaction with a $g\bar{r}$ gluon, its r charge flips from $+\frac{1}{2}$ to $-\frac{1}{2}$ while its g charge flips from $-\frac{1}{2}$ to $+\frac{1}{2}$,

$$\begin{array}{ccc} \uparrow\downarrow\downarrow & \longrightarrow & \downarrow\uparrow\downarrow \\ r \text{ quark} & g\bar{r} \text{ gluon} & g \text{ quark} \end{array} . \quad (48.11)$$

In so doing, the quark loses one unit of r charge, and gains one unit of g charge, consistent with the traditional picture.

Standard-model transformations also preserve baryon number B and lepton number L , quarks being assigned baryon number $\frac{1}{3}$ and zero lepton number, and neutrinos and electrons being assigned lepton number 1 and zero baryon number. $\text{Spin}(10)$ does not preserve baryon and lepton number individually, but it does preserve their difference $B - L$,

$$B - L = -\frac{2}{3}(r + g + b) . \quad (48.12)$$

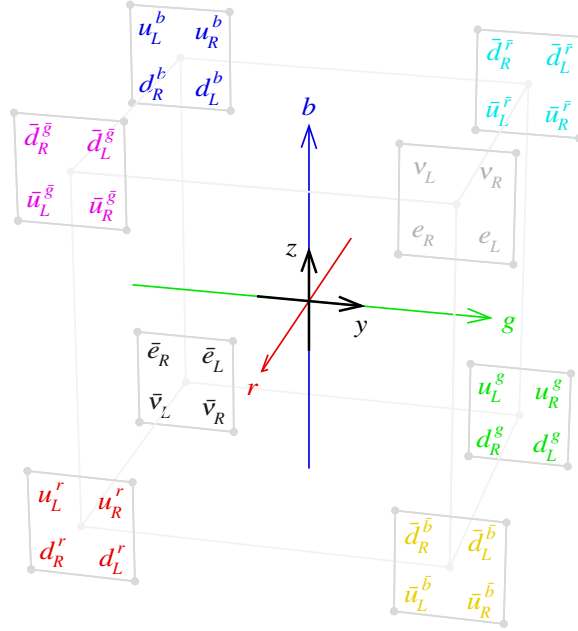


Figure 48.1 A generation (the electron generation) of 32 fermions arranged according to their Spin(10) $yrgb$ charges. The eight squares are distinguished by their colour rgb bits, with r -bit-up directed out of the page, g -bit-up directed rightward, and b -bit-up directed upward. The fermions in each of the eight squares are distinguished by their weak yz bits, with y -bit-up directed rightward, and z -bit-up directed upward. A Spin(11, 1) version of this figure is Figure 49.1.

The sum of Spin(10) charges defines an X -charge (some works normalize X differently)

$$X \equiv y + z + r + g + b = Y - \frac{5}{2}(B - L) . \quad (48.13)$$

After electroweak symmetry breaking, the electroweak group $U(1)_Y \times SU(2)_L$ breaks down to the electromagnetic group $U(1)_Q$. The electromagnetic charge Q is

$$Q = \frac{1}{2}Y + I_L = z - \frac{1}{3}(r + g + b) . \quad (48.14)$$

Electroweak symmetry breaking is a loss of y -symmetry, a loss of conservation of y -charge.

48.2.2 Spin(10) chart of spinors

Figure 48.1 illustrates one generation (the electron generation) of fermions of the standard model arranged according to their Spin(10) $yrgb$ charges.

The following Spin(10) chart (48.15) shows the electron generation of fermions of the standard model arrayed in columns according to the number of up-bits (compare Table 4 of Baez and Huerta (2010); see also Wilczek (1998)). The left element of each entry (before the colon) signifies which bits are up, from $-$ (no bits

up, or $\downarrow\downarrow\downarrow\downarrow$) in the leftmost (0) column, to $yzrgb$ (all bits up, or $\uparrow\uparrow\uparrow\uparrow$) in the rightmost (5) column; the right element of each entry is the corresponding fermion, which comprise (electron) neutrinos ν , electrons e , and up and down quarks u and y , each in right- and left-handed Dirac chiralities R and L , and each in (unbarred) particle and (barred) anti-particle species, a total of $2^5 = 32$ fermions:

Fermions and their Spin(10) bitcodes, arranged by the number of up-bits					
0	1	2	3	4	5
$- : \bar{\nu}_L$	$y : \bar{\nu}_R$	$\bar{c} : \bar{u}_L^c$	$y\bar{c} : \bar{u}_R^c$	$zrgb : \nu_L$	$yzrgb : \nu_R$
	$z : \bar{e}_R$	$yz : \bar{e}_L$	$rgb : e_R$	$yrgb : e_L$	
	$c : d_R^c$	$yc : d_L^c$	$z\bar{c} : \bar{d}_R^c$	$yz\bar{c} : \bar{d}_L^c$	
		$zc : u_L^c$	$yzc : u_R^c$		

(48.15)

Here c denotes any of the three colours r, g , or b (one colour bit up), while $\bar{c}$ denotes any of the three anti-colours gb, br , or rg (two colour bits up, the bit flip of a one-colour-bit-up spinor). Every compact spin group $\text{Spin}(N)$ contains a subgroup $\text{SU}([N/2])$ that preserves the number of up-bits (Atiyah, Bott, and Shapiro, 1964). The columns of the chart (48.15) are $\text{SU}(5)$ multiplets within $\text{Spin}(10)$, with dimensions respectively 1, 5, 10, 10, 5, 1. The standard-model group is a subgroup of $\text{SU}(5)$. All standard-model interactions preserve the number of $\text{Spin}(10)$ up-bits.

Each entry in the chart (48.15) is a 2-component massless Weyl fermion. Lorentz transformations transform the 2 components of a Weyl fermion between each other.

$\text{Spin}(10)$ predicts that the left-handed neutrino ν_L has a right-handed partner ν_R , the all-bit-up spinor $\uparrow\uparrow\uparrow\uparrow$ (whose charge conjugate $\bar{\nu}_L$ is the all-bit-down spinor $\downarrow\downarrow\downarrow\downarrow$). No right-handed neutrino is observed, although the fact that neutrinos have mass, and therefore cannot be purely chiral (purely chiral spinors being necessarily massless), suggests that there may indeed be a right-handed neutrino (see §10.25). Right-handed neutrinos have potential cosmological consequences. A right-handed “sterile” neutrino is a potential candidate for cosmological dark matter (Adhikari et al., 2017). The mass of such a neutrino must exceed the Tremaine-Gunn limit (Tremaine and Gunn, 1979) of 0.4 keV in order for it not to be Fermi degenerate in the dark-matter-dominated cores of dwarf galaxies. Out-of-equilibrium decay of three generations of right-handed neutrino in the early Universe can lead to an excess of leptons over antileptons, a process called leptogenesis (Fukugita, 1986; Buchmüller, Peccei, and Yanagida, 2005; Davidson, Nardi, and Nir, 2008; Blanchet and Di Bari, 2012; Fong, Nardi, and Riotto, 2012; Drewes, 2013; Cline, 2018). Leptogenesis can subsequently promote baryogenesis at the electroweak phase transition.

48.2.3 Naming of bits

This book follows Hamilton and McMaken (2023) in adopting the notation y, z, r, g, b for the 5 bits of $\text{Spin}(10)$ spinors, consisting of two weak bits y, z , and three colour bits r, g, b . Wilczek (1998) used r, w, b (red, white, blue) for the colour bits, and (arbitrarily) p, g (purple, green) for the weak bits. Baez and Huerta (2010) adopted r, g, b (red, green, blue) for the colour bits, and d, u (down, up) for the weak bits, on the grounds

that down and up quarks d_L^c and u_L^c have up bits respectively dc and uc (with c one of r, g, b), per the Spin(10) chart (48.15).

This book follows Baez and Huerta (2010) for the colour bits, since it is consistent with the elegant mental picture that equal amounts of the three colours r, g, b are colourless, as are all free fermionic bound states at low energy (leptons, baryons, mesons). However, Baez and Huerta (2010)'s adoption of d, u for the weak bits is potentially confusing, first because each bit d or u can itself be either down or up, and second because d and u might be misinterpreted as measuring the number of d and u quarks, potentially confusing weak isospin with isospin.

This book chooses y, z for the weak bits in part because it reduces the potential for misinterpretation, and in part because y and z are infrared bands to be used by the Vera Rubin Observatory (formerly the LSST) (Ivezić et al., 2019), for which astronomical first light occurred in 2025. The sequence $yzrgb$ is in (inverse) order of wavelength,

$$\{y, z, r, g, b\} \sim \{1000, 900, 600, 500, 400\} \text{ nm} . \quad (48.16)$$

Note that hypercharge Y carries one unit of y -charge, in addition to other charges, equation (48.10a).

48.3 Spin(10) gauge fields

The Brauer and Weyl (1935) theorem, proved in Exercise 40.3 part 5, states that the algebra of outer (tensor) products of spinors in $N = K+M$ spacetime dimensions is isomorphic to the Clifford algebra $\text{Cl}(K, M)$ (Grassmann, 1877; Clifford, 1878), also known as the geometric algebra (Hestenes, 1966; Hestenes and Sobczyk, 1987).

The orthonormal vectors γ_a , $a = 1, \dots, N$ of a geometric algebra in N dimensions satisfy the product rule

$$\gamma_a \gamma_b = \begin{cases} \delta_{ab} & (a = b) , \\ -\gamma_b \gamma_a & (a \neq b) . \end{cases} \quad (48.17)$$

In 3+1 spacetime dimensions, the geometric algebra $\text{Cl}(3, 1)$ is the Dirac algebra of Dirac γ -matrices and their products (whence the notation γ_a for the vectors of a geometric algebra), §14.7. If a dimension is timelike, as in the Dirac algebra, then it is convenient to treat the timelike dimension as the imaginary i times a spacelike dimension. An antisymmetric product of vectors is commonly written with a wedge sign, $\gamma_a \wedge \gamma_b$, although in this book the wedge sign is often omitted for brevity when there is no ambiguity. Physically, a wedge product $\gamma_{a_1} \wedge \dots \wedge \gamma_{a_p}$ of p vectors, a multivector of grade p , represents a p -dimensional element in the N -dimensional space. The geometric algebra is the algebra of linear combinations of multivectors of all possible grades $p = 0$ to N .

In the presence of spinors, the orthonormal basis vectors of a geometric algebra group naturally into pairs, conveniently denoted $\gamma_i^\pm$. Spinors with i bit up and down transform with opposite phases $e^{\mp i\theta/2}$ (or opposite boosts $e^{\pm\theta/2}$) under rotations in the 2-dimensional $\gamma_i^+ \gamma_i^-$ plane. The 10 orthonormal basis vectors of the $\text{Cl}(10)$ geometric algebra are $\gamma_i^\pm$, $i = y, z, r, g, b$. Chiral combinations of the orthonormal basis vectors are

defined by (do not confuse the index i with the quantum-mechanical imaginary i)

$$\boldsymbol{\gamma}_i \equiv \frac{\boldsymbol{\gamma}_i^+ + i\boldsymbol{\gamma}_i^-}{\sqrt{2}}, \quad \boldsymbol{\gamma}_{\bar{i}} \equiv \frac{\boldsymbol{\gamma}_i^+ - i\boldsymbol{\gamma}_i^-}{\sqrt{2}}. \quad (48.18)$$

The orthonormal vectors $\boldsymbol{\gamma}_i^+$ and $i\boldsymbol{\gamma}_i^-$ can be thought of as, modulo a normalization, the real and imaginary parts of a complex vector $\boldsymbol{\gamma}_i$ whose complex conjugate is $\boldsymbol{\gamma}_{\bar{i}}$. The chiral vectors $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_{\bar{i}}$ carry respectively $+$ and $-$ one unit of i -charge; they transform with opposite phases $e^{\mp i\theta}$ (or opposite boosts $e^{\pm\theta}$) under rotations in the 2-dimensional $\boldsymbol{\gamma}_i^+ \boldsymbol{\gamma}_i^-$ plane, and they raise and lower the charge of the object (spinor or multivector) that they act on by one unit of i -charge.

The gauge fields associated with any gauge group form a multiplet labelled by the generators of the group. The generators of the $\text{Spin}(2n)$ group with $n = 5$ are its $n(2n-1) = 45$ orthonormal bivectors (products of two orthonormal vectors), comprising the $2n(n-1) = 40$ bivectors

$$\boldsymbol{\gamma}_i^+ \wedge \boldsymbol{\gamma}_j^+ = \frac{1}{2}(\boldsymbol{\gamma}_i + \boldsymbol{\gamma}_{\bar{i}}) \wedge (\boldsymbol{\gamma}_j + \boldsymbol{\gamma}_{\bar{j}}), \quad (48.19a)$$

$$\boldsymbol{\gamma}_i^+ \wedge \boldsymbol{\gamma}_j^- = \frac{1}{2i}(\boldsymbol{\gamma}_i + \boldsymbol{\gamma}_{\bar{i}}) \wedge (\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{\bar{j}}), \quad (48.19b)$$

$$\boldsymbol{\gamma}_i^- \wedge \boldsymbol{\gamma}_j^+ = \frac{1}{2i}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_{\bar{i}}) \wedge (\boldsymbol{\gamma}_j + \boldsymbol{\gamma}_{\bar{j}}), \quad (48.19c)$$

$$\boldsymbol{\gamma}_i^- \wedge \boldsymbol{\gamma}_j^- = -\frac{1}{2}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_{\bar{i}}) \wedge (\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{\bar{j}}), \quad (48.19d)$$

with distinct indices i and j running over y, z, r, g, b , together with the $n = 5$ diagonal bivectors

$$\frac{1}{2} \boldsymbol{\gamma}_i^+ \wedge \boldsymbol{\gamma}_i^- = \frac{i}{2} \boldsymbol{\gamma}_i \wedge \boldsymbol{\gamma}_{\bar{i}}, \quad (48.20)$$

with indices i running over y, z, r, g, b .

The generators of a gauge group serve two roles. On the one hand they generate the symmetries that rotate fields. On the other hand, the generators are themselves fields that are rotated by the symmetries they generate. To appreciate the distinction, consider the diagonal chiral bivector

$$\frac{1}{2} \boldsymbol{\gamma}_i \wedge \boldsymbol{\gamma}_{\bar{i}}. \quad (48.21)$$

On the one hand, the diagonal bivector (48.21) acts as an operator whose eigenvalues equal the i charge of the objects it acts on (the normalization factor of $\frac{1}{2}$ serves to ensure that is true). As an operator, a generator acts on its argument by commutation (equivalent to left multiplication if the argument is a column spinor, or right multiplication if the argument is a row spinor). On the other hand, the diagonal bivector (48.21) is itself an object, a field, whose i charge is zero. As a field, a generator is itself acted on by commutation.

Expressions for gauge fields of subgroups of $\text{Spin}(10)$, including for $\text{SU}(5)$ and the standard-model, are collected in Appendix 48.5.

48.4 Masses of fundamental fermions

What is mass? Mass remains one of the most mysterious ingredients of the Standard Model (Quigg, 2007). In the conventional picture, the chiral (right- or left-handed) fundamental fermions of the SM are taken to be

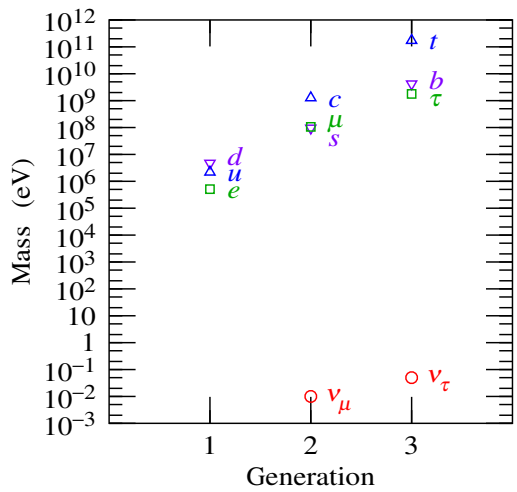


Figure 48.2 Masses of fundamental fermions, Table 48.2. Neutrino masses, and their assignment to generation, remain uncertain.

natively massless, since chirality is a property only of massless spinors. A massive spinor is a superposition of two chiral spinors of opposite chirality, a linear combination of right- and left-handed chiral spinors. A massive spinor at rest is an equal superposition of right- and left-handed spinors. For example, in the chiral representation (41.12), an electron at rest is $e_{\top} = (e_R - ie_L)/\sqrt{2}$, while a positron (an antielectron) at rest is $e_{\perp} = (-ie_R + e_L)/\sqrt{2}$.

The Spin(10) chart (48.15) of fermions shows that right- and left-handed versions of each species of fermion (for example, e_R and e_L) differ by the y -bit. The SM postulates that fermions flip their y -bit as a result of interaction with the electroweak Higgs field, §50.3, giving the fermions their fundamental masses. Spinors that come in right- and left-handed versions are called Dirac spinors, and the mass that arises from flipping between the massless right- and left-handed components is called Dirac mass. A Dirac mass that results from flipping the y -bit is possible only after electroweak symmetry breaking, where y charge is not conserved.

Table 48.2: Masses of leptons (NIST, 2022) and quarks (Particle Data Group, 2024, p. 32)

Generation								
1			2			3		
e -nu	ν_e	?	μ -nu	ν_{μ}	0.01 eV?	τ -nu	ν_{τ}	0.05 eV?
electron	e	0.510 998 950 69(16) MeV	muon	μ	105.658 375 5(23) MeV	tauon	τ	1.776 86(12) GeV
up	u	2.16(7) MeV	charm	c	1 273.0(46) MeV	top	t	172.57(29) GeV
down	d	4.70(7) MeV	strange	s	93.5(8) MeV	bottom	b	4.183(7) GeV

Table 48.2 shows the measured rest masses of the fundamental fermions, with leptons in the top two rows, quarks in the bottom two. The fundamental fermions come in 3 generations, electron, muon, and tauon (or 1, 2, and 3), each generation repeating the same pattern of charges, Table 48.1, but with different masses. Whereas the charges of fermions fit a delightfully simple pattern, their masses follow no clear design, except that higher generations are more massive, and neutrino masses are substantially smaller than other fermion masses, as illustrated in Figure 48.2. Neutrino masses, and their assignment to generation, remain as yet uncertain; neutrino oscillations, §48.4.2, yield mass squared differences, and cosmological constraints yield only an upper limit $\sum m_\nu < 0.12 \text{ eV}$ on the sum of the three neutrino masses, equation (10.111).

Most of the mass of objects in the familiar world comes not from the masses of fundamental fermions, but from protons and neutrons, which are bound states of quarks. Protons and neutrons, along with other strongly interacting particles containing an odd number of quarks, are collectively called baryons. Baryons themselves combine into nuclei, and thence with electrons into atoms and molecules. A proton is a colourless combination uud of two up quarks and one down quark, while a neutron is a colourless combination udd of one up quark and two down quarks. Colourless means that the combination is a symmetric superposition of equal contributions of r , g , and b colours. Numerical calculation of quantum chromodynamics on a lattice (lattice QCD) reveals that protons and neutrons should be thought of not as three quarks somehow stuck together, but rather as a seething maelstrom of strongly interacting relativistic quarks and gluons bound together by the colour force (Yang et al., 2018). The rest masses of the three “valence” uud or udd quarks contribute only about 1% of the $\approx 1 \text{ GeV}$ mass of a proton or neutron.

48.4.1 Neutrino mass and the see-saw mechanism

Neutrinos cannot acquire their mass in the same way as the other fundamental fermions, since only left-handed neutrinos (and right-handed antineutrinos) are observed. There is no experimental evidence for a right-handed neutrino. Evidence from particle accelerator experiments indicates that there are only 3 neutrino types with masses less than half the mass of the Z neutral weak gauge boson, $\frac{1}{2}m_Z \approx 45 \text{ GeV}$ (Particle Data Group, 2024, eq. (10.79)),

$$N_\nu = 3.003 \pm 0.006 . \quad (48.22)$$

Evidence from the CMB indicates that there are only 3 neutrino types with masses less than about the electron mass (the observations set limits on the number of neutrino types post electron-positron annihilation) (Planck Collaboration, 2018),

$$N_{\text{eff}} = 3.0 \pm 0.5 . \quad (48.23)$$

Yet neutrinos are observed to have (small) masses. How can neutrinos have mass if they are purely chiral?

A leading idea is the see-saw mechanism (Minkowski, 1977; Yanagida, 1980; Gell-Mann, Ramond, and Slansky, 1979). According to this idea, the right-handed neutrino, alone among all the fundamental fermions, could be a superposition of itself ν_R and its charge conjugate $\bar{\nu}_L$. The mass acquired by flipping between a massless particle and its charge conjugate is called a Majorana mass (Majorana, 1937). The right-handed neutrino can have a Majorana mass because it has no SM charge (no hypercharge Y , no isospin I_L , and no

colour, Table 48.1). The right-handed neutrino ν_R is the all-bit-up spinor $\uparrow\uparrow\uparrow\uparrow$ (and its charge conjugate $\bar{\nu}_L$ is the all-bit-down spinor $\downarrow\downarrow\downarrow\downarrow$), which has zero SM charge because the SM excludes the generator $i \sum_a \gamma_a \wedge \gamma_a$ that would give ν_R a charge, equation (48.42). The right-handed neutrino is the only fundamental fermion with zero SM charge. The right-handed neutrino could escape observation provided that it has a sufficiently large Majorana mass, greater than the electroweak scale ~ 1 TeV.

Although a right-handed neutrino has no SM charge, it does have lepton number L . A Majorana mass that flips between the right-handed neutrino and its charge conjugate, the left-handed antineutrino, violates conservation of lepton number. It also violates conservation of the difference $B - L$ of baryon and lepton number. SM transformations conserve both baryon B and lepton L number, and Spin(10) conserves $B - L$, though not B and L individually, equation (48.12). Does Nature allow a lepton-non-conserving Majorana mass? That is a secret that at present only Nature knows. But if it does, then out-of-equilibrium decay of three generations of right-handed neutrino in the early Universe could lead to an excess of leptons over antileptons, a process called leptogenesis.

Gell-Mann, Ramond, and Slansky (1979) proposed that neutrinos, alone among the fundamental fermions, acquire both kinds of masses, a Majorana mass M that flips the right-handed neutrino and its charge conjugate into each other $\nu_R \leftrightarrow \bar{\nu}_L$, or equivalently $\nu_{\uparrow\uparrow} \leftrightarrow \bar{\nu}_{\downarrow\downarrow}$ and $\nu_{\downarrow\downarrow} \leftrightarrow \bar{\nu}_{\uparrow\uparrow}$, and a Dirac mass m that flips right- and left-handed neutrinos into each other, $\nu_R \leftrightarrow \nu_L$, or equivalently $\nu_{\uparrow\uparrow} \leftrightarrow \nu_{\downarrow\downarrow}$ and $\nu_{\uparrow\downarrow} \leftrightarrow \nu_{\downarrow\uparrow}$. The result is that neutrino spinors are coupled to each other by a Hermitian mass matrix $\mathbf{M}$ that, in the chiral representation (41.13), is, for spin up $\uparrow$ spinors,

$$\nu^\dagger \mathbf{M} \nu = -i \begin{pmatrix} \nu_{\uparrow\uparrow} & \nu_{\downarrow\uparrow} & \bar{\nu}_{\uparrow\uparrow} & \bar{\nu}_{\downarrow\uparrow} \end{pmatrix}^* \begin{pmatrix} 0 & m & 0 & M \\ -m & 0 & 0 & 0 \\ 0 & 0 & 0 & m \\ -M & 0 & -m & 0 \end{pmatrix} \begin{pmatrix} \nu_{\uparrow\uparrow} \\ \nu_{\downarrow\uparrow} \\ \bar{\nu}_{\uparrow\uparrow} \\ \bar{\nu}_{\downarrow\uparrow} \end{pmatrix}. \quad (48.24)$$

The same mass matrix $\mathbf{M}$ holds for spinors with all boost and spin bits flipped, $\uparrow \leftrightarrow \downarrow$ and $\uparrow \rightarrow \downarrow$. The signs and normalization of equation (48.24) stem from the fact that the Dirac mass term is $m \bar{\nu} \cdot \nu = -im \nu^\dagger \gamma_0 \nu$, equation (41.102). The mass matrix $\mathbf{M}$ has 4 eigenvalues $\pm m_+$ and $\pm m_-$ with

$$m_{\pm} = \pm \frac{M}{2} + \sqrt{\left(\frac{M}{2}\right)^2 + m^2}, \quad (48.25)$$

satisfying $m_+ m_- = m^2$, or equivalently

$$\frac{m_-}{m} = \frac{m}{m_+}. \quad (48.26)$$

The condition (48.26) is called the see-saw condition. The mass eigenstates $\nu_{\pm}$ and their antiparticles $\bar{\nu}_{\pm}$ are related to the chiral eigenstates by a unitary matrix,

$$\begin{pmatrix} m_+ \\ m_- \\ -m_- \\ -m_+ \end{pmatrix} : \begin{pmatrix} \nu_+ \\ \nu_- \\ \bar{\nu}_- \\ \bar{\nu}_+ \end{pmatrix} = \frac{1}{\sqrt{2(1+a^2)}} \begin{pmatrix} 1 & -ia & a & -i \\ ia & 1 & -i & -a \\ -a & -i & 1 & ia \\ -i & a & -ia & 1 \end{pmatrix} \begin{pmatrix} \nu_{\uparrow\uparrow} \\ \nu_{\downarrow\uparrow} \\ \bar{\nu}_{\uparrow\uparrow} \\ \bar{\nu}_{\downarrow\uparrow} \end{pmatrix}, \quad (48.27)$$

where

$$a \equiv \frac{m_-}{m} = \frac{m}{m_+} . \quad (48.28)$$

If the Majorana mass M is much larger than the Dirac mass m , then the large mass m_+ approximates the Majorana mass and is much larger than the Dirac mass, $m_+ \approx M \gg m$, while the small mass m_- is much less than the Dirac mass, $m_- \approx m^2/M \ll m$. For example, if the muon neutrino has mass $m_- = m_{\nu_\mu} \approx 10^{-2}$ eV and the Dirac mass of the muon neutrino approximates the mass of the muon, $m \approx m_\mu \approx 100$ MeV, then the Majorana mass of the right-handed muon neutrino is $m_+ \approx 10^9$ GeV, well above the electroweak symmetry breaking scale, and large enough to make the right-handed neutrino inaccessible to current experiment.

If the Majorana mass M is zero, which is true for fundamental fermions other than the neutrino, then the two masses $m_\pm$ degenerate to the same Dirac mass, $m_\pm = m$. The two degenerate mass eigenstates correspond to spin up and down versions of the same spinor, and the negative mass eigenstates are their antiparticles; for example the electron $e_{\uparrow\uparrow}$ and $e_{\uparrow\downarrow}$, and its antiparticle the positron $e_{\downarrow\uparrow}$ and $e_{\downarrow\downarrow}$.

48.4.2 Neutrino oscillations

A remarkable property of fundamental fermions is that weak eigenstates are misaligned with mass eigenstates. This is true for neutrinos and quarks, and it could well be true also for the charged leptons (electrons, muons, tauons). The weak eigenstates are often called flavours, to distinguish them from mass eigenstates.

The misalignment of weak and mass eigenstates is evidenced most spectacularly by oscillations between the three generations of neutrino (Xing, 2020). Weak eigenstates ν_w , $w = e, \mu, \tau$ of neutrinos are linear combinations of mass eigenstates ν_i , $i = 1, 2, 3$,

$$\nu_w = \sum_i U_{wi} \nu_i , \quad (48.29)$$

where U_{wi} is a unitary matrix called the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix (Pontecorvo, 1958; Maki, Nakagawa, and Sakata, 1962; Gribov and Pontecorvo, 1969). When a neutrino is created, for example by the decay of a pion $\pi^+ \rightarrow \bar{\mu} + \nu_\mu$, it is created as a result of a weak interaction in a definite weak eigenstate, in this example a muon neutrino ν_μ . But that weak eigenstate is a superposition of 3 mass eigenstates, which propagate with slightly different frequencies and wavevectors. When the neutrino is then detected some distance from its creation, it has oscillated into a superposition of weak eigenstates, and may be detected as a different weak eigenstate from the one in which it was created.

Neutrino oscillations result from interference between mass eigenstates. The condition for detectable interference between a pair (or more) of propagating waves is that they differ slightly in frequency ω and/or wavevector k_a ,

$$\delta\omega \ll \omega , \quad |\delta k_a| \ll |k_a| , \quad (48.30)$$

or equivalently in energy E and/or momentum p_a ,

$$\delta E \ll E , \quad |\delta p_a| \ll |p_a| . \quad (48.31)$$

Neutrinos, whether created in the Sun or in a particle accelerator, are typically highly relativistic, and naturally satisfy the conditions (48.31). By contrast, charged leptons are typically not highly relativistic, and moreover are constantly interacting electromagnetically with other charged particles in the environment, decohering them into one or other definite mass eigenstate.

As discussed by Kayser (1981), neutrino oscillations would be destroyed if the energy-momentum of the neutrino were measured at source sufficiently accurately to determine its mass eigenstate, in much the same way that the interference pattern from a two-slit experiment is destroyed if the wave/particle is located with sufficient accuracy to determine through which slit it passed. For example, the mass of the neutrino in pion decay $\pi^+ \rightarrow \bar{\mu} + \nu_\mu$ could be determined by measuring the energy-momenta of the pion and anti-muon sufficiently accurately. Kayser (1981) concludes that a necessary condition for neutrino oscillations is a distribution of neutrino energy-momenta broad enough to admit multiple mass eigenstates. An estimate of the minimum range in energy-momentum comes from assuming that the eigenstates have the same energy, in which case the difference in their momenta is $\delta p = \delta\sqrt{E^2 - m^2} \approx -\delta(m^2)/2E$ in the relativistic approximation $E \gg m$. The resulting minimum range of momentum is

$$\frac{|\delta p|}{p} \gtrsim \frac{|\delta(m^2)|}{2p^2} . \quad (48.32)$$

A basic tenet of quantum field theory (qft) is that interactions occur at points of spacetime, and that fields propagate as waves between those points. The interaction at points means that a neutrino can be considered to be created at the origin at time zero, and then detected at position $\{t, x^a\}$. These creation and detection points are not necessarily known nor unique; qft demands integrating over whatever is not known or specified. What is meant by interactions happening at spacetime points is that an entire neutrino, including all its mass components, is created at a spacetime point, and then an entire neutrino, including all its mass components, is detected at another spacetime point. In classical mechanics, a particle moving on a straight line from the origin to $\{t, x^a\}$ has energy-momentum vector proportional to the spacetime distance along the line, $\{E, p_a\} \propto \{t, x^a\}$ (with constant of proportionality mass over proper time, m/τ). In qft by contrast, waves of all energy-momenta are permitted between interaction points. The classical energy-momentum is merely the most probable of a range of possibilities.

A plane wave of a mass eigenstate i with energy-momentum $\{E_i, p_{ai}\}$ that propagates over spacetime distance $\{t, x^a\}$ changes by a quantum mechanical phase factor $e^{i\phi_i}$ with phase $\phi_i = -E_i t + p_{ai} x^a$. The phase difference $\phi_{21} \equiv \phi_2 - \phi_1$ between two mass eigenstates 1 and 2 is

$$\phi_{21} = -E_{21}t + p_{a21}x^a , \quad (48.33)$$

with $E_{21} \equiv E_2 - E_1$ and $p_{a21} \equiv p_{a2} - p_{a1}$. Akhmedov and Smirnov (2009) give a careful exposition of the evaluation of the phase difference ϕ_{21} . The first point is that the neutrino travels many wavelengths from creation to detection. For example, a 1 MeV relativistic neutrino has a wavelength of $hc/\text{MeV} \approx 10^{-12}$ m, which is tiny compared to the distances of kilometers and more over which neutrino oscillations are measured. Consequently each neutrino mass eigenstate is well approximated as a plane wave with momentum p_{ai} aligned with the direction x^a , that is, the transverse components of momentum can be neglected. Moreover, by the time it has travelled many wavelengths, each mass eigenstate is to an excellent approximation on-shell,

meaning that the energy and momentum of a mass eigenstate of mass m_i are related by $E_i = \sqrt{p_i^2 + m_i^2}$. The second point is that each mass eigenstate i should be described by a wavepacket with a small but finite range of momentum p_i about some central momentum $\bar{p}_i$. The group velocity of the wavepacket is $v_i = \partial E_i / \partial p_i|_{p_i=\bar{p}_i}$. The third point is that the velocity $v = x/t$ between creation and detection equals the mean group velocity $\bar{v} \equiv \frac{1}{2}(v_1 + v_2)$ with an uncertainty of order the difference $v_{21} \equiv v_2 - v_1$ of group velocities, $v = \bar{v} + O(v_{21})$. Under these conditions, the phase difference between the mass eigenstates is

$$\phi_{21} = \frac{-m_{21}^2 x}{2\bar{p}} . \quad (48.34)$$

More precisely, including leading uncertainties, the phase difference is

$$\phi_{21} = \frac{-m_{21}^2 x}{2\bar{p}} \left(1 + \alpha \left(\frac{m_{21}^2}{2\bar{p}^2} + \frac{p_{21}}{\bar{p}} \right) + O\left(\frac{p_{21}}{\bar{p}}\right)^2 \right) , \quad (48.35)$$

where α , a number of order unity, measures the departure of the spacetime velocity from the group velocity, $v = \bar{v} + \alpha v_{21}$. Equation (48.34) implies that the wavelength of a neutrino oscillation is

$$\lambda = \frac{4\pi\bar{p}}{m_{21}^2} = 24.8 \text{ km} \left(\frac{m_{21}}{0.01 \text{ eV}} \right)^{-2} \left(\frac{\bar{p}}{1 \text{ MeV}} \right) . \quad (48.36)$$

To illustrate how the calculation of neutrino oscillations works out, consider the example of just two neutrino eigenstates. The unitary matrix (48.29) is then a 2×2 matrix. Three arbitrary phases can be absorbed into a rephasing of the weak and mass eigenstates ν_w and ν_i , which reduces the matrix without loss of generality to

$$U_{wi} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} . \quad (48.37)$$

The quantum-mechanical amplitude for a neutrino created in weak eigenstate ν_w to be detected as the weak eigenstate $\nu_{w'}$ is

$$\langle \nu_{w'} | \nu_w \rangle = \sum_i \langle \nu_{w'} | \nu_i \rangle e^{-i\phi_i} \langle \nu_i | \nu_w \rangle = \sum_i U_{w'i} e^{-i\phi_i} U_{iw}^* , \quad (48.38)$$

where ϕ_i is the change of phase of mass eigenstate i from creation to detection. The probability that the initial weak state ν_w is detected as the other weak state $\nu_{w'}$ is the square of the amplitude (48.38), which simplifies to

$$P(\nu_w \rightarrow \nu_{w'}) = |\langle \nu_{w'} | \nu_w \rangle|^2 = \sin^2 2\theta \sin^2 \left(\frac{1}{2} \phi_{21} \right) , \quad (48.39)$$

where ϕ_{21} is the phase difference of the mass eigenstates, equation (48.34). The probability of no change in the weak state is 1 minus the probability (48.39) of a change,

$$P(\nu_w \rightarrow \nu_w) = P(\nu_{w'} \rightarrow \nu_{w'}) = 1 - P(\nu_w \rightarrow \nu_{w'}) . \quad (48.40)$$

48.5 Appendix: Gauge fields of subgroups of Spin(10)

The standard-model gauge group $U(1)_Y \times SU(2)_L \times SU(3)_c$ is a subgroup of the $SU(5)$ subgroup of $Spin(10)$. The gauge fields (generators) of $SU(5)$ comprise the subset of gauge fields of $Spin(10)$ that leave the number of up-bits of a spinor unchanged. The gauge bivectors of $SU(n)$ with $n = 5$ constitute $(n+1)(n-1) = 24$ bivectors comprising the $n(n-1) = 20$ off-diagonal bivectors

$$\frac{1}{2}(1 - \kappa_{ij})\gamma_i^+\gamma_j^+ = \frac{1}{2}(\gamma_i^+\gamma_j^+ + \gamma_i^-\gamma_j^-) = \frac{1}{2}(\gamma_i \wedge \gamma_j + \gamma_{\bar{i}} \wedge \gamma_{\bar{j}}) , \quad (48.41a)$$

$$\frac{1}{2}(1 - \kappa_{ij})\gamma_i^+\gamma_j^- = \frac{1}{2}(\gamma_i^+\gamma_j^- - \gamma_i^-\gamma_j^+) = \frac{i}{2}(\gamma_i \wedge \gamma_j - \gamma_{\bar{i}} \wedge \gamma_{\bar{j}}) , \quad (48.41b)$$

and the $n-1 = 4$ diagonal bivectors

$$\frac{1}{2}\gamma_i^+\gamma_i^- = \frac{i}{2}\gamma_i \wedge \gamma_{\bar{i}} \quad \text{modulo} \quad \frac{1}{2}\sum_i \gamma_i^+\gamma_i^- = \frac{i}{2}\sum_i \gamma_i \wedge \gamma_{\bar{i}} , \quad (48.42)$$

with indices i and j running over y, z, r, g, b . The quantity $\kappa_{ij} \equiv \gamma_i \wedge \gamma_{\bar{i}} \wedge \gamma_j \wedge \gamma_{\bar{j}} = -\gamma_i^+\gamma_i^-\gamma_j^+\gamma_j^-$ in equations (48.41) is the ij chiral operator. The factor $\frac{1}{2}(1 - \kappa_{ij})$ is a projection operator, whose square is itself, which serves to project its argument into the space where the sum of i and j charges is zero. The S in $SU(5)$ restricts to $U(5)$ matrices of unit determinant, effectively removing the bivector $\frac{1}{2}\sum_i \gamma_i^+\gamma_i^-$ that rotates all spinors in an $SU(5)$ multiplet by a common phase.

The bivectors of $Spin(10)$ that are not in $SU(5)$ are the 20 off-diagonal bivectors

$$\frac{1}{2}(1 + \kappa_{ij})\gamma_i^+\gamma_j^+ = \frac{1}{2}(\gamma_i^+\gamma_j^+ - \gamma_i^-\gamma_j^-) = \frac{1}{2}(\gamma_i \wedge \gamma_j + \gamma_{\bar{i}} \wedge \gamma_{\bar{j}}) , \quad (48.43a)$$

$$\frac{1}{2}(1 + \kappa_{ij})\gamma_i^+\gamma_j^- = \frac{1}{2}(\gamma_i^+\gamma_j^- + \gamma_i^-\gamma_j^+) = -\frac{i}{2}(\gamma_i \wedge \gamma_j - \gamma_{\bar{i}} \wedge \gamma_{\bar{j}}) , \quad (48.43b)$$

and the 1 diagonal bivector

$$iX \equiv \frac{1}{2}\sum_i \gamma_i^+\gamma_i^- = \frac{i}{2}\sum_i \gamma_i \wedge \gamma_{\bar{i}} , \quad (48.44)$$

with indices i running over y, z, r, g, b . The bivector X measures total $yzrgb$ charge $y + z + r + g + b$, and iX is the generator of the $U(1)$ factor that would complete $SU(5)$ to $U(5)$.

The 6 bivectors involving the weak charges y and z generate the weak group $Spin(4)_w = SU(2)_R \times SU(2)_L$, a product of right- and left-handed components. The subgroup of $Spin(4)_w$ that preserves the number of yz up-bits is the standard-model left-handed component $SU(2)_L$. The bivectors of the left-handed weak group $SU(2)_L$ comprise the $2+(2-1) = 3$ bivectors (48.41) and (48.42) with i and j running over y and z . The bivectors are equivalent to $i\tau_i$ with τ_i satisfying the Pauli algebra,

$$i\tau_1 \equiv \frac{1}{2}(1 - \kappa_{yz})\gamma_y^+\gamma_z^- = \frac{1}{2}(\gamma_y^+\gamma_z^- - \gamma_y^-\gamma_z^+) = \frac{i}{2}(\gamma_y \wedge \gamma_{\bar{z}} + \gamma_z \wedge \gamma_{\bar{y}}) , \quad (48.45a)$$

$$i\tau_2 \equiv -\frac{1}{2}(1 - \kappa_{yz})\gamma_y^+\gamma_z^+ = -\frac{1}{2}(\gamma_y^+\gamma_z^+ + \gamma_y^-\gamma_z^-) = -\frac{i}{2}(\gamma_y \wedge \gamma_{\bar{z}} - \gamma_z \wedge \gamma_{\bar{y}}) , \quad (48.45b)$$

$$i\tau_3 \equiv -\frac{1}{2}(1 - \kappa_{yz})\gamma_y^+\gamma_y^- = -\frac{1}{2}(\gamma_y^+\gamma_y^- - \gamma_z^+\gamma_z^-) = \frac{i}{2}(\gamma_z \wedge \gamma_{\bar{z}} - \gamma_y \wedge \gamma_{\bar{y}}) , \quad (48.45c)$$

where $\kappa_{yz} \equiv \gamma_y \wedge \gamma_{\bar{y}} \wedge \gamma_z \wedge \gamma_{\bar{z}}$ is the weak chiral operator. The left-handed projection operator $\frac{1}{2}(1 - \kappa_{yz})$ equals 1 acting on left-handed weak chiral states, and vanishes acting on right-handed weak chiral states. The

bivectors of the right-handed weak group $SU(2)_R$ comprise 3 bivectors equivalent to $i\rho_i$ with ρ_i satisfying the Pauli algebra,

$$i\rho_1 \equiv -\frac{1}{2}(1 + \kappa_{yz})\gamma_y^+\gamma_z^- = -\frac{1}{2}(\gamma_y^+\gamma_z^- + \gamma_y^-\gamma_z^+) = \frac{i}{2}(\gamma_y \wedge \gamma_z - \gamma_{\bar{y}} \wedge \gamma_{\bar{z}}) , \quad (48.46a)$$

$$i\rho_2 \equiv \frac{1}{2}(1 + \kappa_{yz})\gamma_y^+\gamma_z^+ = \frac{1}{2}(\gamma_y^+\gamma_z^+ - \gamma_y^-\gamma_z^-) = \frac{1}{2}(\gamma_y \wedge \gamma_z + \gamma_{\bar{y}} \wedge \gamma_{\bar{z}}) , \quad (48.46b)$$

$$i\rho_3 \equiv \frac{1}{2}(1 + \kappa_{yz})\gamma_y^+\gamma_y^- = \frac{1}{2}(\gamma_y^+\gamma_y^- + \gamma_z^+\gamma_z^-) = \frac{i}{2}(\gamma_y \wedge \gamma_{\bar{y}} + \gamma_z \wedge \gamma_{\bar{z}}) . \quad (48.46c)$$

The right-handed projection operator $\frac{1}{2}(1 + \kappa_{yz})$ equals 1 acting on right-handed weak chiral states, and vanishes acting on left-handed weak chiral states. The expressions for the bivectors $i\rho_1$ and $i\rho_2$ are equivalent to equations (48.43b) and (48.43a) with indices i and j equal to y and z . The diagonal bivector $i\rho_3$ measures the total yz charge $y + z$,

$$i\rho_3 = \frac{1}{2} \sum_{i=y,z} \gamma_i^+ \gamma_i^- = \frac{i}{2} \sum_{i=y,z} \gamma_i \wedge \gamma_{\bar{i}} . \quad (48.47)$$

The 15 bivectors involving the colour charges r , g , and b generate the colour group $Spin(6)_c$. The subgroup of $Spin(6)_c$ that preserves the number of rgb up-bits is the standard-model colour group $SU(3)_c$. The bivectors of the colour group $SU(3)_c$ comprise the $6+(3-1) = 8$ bivectors (48.41) and (48.42) with i and j running over r , g , and b . The 7 other bivectors of $Spin(6)_c$ consist of the 6 leptoquark bivectors given by equations (48.43) with i, j drawn from pairs of distinct colour indices r, g, b , and the 1 diagonal bivector that measures the total rgb charge $r + g + b$,

$$-i\frac{3}{2}(B - L) = \frac{1}{2} \sum_{i=r,g,b} \gamma_i^+ \gamma_i^- = \frac{i}{2} \sum_{i=r,g,b} \gamma_i \wedge \gamma_{\bar{i}} . \quad (48.48)$$

The 1 hypercharge bivector, the generator of $U(1)_Y$, is defined to be the bivector whose eigenvalue is iY where Y is the hypercharge, equation (48.10a),

$$iY \equiv \frac{1}{2} \sum_{i=y,z} \gamma_i^+ \gamma_i^- - \frac{1}{3} \sum_{i=r,g,b} \gamma_i^+ \gamma_i^- = i \left(\frac{1}{2} \sum_{i=y,z} \gamma_i \wedge \gamma_{\bar{i}} - \frac{1}{3} \sum_{i=r,g,b} \gamma_i \wedge \gamma_{\bar{i}} \right) . \quad (48.49)$$

The 1 electromagnetic charge bivector, the generator of $U(1)_Q$, is defined to be the bivector whose eigenvalue is iQ where Q is the electric charge, equation (48.14),

$$iQ \equiv \frac{1}{2} \gamma_z^+ \gamma_z^- - \frac{1}{6} \sum_{i=r,g,b} \gamma_i^+ \gamma_i^- = i \left(\frac{1}{2} \gamma_z \wedge \gamma_{\bar{z}} - \frac{1}{6} \sum_{i=r,g,b} \gamma_i \wedge \gamma_{\bar{i}} \right) . \quad (48.50)$$

Unification of $\text{Cl}(10)$ and the Dirac algebra in $\text{Cl}(11,1)$

The previous Chapter 48 showed, §48.2, that each generation of fermions of the standard model fits neatly into a spinor representation of the group $\text{Spin}(10)$, and that the charges of fermions are elegantly described by a $(10/2 = 5)$ -bit bitcode *yzrgb* with 2 weak bits *yz* and 3 colour bits *rgb*, Figure 48.1 and chart (48.15). The Brauer-Weyl (1935) theorem, proved in Exercise 40.3 part 5, states that the algebra of outer (tensor) products of spinors in $N = K+M$ spacetime dimensions is isomorphic to the geometric (Clifford) algebra $\text{Cl}(K, M)$ of multivectors of all grades 0 to N . The gauge bosons of the standard model constitute a subset of the bivectors of the geometric algebra $\text{Cl}(10)$, §48.3.

Each of the $2^5 = 32$ spinors of a generation of $\text{Spin}(10)$ is actually a Weyl spinor, a massless chiral Dirac spinor, with 2 degrees of freedom. A Weyl spinor transforms under the Lorentz group $\text{Spin}(3, 1)$. Including both $\text{Spin}(10)$ and Weyl degrees of freedom, each generation of fermions contains $2^6 = 64$ spinors. The smallest spinor group that contains 2^6 spinors and 1 time dimension is $\text{Spin}(11, 1)$. The Dirac algebra of gamma matrices required by the Dirac Lagrangian (44.3) is the geometric algebra $\text{Cl}(3, 1)$ in 3+1 spacetime dimensions. This chapter shows that the algebra of the standard model and the Dirac algebra $\text{Cl}(3, 1)$ are commuting subalgebras of the geometric algebra $\text{Cl}(11, 1)$ in 11+1 spacetime dimensions, in compliance with the Coleman-Mandula no-go theorem (Coleman and Mandula, 1967; Mandula, 2015). The Coleman-Mandula theorem states that if a symmetry group admits nontrivial analytic scattering amplitudes, then generators of those symmetries must commute with those of spacetime. The Coleman-Mandula generalizes to higher dimensions (Pelc and Horwitz, 1997).

This chapter is based largely on Hamilton and McMaken (2023). However, there has been a major evolution in our thinking since Hamilton and McMaken (2023) was written. There it seemed natural that, if the grand unified group was $\text{Spin}(11, 1)$, then before grand symmetry breaking spacetime would have 11+1 large spacetime dimensions. However, this led to some puzzles, highlighted in §10.2 of Hamilton and McMaken (2023), leading to their suggestion that the grand unified group might have been $\text{Spin}(10, 1)$ rather than $\text{Spin}(11, 1)$. Our more recent thinking is motivated by string theory, Chapter ??, where the Chan-Paton condition, §??, requires that the internal gauge group before grand symmetry breaking was $\text{SU}(16)$ rather than $\text{Spin}(10)$. As shown in §49.3, this in turn indicates that spacetime had 3+1 large spacetime dimensions before as well as after grand symmetry breaking, resolving some of the puzzles in §10.2 of Hamilton and McMaken (2023).

String theory, Chapter ??, presents a more complicated picture than a simple geometry of 11+1 dimensions. The 11+1 spacetime dimensions considered in the present chapter prove to be the dimensions of a D-brane that carries the fermions of the standard model. Bosonic string theory requires 26 spacetime dimensions, and the extra dimensions over and above the 11+1 dimensions of the D-brane must be supplied by a bosonic torus. Most of the stringy details do not matter to the present chapter. What is relevant is that the internal symmetry group was SU(16), and internal and spacetime symmetries were separate already before grand symmetry breaking.

The plan of this chapter is as follows. Section 49.1 points out some striking features of the Spin(10) chart (48.15) that suggest unification of internal and Lorentz algebras in the Cl(11,1) geometric algebra. Section 49.2 carries out that unification. The important results are the expressions (49.14) for Dirac vectors γ_m in terms of Cl(11,1) multivectors, and the Spin(11,1) chart (49.7) of spinors. Section 49.3 derives the corresponding expressions (49.20) for Dirac vectors γ_m in terms of Cl(11,1) multivectors before SU(8) symmetry breaking.

49.1 Striking features of the Spin(10) chart

The Spin(10) chart (48.15) shows some striking features.

The most striking feature is that Spin(10) chirality coincides with Dirac chirality. Physically, this reflects the well known fact that left-handed from the perspective of weak interactions coincides with left-handed from the perspective of spacetime. Chirality counts whether the number of up-bits is even or odd, with right-handed defined as all bits up. The odd and even columns of the Spin(10) chart (48.15) have respectively right-handed (R) and left-handed (L) Dirac chirality. Modulo a phase, chirality is (the eigenvalue of) the pseudoscalar of the algebra, the product of all the N vectors in the N -dimensional geometry. The Dirac chiral operator is traditionally denoted γ_5 , which equals $-i$ times the Dirac pseudoscalar I . The Spin(10) chiral operator can be denoted $\varkappa_{yzrgb}$, which equals $-i$ times the Spin(10) pseudoscalar I_{yzrgb} . The coincidence (signified $\stackrel{!}{=}$) between Dirac and Spin(10) chirality in the chart (48.15) is

$$\begin{aligned} \gamma_5 &\equiv -iI && \equiv -i\gamma_0\gamma_1\gamma_2\gamma_3 \\ &\stackrel{!}{=} \varkappa_{yzrgb} && \equiv -iI_{yzrgb} \equiv -i\gamma_y^+\gamma_y^-\gamma_z^+\gamma_z^-\gamma_r^+\gamma_r^-\gamma_g^+\gamma_g^-\gamma_b^+\gamma_b^- . \end{aligned} \quad (49.1)$$

The coincidence of Dirac and Spin(10) chiralities suggests that the vectors γ_m , ($m = 0, 1, 2, 3$) of spacetime are related to the vectors $\gamma_i^\pm$ ($i = y, z, r, g, b$) of Spin(10), in contrast to the usual assumption that the generators of spacetime are unrelated to (commute with) those of grand unified symmetries.

A second striking feature of the chart (48.15) is that standard-model transformations are arrayed vertically, whereas the four components of fermions of the same species, such as electrons e_R and e_L and their positron partners $\bar{e}_L$ and $\bar{e}_R$, are arrayed (mostly) horizontally. In Dirac theory, a Dirac spinor such as an electron has four complex components that are distinguished by their properties under Lorentz transformations. With standard-model transformations vertical and spacetime transformations horizontal, the chart (48.15) seems

to invite the idea that $\text{Spin}(10)$ somehow contains both standard-model and spacetime transformations under one tent.

A third striking feature is that flipping the y -bit preserves the identity of the spinor but flips its chirality; for example the electron is flipped $e_R \leftrightarrow e_L$. Electroweak symmetry breaking is a loss of y -symmetry and a loss of conservation of y -charge. In the standard model, electroweak symmetry breaking is accomplished by a Higgs field that carries y -charge and gives mass to fermions by flipping their y -bit.

A fourth tea leaf is that fermions (unbarred) and antifermions (barred) in the chart (48.15) have colour chirality $\varkappa_{rgb}$ respectively positive and negative,

$$\varkappa_{rgb} \equiv iI_{rgb} \equiv i\gamma_r^+ \gamma_r^- \gamma_g^+ \gamma_g^- \gamma_b^+ \gamma_b^- . \quad (49.2)$$

Colour chirality $\varkappa_{rgb}$ is positive or negative as the number of colour up-bits is odd or even.

49.1.1 Two extra dimensions, and one extra bit

Each of the $2^{[10/2]} = 2^5 = 32$ fermions of a generation of $\text{Spin}(10)$ is a Weyl (massless, chiral) spinor, containing 2 degrees of freedom, for a total of $2^6 = 64$ fermions in a generation. The smallest spin group containing 2^6 fermions and one time dimension is $\text{Spin}(11,1)$. If the spinors of $\text{Spin}(10)$ and the Lorentz group $\text{Spin}(3,1)$ are to be unified in a minimal fashion, in a common spin group with no superfluous degrees of freedom, then that group is $\text{Spin}(11,1)$.

The reason it is possible to embed $\text{Spin}(10)$ and $\text{Spin}(3,1)$ in a group $\text{Spin}(11,1)$ of just 12 dimensions rather than 14 dimensions is precisely that $\text{Spin}(10)$ chirality coincides with $\text{Spin}(3,1)$ (Dirac) chirality.

Adding two extra dimensions adjoins an additional, 6th, t -bit, or time bit, to the 5 $yzrgb$ bits of a $\text{Spin}(10)$ spinor. Like the other 5 bits, the t -charge of a $\text{Spin}(11,1)$ spinor can be either $+\frac{1}{2}$ (t -bit up) or $-\frac{1}{2}$ (t -bit down).

In Dirac theory, massive spinors at rest are eigenvectors of the time axis, equations (14.109), and massive spinors and antispinors are conjugates of each other, equation (41.100). These two conditions require interpreting the 12th dimension γ_t^- , not the 11th dimension γ_t^+ , as providing a timelike dimension γ_0 ,

$$\gamma_0 \equiv i\gamma_t^- . \quad (49.3)$$

49.1.2 The spinor metric and the conjugation operator

Any spin algebra contains two operators, the spinor metric ε , Exercise 40.3 part 3, and the conjugation operator C , Exercise 41.5 part 14, that are invariant under rotations. A consistent translation between Dirac and $\text{Spin}(11,1)$ representations must agree on the behaviour of these two operators.

The Dirac spinor metric ε , equation (41.38), and conjugation operator C , equation (41.87), are respectively antisymmetric ($\varepsilon^2 = -1$) and symmetric ($CC^* = 1$). Consistency requires that the $\text{Spin}(11,1)$ spinor metric and conjugation operator be similarly antisymmetric and symmetric. Consultation of Tables 40.1 and 41.1 shows that in 11+1 dimensions only the standard choice ε of spinor metric and associated conjugation

operator C possess the desired antisymmetry and symmetry. The standard spinor metric in 11+1 dimensions is

$$\varepsilon = \gamma_t^+ \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ . \quad (49.4)$$

Below it will be found that the representation of the spatial rotation generator $J\sigma_2$, equation (49.12), coincides with the representation of the spinor metric (49.4), which is similar to the coincidence (41.38) between $I\sigma_2$ and the spinor metric ε in the chiral representation of the Dirac algebra. Given the Spin(11,1) spinor metric (49.4), and with the time axis $\gamma_0 = i\gamma_t^-$, the Spin(11,1) conjugation operator is

$$C = -i\varepsilon\gamma_0^\top = i\varepsilon\gamma_0 = -\varepsilon\gamma_t^- = \gamma_t^+ \gamma_t^- \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ . \quad (49.5)$$

Whereas the Spin(11,1) spinor metric (49.4) flips all bits, the Spin(11,1) conjugation operator (49.5) flips all bits except t , that is, it flips $yzrgb$. This is the same as the Spin(10) conjugation operator, which flips the five $yzrgb$ bits.

That the Cl(11,1) and Dirac (Cl(3,1)) geometric algebras share the same symmetries is no coincidence: it stems from the period-8 Cartan-Bott periodicity (Study and Cartan, 1908; Bott, 1970; Coquereaux, 1982) of geometric algebras, evident in Tables 40.1 and 41.1.

49.1.3 Spin(11,3) or Spin(13,1)?

Nesti and Percacci (2010) and Krasnov (2022) have proposed that Spin(10) and the Lorentz group Spin(3,1) are unified in Spin(11,3), and that the 64 spinors of a generation comprise one of the two chiral components of the $2^7 = 128$ spinors of Spin(11,3). The 64 spinors comprise 32 spinors that are direct products of right-handed Spin(10) spinors with right-handed Dirac spinors, and their 32 antispinor conjugate partners that are direct products of left-handed Spin(10) spinors with left-handed Dirac spinors. It is precisely the coincidence of Spin(10) and Dirac chirality that ensures the consistency of the construction.

Percacci (1991) (see Nesti and Percacci (2008)) previously proposed that the grand unified group SO(10) and the Lorentz group SO(3,1) are unified in SO(13,1). That proposal runs into the difficulty that the conjugation operator in 13+1 spacetime dimensions is antisymmetric, whereas the conjugation operator in the Dirac algebra is symmetric, Table 41.1.

49.2 Spin(11,1)

49.2.1 Spin(11,1) chart of spinors

The challenge now is to reinterpret the Spin(10) chart (48.15) of fermions as living in Spin(11,1). The conventional interpretation of the Spin(10) chart (48.15) is that each spinor is a 2-component Weyl spinor. For example, the right-handed electron e_R is the Weyl spinor with components $e_{\uparrow\uparrow}$ and $e_{\downarrow\downarrow}$. The conventional interpretation is tantamount to assuming that the Dirac and Spin(10) algebras are distinct. If instead the Dirac and Spin(10) algebras combine into Spin(11,1), then each entry in the Spin(10) chart (48.15) must

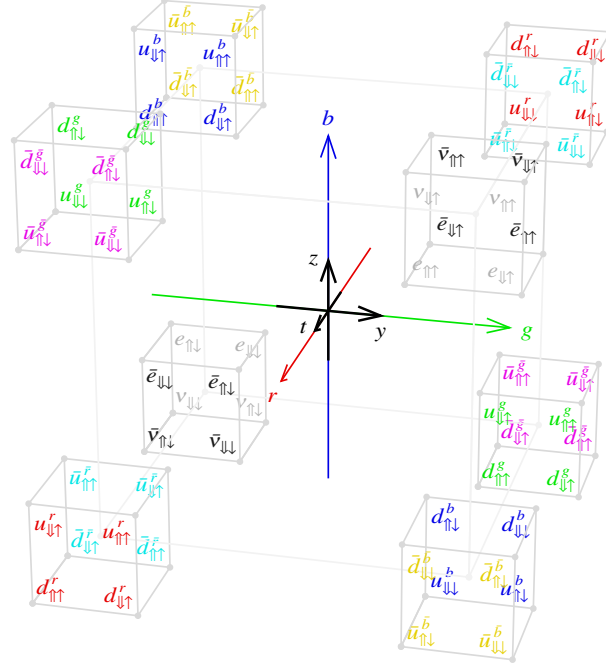


Figure 49.1 A generation (the electron generation) of 64 fermions arranged according to their $\text{Spin}(11,1)$ *tyzrgb* charges. This is similar to Figure 48.1, but with the addition of the *t*-bit, with *t*-bit-up directed out of the page. Whereas in the $\text{Spin}(10)$ Figure 48.1 each fermion was a 2-component Weyl fermion, here the Weyl components are distinguished by their Dirac boost ($\uparrow$ or $\downarrow$) and spin ($\uparrow$ or $\downarrow$) bits. A right-handed Weyl fermion has Dirac bits $\uparrow\uparrow$ or $\downarrow\downarrow$, while a left-handed Weyl fermion has Dirac bits $\downarrow\uparrow$ or $\uparrow\downarrow$. Flipping the *t*-bit of a fermion flips the fermion to its antifermionic partner of opposite boost and the same spin. Flipping all 6 *tyzrgb* bits of a fermion flips its Dirac boost and spin bits, thereby transforming the fermion to its Weyl companion.

still contain 2 components, one with *t*-bit up, the other with *t*-bit down, but those 2 components do not Lorentz-transform into each other.

At first sight it might seem that it would be impossible to distribute the two components of a Weyl spinor in separate entries, because the two components of a Weyl spinor necessarily have the same charge (being related by a proper Lorentz transformation), whereas different entries in the $\text{Spin}(10)$ chart (48.15) carry different charges. However, there is a trick that gets around that difficulty. The trick is suggested by the fourth tea leaf of §49.1, that the spinors of the $\text{Spin}(10)$ chart (48.15) are fermions (unbarred) or antifermions (barred) as the colour chirality $\varkappa_{rgb}$, equation (49.2), is positive or negative. The five $\text{Spin}(10)$ charges of a spinor are eigenvalues of the five diagonal bivectors (48.20). If these diagonal bivectors are modified by multiplying them by $\varkappa_{rgb}$, then their eigenvalues will measure the charge of the fermion, not the antifermion, in all entries of the $\text{Spin}(10)$ chart. A key point that allows this adjustment to be made consistently is that $\varkappa_{rgb}$ commutes with all standard-model bivectors. Notably, $\varkappa_{rgb}$ does not commute with $\text{SU}(5)$ bivectors

that transform between leptons and quarks; but that is fine, because SU(5) is not an unbroken symmetry of the standard model. A consistent way to implement this modification, that leaves the bivector algebra of the standard model (but not of SU(5)) unchanged, is to multiply all imaginary bivectors $\gamma_i^+ \gamma_j^-$ by $\varkappa_{rgb}$, while leaving all real bivectors $\gamma_i^+ \gamma_j^+$ and $\gamma_i^- \gamma_j^-$ unchanged,

$$\gamma_i^+ \gamma_j^- \rightarrow \gamma_i^+ \gamma_j^- \varkappa_{rgb}, \quad i, j = t, y, z, r, g, b. \quad (49.6)$$

Equivalently, replace the imaginary i in all Spin(10) multivectors by the colour pseudoscalar $-I_{rgb} = i\varkappa_{rgb}$, equation (49.2). Although at this point the modification (49.6) is needed only for bivectors in the Spin(10) algebra, for which $i, j = y, z, r, g, b$ (excluding t), it turns out that the electroweak and grand Higgs fields, equations (50.4) and (50.31), have the correct symmetry-breaking behaviour only if the modification (49.6) is extended to all Spin(11,1) bivectors, that is, for all of $i, j = t, y, z, r, g, b$.

The purpose of making the modification (49.6) was to allow Lorentz transformations to connect fermions across different entries of the Spin(10) chart (48.15), and that works, as will now be shown. The modification (49.6) serves to replace each antifermion in the chart with the corresponding fermion. For example, the positron entries $\bar{e}_R$ and $\bar{e}_L$ are replaced by electrons e_L and e_R . What about antifermions? Where have they gone? The answer is that antifermions are obtained from fermions in the usual way, by taking their complex conjugates and multiplying by the conjugation operator C , equation (49.5), $\bar{\psi} \equiv C\psi^*$. Thus antifermions appear in a second copy of the Spin(10) chart (48.15), a conjugated version in which all fermions are replaced by antifermions.

It requires some work, §49.2.2, to establish the correct assignment of Dirac boost ($\uparrow$ or $\downarrow$) and spin ($\uparrow$ or $\downarrow$) bits, but the end result is the following Spin(11,1) chart of spinors, arranged in columns by the number of Spin(10) up-bits as in the earlier chart (48.15):

0	1	2	3	4	5
$- : \begin{smallmatrix} \bar{\nu}_{\uparrow\downarrow} \\ \nu_{\downarrow\downarrow} \end{smallmatrix}$	$y : \begin{smallmatrix} \bar{\nu}_{\downarrow\downarrow} \\ \nu_{\uparrow\downarrow} \end{smallmatrix}$ $z : \begin{smallmatrix} \bar{e}_{\downarrow\downarrow} \\ e_{\uparrow\downarrow} \end{smallmatrix}$ $c : \begin{smallmatrix} d_{\uparrow\uparrow}^c \\ \bar{d}_{\downarrow\uparrow}^c \end{smallmatrix}$	$\bar{c} : \begin{smallmatrix} \bar{u}_{\uparrow\downarrow}^c \\ u_{\downarrow\downarrow}^c \end{smallmatrix}$ $yz : \begin{smallmatrix} \bar{e}_{\uparrow\downarrow} \\ e_{\downarrow\downarrow} \end{smallmatrix}$ $yc : \begin{smallmatrix} d_{\downarrow\uparrow}^c \\ \bar{d}_{\uparrow\uparrow}^c \end{smallmatrix}$ $zc : \begin{smallmatrix} u_{\downarrow\uparrow}^c \\ \bar{u}_{\uparrow\uparrow}^c \end{smallmatrix}$	$y\bar{c} : \begin{smallmatrix} \bar{u}_{\downarrow\downarrow}^c \\ u_{\uparrow\downarrow}^c \end{smallmatrix}$ $rgb : \begin{smallmatrix} e_{\uparrow\uparrow} \\ \bar{e}_{\downarrow\uparrow} \end{smallmatrix}$ $z\bar{c} : \begin{smallmatrix} \bar{d}_{\downarrow\downarrow}^c \\ d_{\uparrow\downarrow}^c \end{smallmatrix}$ $yzc : \begin{smallmatrix} u_{\uparrow\uparrow}^c \\ \bar{u}_{\downarrow\uparrow}^c \end{smallmatrix}$	$zrgb : \begin{smallmatrix} \nu_{\downarrow\uparrow} \\ \bar{\nu}_{\uparrow\uparrow} \end{smallmatrix}$ $yrgb : \begin{smallmatrix} e_{\downarrow\uparrow} \\ \bar{e}_{\uparrow\uparrow} \end{smallmatrix}$ $yz\bar{c} : \begin{smallmatrix} \bar{d}_{\uparrow\downarrow}^c \\ d_{\downarrow\downarrow}^c \end{smallmatrix}$	$yzrgb : \begin{smallmatrix} \nu_{\uparrow\uparrow} \\ \bar{\nu}_{\downarrow\uparrow} \end{smallmatrix}$

(49.7)

Whereas in the original Spin(10) chart (48.15) each entry was a 2-component Weyl spinor, in the Spin(11,1) chart (49.7) the 2 components of each Weyl spinor appear in bit-flipped entries. For example, the right-handed electron e_R of the original chart is replaced by $e_{\uparrow\uparrow}$, and its spatially rotated partner $e_{\downarrow\downarrow}$ of the same chirality appears in the all-bit-flipped entry. Each entry still has two components, but in the Spin(11,1) chart those two components differ by their t -bit: the upper component has t -bit up, the lower t -bit down. The net number of degrees of freedom remains the same, $2^6 = 64$. Flipping the t -bit transforms a fermion into its antifermionic partner of opposite boost. Flipping all bits except the t -bit transforms a fermion into its antifermionic partner of opposite spin. The Dirac boost bit in the Spin(11,1) chart (49.7) is $\uparrow$ or $\downarrow$ as

$\varkappa_{tyz}$ is positive or negative, that is, as the number of tyz up-bits is odd or even. The Dirac spin bit is $\uparrow$ or $\downarrow$ as $\varkappa_{rgb}$ is positive or negative, that is, as the number of rgb up-bits is odd or even. Whereas in the original $\text{Spin}(10)$ chart (48.15) a spinor was a fermion or antifermion as $\varkappa_{rgb}$ was positive or negative, in the $\text{Spin}(11,1)$ chart (49.7), a spinor is a fermion or antifermion as the time-colour chirality $\varkappa_{trgb}$ is positive or negative, that is, as the number of $trgb$ up-bits is even or odd.

A Dirac spinor has 4 complex components, for a total of 8 degrees of freedom. As previously in the $\text{Spin}(10)$ chart (48.15), each Dirac spinor in the $\text{Spin}(11,1)$ chart (49.7) has 8 components, each component representing one real degree of freedom. In any geometric algebra with one time dimension, conjugation flips all bits except the boost bit, Exercise 41.5 part 14. This is true in the Dirac algebra, equation (41.101), and it is also true in the $\text{Spin}(11,1)$ algebra, where the boost bit is the t -bit.

Figure 49.1 illustrates one generation (the electron generation) of fermions of the standard model arranged according to their $\text{Spin}(11,1)$ $tyzrgb$ charges.

The correctness of the assignment of Dirac boost and spin bits in the $\text{Spin}(11,1)$ chart (49.7), and the consistency of the entire construction, will now be established.

49.2.2 Justification of the $\text{Spin}(11,1)$ chart

The assignment of Dirac boost and spin bits for each individual species (electrons, for example) in the $\text{Spin}(11,1)$ chart (49.7) is determined by two conditions, that conjugation flips spin $\uparrow \leftrightarrow \downarrow$, while flipping the y -bit flips boost $\uparrow \leftrightarrow \downarrow$. The first condition holds because the expression (41.87) for the Dirac conjugation operator C in the chiral representation shows that Dirac conjugation flips chirality and spin, but not boost, equation (41.101). The second condition holds because, after electroweak symmetry breaking, flipping the y -bit flips spinors between right- and left-handed Dirac chiralities of the same species, for example $e_R \leftrightarrow e_L$. Massive spinors are linear combinations of the two chiralities. Since massive spinors have definite spin, either $\uparrow$ or $\downarrow$, flipping the y -bit must flip the Dirac boost bit while preserving the spin bit, for example, $e_{\uparrow\uparrow} \leftrightarrow e_{\downarrow\uparrow}$.

The two conditions suffice to determine the translation between Dirac and $\text{Spin}(11,1)$ spinors of the same species, but they do not fix the translation across different species. The translation across different species is determined by the condition that Lorentz transformations commute with standard-model transformations, in accordance with the Coleman-Mandula theorem.

The Dirac algebra contains two mutually commuting operators, $\gamma_0\gamma_3$ and $\gamma_1\gamma_2$, that respectively generate a boost and a spatial rotation without flipping any bits. The operator $\gamma_0\gamma_3$ generates a boost in the $\uparrow$ - $\downarrow$ boost plane: a boost by rapidity θ multiplies $\uparrow$ and $\downarrow$ spinors by a real number $e^{\pm\theta/2}$. The operator $\gamma_1\gamma_2$ generates a spatial rotation in the $\uparrow$ - $\downarrow$ spin plane: a rotation by angle θ multiplies $\uparrow$ and $\downarrow$ spinors by a phase $e^{\mp i\theta/2}$.

The $Cl(11,1)$ geometric algebra contains three mutually commuting operators that flip no bits, and at the same time commute with all standard-model transformations, namely the time pseudoscalar I_t , the weak pseudoscalar I_{yz} , and the colour pseudoscalar I_{rgb} (the definition (49.8c) essentially repeats the earlier

definition (49.2)),

$$I_t \equiv -i\gamma_t^+ \gamma_t^- = \varkappa_t \equiv \gamma_t \wedge \gamma_{\bar{t}} , \quad (49.8a)$$

$$I_{yz} \equiv \gamma_y^+ \gamma_y^- \gamma_z^+ \gamma_z^- = -\varkappa_{yz} \equiv -\gamma_y \wedge \gamma_{\bar{y}} \wedge \gamma_z \wedge \gamma_{\bar{z}} , \quad (49.8b)$$

$$I_{rgb} \equiv \gamma_r^+ \gamma_r^- \gamma_g^+ \gamma_g^- \gamma_b^+ \gamma_b^- = -i\varkappa_{rgb} \equiv -i\gamma_r \wedge \gamma_{\bar{r}} \wedge \gamma_g \wedge \gamma_{\bar{g}} \wedge \gamma_b \wedge \gamma_{\bar{b}} . \quad (49.8c)$$

The weak pseudoscalar I_{yz} changes sign when an odd number of yz bits are flipped, and, as remarked above, flipping an odd number of yz bits flips the Dirac boost bit. But I_{yz} is a spacelike operator (it generates a rotation by a phase), so cannot by itself generate a boost. On the other hand the time pseudoscalar I_t does generate a Lorentz boost, whose action is independent of the yz bits. Thus the combination I_{tyz} generates a Lorentz boost that acts oppositely on spinors of opposite weak chirality, consistent with the behaviour of the Dirac boost operator $\gamma_0\gamma_3$. The Spin(11,1) boost operator that can be identified with the Dirac boost operator $\gamma_0\gamma_3$ is

$$I_{tyz} \equiv -i\gamma_t^+ \gamma_t^- \gamma_y^+ \gamma_y^- \gamma_z^+ \gamma_z^- = -\varkappa_{tyz} \equiv -\gamma_t \wedge \gamma_{\bar{t}} \wedge \gamma_y \wedge \gamma_{\bar{y}} \wedge \gamma_z \wedge \gamma_{\bar{z}} . \quad (49.9)$$

It was remarked just before the definition (49.2) of colour chirality that Spin(10)-conjugate spinors have opposite colour chirality $\varkappa_{rgb}$. As remarked above, in the Dirac algebra, conjugation flips spin but not boost. Therefore the colour pseudoscalar I_{rgb} can be identified with the Dirac spin operator $\gamma_1\gamma_2$.

The product of the boost operator I_{tyz} and the spin operator I_{rgb} equals the 12-dimensional pseudoscalar J ,

$$\begin{aligned} J \equiv I_{tyzrgb} &= I_{tyz} \underset{\text{boost}}{I_{rgb}} = -i\gamma_t^+ \gamma_t^- \gamma_y^+ \gamma_y^- \gamma_z^+ \gamma_z^- \gamma_r^+ \gamma_r^- \gamma_g^+ \gamma_g^- \gamma_b^+ \gamma_b^- \\ &= i\varkappa_{tyzrgb} \equiv i\gamma_t \wedge \gamma_{\bar{t}} \wedge \gamma_y \wedge \gamma_{\bar{y}} \wedge \gamma_z \wedge \gamma_{\bar{z}} \wedge \gamma_r \wedge \gamma_{\bar{r}} \wedge \gamma_g \wedge \gamma_{\bar{g}} \wedge \gamma_b \wedge \gamma_{\bar{b}} . \end{aligned} \quad (49.10)$$

The 12-dimensional chiral operator $\varkappa_{tyzrgb}$ corresponding to the Dirac chiral operator $\gamma_5 = -iI$ is

$$\varkappa_{tyzrgb} = -iJ . \quad (49.11)$$

It is the 12-dimensional chiral operator $\varkappa_{tyzrgb}$ that should be identified with the Dirac chiral operator γ_5 , not the 10-dimensional chiral operator $\varkappa_{yzrgb}$ as suggested by the coincidence (49.1).

The final ingredient to complete the translation between Dirac and Spin(11,1) algebras is to identify an operator that connects the two Weyl components of each spinor in the Spin(11,1) chart (49.7) to each other. In the Dirac algebra, the two Weyl components are connected by flipping both bits. A Dirac operator that flips both bits is $I\sigma_2 = -\gamma_1\gamma_3$, equation (41.14), which is also the generator of a spatial rotation. The corresponding operator that flips all bits in the Spin(11,1) algebra is

$$J\sigma_2 \equiv \gamma_t^+ \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.12)$$

where J is the pseudoscalar (49.10). Equation (49.12) can be regarded as defining σ_2 . Below, equation (49.13e), σ_2 will be identified as a generator of a Lorentz boost. The expression (49.12) for $J\sigma_2$ coincides with that for the Spin(11,1) spinor metric ε , equation (49.4), but the two are not the same because $J\sigma_2$ transforms as a multivector whereas the spinor metric ε transforms as a (Lorentz-invariant) spinor tensor. The coincidence

of the expressions for $J\sigma_2$ and ε is similar to the coincidence (41.38) between $I\sigma_2$ and the spinor metric ε in the chiral representation of the Dirac algebra.

To qualify as a satisfactory operator that generates a spatial rotation, the operator $J\sigma_2$ defined by equation (49.12) must satisfy two conditions. First, consistent with the expected anticommutation of generators of spatial rotations, $J\sigma_2$ must anticommute with I_{rgb} , which was identified above as generating a spatial rotation. Second, in accordance with the Coleman-Mandula theorem, $J\sigma_2$ must commute with all standard-model bivectors modified per (49.6). Both conditions hold, and that should not be too surprising. Recall that the modification (49.6) of $\text{Spin}(10)$ bivectors was done precisely to enable the existence of an operator that connects the two Weyl components of a spinor by a bit-flip; and the two Weyl components of a spinor are related by a spatial rotation.

Note that whereas the generator $J\sigma_2$, equation (49.12), of a spatial rotation flips all six *tyzrgb* bits, and preserves standard-model charges, the $\text{Spin}(11,1)$ conjugation operator C , equation (49.5), which transforms a fermion into its antifermionic partner, flips all bits except *t*, and flips all standard-model charges, consistent with the assignment of fermions in the $\text{Spin}(11,1)$ chart (49.7).

49.2.3 The Dirac algebra as a subalgebra of the $Cl(11,1)$ geometric algebra

The Dirac algebra can now be expressed in terms of the $Cl(11,1)$ geometric algebra.

The generators I_{rgb} and $J\sigma_2$, equations (49.8c) and (49.12), and their product constitute a set of 3 anticommuting generators of spatial rotations that commute with all standard-model bivectors modified per (49.6). The pseudoscalar J is given by equation (49.10). The full set of 6 Lorentz generators, consisting of 3 spatial generators $J\sigma_a$ and 3 boost generators σ_a , is

$$J\sigma_1 = -\gamma_t^+ \gamma_y^+ \gamma_z^+ \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.13a)$$

$$J\sigma_2 = \gamma_t^+ \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.13b)$$

$$I_{rgb} = J\sigma_3 = \gamma_r^+ \gamma_r^- \gamma_g^+ \gamma_g^- \gamma_b^+ \gamma_b^- , \quad (49.13c)$$

$$\sigma_1 = i\gamma_t^- \gamma_y^- \gamma_z^- \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.13d)$$

$$\sigma_2 = i\gamma_t^- \gamma_y^- \gamma_z^- \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.13e)$$

$$I_{tyz} = \sigma_3 = -i\gamma_t^+ \gamma_t^- \gamma_y^+ \gamma_y^- \gamma_z^+ \gamma_z^- . \quad (49.13f)$$

The 6 Lorentz generators all have grade 6. They are not bivectors, but they nevertheless generate Lorentz transformations. The 8 basis elements of the complete Lie algebra of Lorentz transformations comprise the 6 Lorentz generators (49.13) along with the unit element 1 and the pseudoscalar J given by equation (49.10). The commutation rules of the elements of the Lie algebra are those of the Lorentz algebra. With the modification (49.6) to standard-model generators, all the Lorentz generators commute with all standard-model generators.

Given a time vector γ_0 and a set of generators σ_a of Lorentz boosts, spatial vectors γ_a can be deduced by Lorentz transforming γ_0 appropriately. Since the boost generators satisfy $\sigma_a = \gamma_0 \gamma_a$, spatial vectors satisfy $\gamma_a = -\gamma_0 \sigma_a$. With the time axis $\gamma_0 = i\gamma_t^-$ and the expressions (49.13) for σ_a , the full set of 4 orthonormal

spacetime vectors γ_m is

$$\gamma_0 = i\gamma_t^- , \quad (49.14a)$$

$$\gamma_1 = \gamma_y^- \gamma_z^- \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.14b)$$

$$\gamma_2 = \gamma_y^- \gamma_z^- \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.14c)$$

$$\gamma_3 = \gamma_t^+ \gamma_y^+ \gamma_g^- \gamma_z^+ \gamma_z^- . \quad (49.14d)$$

The vectors (49.14) all have grade 1 mod 4. The multiplication rules for the vectors γ_m given by equations (49.14) agree with the usual multiplication rules for Dirac γ -matrices: the vectors γ_m anticommute, and their scalar products form the Minkowski metric. All the spacetime vectors γ_m commute with all standard-model bivectors modified per (49.6). The Dirac pseudoscalar I coincides with the Spin(11, 1) pseudoscalar J defined by equation (49.10),

$$I \equiv \gamma_0 \gamma_1 \gamma_2 \gamma_3 = J . \quad (49.15)$$

Equivalently, the Dirac chiral operator $\gamma_5 \equiv -iI$ coincides with the Spin(11, 1) chiral operator $\varkappa_{tyzrgb} \equiv -iJ$.

Thus the Dirac and standard-model algebras are subalgebras of the Cl(11, 1) geometric algebra, such that all Dirac generators commute with all standard-model generators modified per (49.6), as was to be proved.

The time dimension (49.14a) is just a simple vector in the Spin(11, 1) algebra, but the 3 spatial dimensions (49.14b)–(49.14d) are all 5-dimensional. The spatial dimensions share a common 2-dimensional factor $\gamma_y^- \gamma_z^-$. Aside from that common factor, each of the 3 spatial dimensions is itself 3-dimensional: $\gamma_r^+ \gamma_g^+ \gamma_b^+$, $\gamma_r^- \gamma_g^- \gamma_b^-$, and $\gamma_t^+ \gamma_y^+ \gamma_z^+$.

49.3 Spatial readjustment of geometry following grand symmetry breaking

Bosonic string theory requires that the gauge group satisfy the Chan-Paton condition, §??, that it transform in a tensor representation of the spinor representation. If the spinors are those of Spin(11, 1), then the compact gauge group is not the group Spin(10) generated by the bivectors of Cl(10), but rather the larger gauge group $G^2(10) = U(16)$ generated by multivectors of grade $\hat{2} \equiv 2 \pmod{4}$, Table ??.

The generators of the gauge group U(16) are multivectors of grades 2, 6, 10. The grade-10 multivector is the 10-dimensional pseudoscalar I_{yzrgb} . The pseudoscalar I_{yzrgb} can be recognized as generating the abelian factor $U(1)_{yzrgb}$ of the gauge group,

$$G^2(10) = U(16) = U(1)_{yzrgb} \times SU(16) . \quad (49.16)$$

If the chirality of the gauge group coincides with spacetime chirality, as it does in the case of Spin(10), then the pseudoscalar I_{yzrgb} can consistently be interpreted as generating a spatial rotation in the transverse plane, leaving SU(16) as the internal compact gauge group. The transverse spatial rotation group $U(1)_{yzrgb}$ commutes with the internal gauge group SU(16), consistent with the Coleman-Mandula no-go theorem (Coleman and Mandula, 1967; Mandula, 2015).

The conclusion that a transverse spatial rotation is generated by the 10-dimensional pseudoscalar I_{yzrgb}

appears to be at odds with the earlier conclusion that a transverse spatial rotation is generated by the 6-dimensional colour pseudoscalar I_{rgb} , equation (49.13c). The two pictures can be reconciled provided that the gauge group $SU(16)$ required by bosonic string theory is interpreted as the primordial gauge group prior to grand symmetry breaking, and that at some point there is an adjustment in the configuration of 3+1 large spacetime dimensions, with I_{yzrgb} and I_{rgb} generating a transverse spatial rotation respectively before and after that point.

When did the readjustment of large spacetime dimensions occur? According to the minimal symmetry breaking chain (50.1) found in Chapter 50, $SU(16)$ breaks to $SU(8)_R \times SU(8)_L$, which then breaks to the Pati-Salam group,

$$SU(16) \rightarrow SU(8)_R \times SU(8)_L \rightarrow Spin(4)_w \times Spin(6)_c . \quad (49.17)$$

The generators of $SU(8)_R \times SU(8)_L$ commute with the Dirac vectors (49.20) but not (49.14), while the generators of $Spin(4)_w \times Spin(6)_c$ commute with both sets of Dirac vectors. Therefore the reconfiguration of 3+1 large spacetime dimensions could occur at $SU(8)$ symmetry breaking, where $SU(8)_R \times SU(8)_L$ breaks to $Spin(4)_w \times Spin(6)_c$, but it could also have happened at a later point. The spatial spin of a spinor, $\uparrow$ or $\downarrow$, is the eigenvalue of the generator of a transverse spatial rotation. The earlier argument that I_{rgb} generated a spatial rotation rested in part on the fact that right- and left-handed fermions of the same species (e.g. electrons) differ in their y -bit, and that at electroweak symmetry breaking those massless right- and left-handed components combined into massive Dirac fermions of definite spin. Therefore flipping the y -bit flipped chirality but not spin. Thus spin could have been the eigenvalue of I_{rgb} , which is unchanged when the y -bit is flipped, but not the eigenvalue of I_{yzrgb} , which changes sign when the y -bit is flipped.

Thus the readjustment of large spacetime dimensions could have happened at any time between $SU(8)$ symmetry breaking and electroweak symmetry breaking. Readjustment at $SU(8)$ symmetry breaking is plausible because there is no other marker of that symmetry breaking: the grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), breaks $SU(16)$ to $SU(8)_R \times SU(8)_L$, and $Spin(10)$ to $Spin(4)_w \times Spin(6)_c$, but it does not directly mediate the transition from $SU(8)_R \times SU(8)_L$ to $Spin(4)_w \times Spin(6)_c$. An alternative possibility is that the readjustment of large spacetime dimensions occurred at electroweak symmetry breaking, which is intriguing because that energy scale is accessible to experiment, though it is not clear what the experimental signature might be.

The readjustment of large spacetime dimensions, whenever it happened, will here be called spatial readjustment.

The reconfiguration of the 3+1 large spacetime dimensions leaves the internal $yzrgb$ charges of a spinor unchanged. The reconfiguration also leaves the $Cl(11,1)$ chirality of a spinor unchanged, since the product of the boost and spin operators before spatial readjustment is the same $Cl(11,1)$ pseudoscalar J as after spatial readjustment equation (49.10),

$$J \equiv I_{tyzrgb} = \underset{\text{boost}}{I_t} \underset{\text{rotation}}{I_{yzrgb}} . \quad (49.18)$$

The construction of the full Lorentz algebra before spatial readjustment proceeds similarly to that after spatial readjustment. To ensure that all charges of a spinor are invariant under Lorentz transformations, each generator of a Lorentz transformation must preserve the orientation of each bit (up or down) relative to the t -bit (so guaranteeing that the charge is the same as it is in the t -bit-up frame). The two commuting

generators I_t and I_{yzrgb} have this property by virtue of leaving all bits unchanged. Other Lorentz generators must flip some bits. If a Lorentz generator flips the t -bit, then it must simultaneously flip all bits. Such a generator must have a full set of 6 charges (including t -charge), and must therefore be a multivector of grade 6. Each i -factor of the multivector must flip both i -bit up to i -bit down and vice versa, so must be a linear combination of both chiral vectors γ_i and γ_i^- ; with a suitable choice of phase, each i -factor is γ_i^+ or γ_i^- . A suitable Lorentz generator that flips all bits is the same generator $J\sigma_2$, equation (49.12), as after spatial readjustment.

The full set of Lorentz generators of the 3+1 large spacetime dimensions before spatial readjustment is

$$J\sigma_1 = \gamma_t^+ \gamma_y^- \gamma_z^- \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.19a)$$

$$J\sigma_2 = \gamma_t^+ \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.19b)$$

$$I_{yzrgb} = J\sigma_3 = \gamma_y^+ \gamma_y^- \gamma_z^+ \gamma_z^- \gamma_r^+ \gamma_r^- \gamma_g^+ \gamma_g^- \gamma_b^+ \gamma_b^- , \quad (49.19c)$$

$$\sigma_1 = i\gamma_t^- \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.19d)$$

$$\sigma_2 = i\gamma_t^- \gamma_y^- \gamma_z^- \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.19e)$$

$$I_t = \sigma_3 = -i\gamma_t^+ \gamma_t^- . \quad (49.19f)$$

The generators $J\sigma_i$ and σ_i satisfy the usual Lie algebra of Lorentz transformations in 3+1 spacetime dimensions. The Lorentz generators (49.19) all have grade $\hat{2}$. The Lorentz generators (49.19) before spatial readjustment is can be comared to those (49.13) after spatial readjustment.

The 3+1 Dirac vectors γ_m , $m = 0, 1, 2, 3$, that yield the Lorentz generators (49.19) can be deduced from $\sigma_a = \gamma_0 \gamma_a$, $a = 1, 2, 3$, and are

$$\gamma_0 = i\gamma_t^- , \quad (49.20a)$$

$$\gamma_1 = \gamma_y^+ \gamma_z^+ \gamma_r^+ \gamma_g^+ \gamma_b^+ , \quad (49.20b)$$

$$\gamma_2 = \gamma_y^- \gamma_z^- \gamma_r^- \gamma_g^- \gamma_b^- , \quad (49.20c)$$

$$\gamma_3 = \gamma_t^+ . \quad (49.20d)$$

The Dirac vectors (49.20) all have grade $\hat{1}$. The Dirac vectors (49.20) satisfy the Dirac algebra: they anticommute, and their scalar products form the Minkowski metric. The Dirac vectors (49.20) before spatial readjustment can be compared to those (49.14) before spatial readjustment.

The Coleman-Mandula theorem requires that the Dirac algebra commute with the internal symmetry algebra, the algebra of the subgroup $SU(16)$ of $G^2(10) = U(10)$. But whereas real grade $\hat{2}$ multivectors (those having an even number of imaginary factors γ_i^- and an even number of real factors γ_i^+) commute with all Dirac vectors (49.20), imaginary grade $\hat{2}$ multivectors (those having an odd number of imaginary factors γ_i^- and an odd number of real factors γ_i^+) anticommute with the Dirac vectors γ_1 and γ_2 .

The solution to the problem is to adjust all imaginary generators γ_A of $SU(16)$ by multiplying them by the chiral operator $\varkappa_{yzrgb}$, while leaving all real generators unchanged:

$$\begin{array}{ccc} \gamma_A & \rightarrow & \gamma_A , \\ \text{real} & & \\ \gamma_A & \rightarrow & \gamma_A \varkappa_{yzrgb} , \\ \text{imaginary} & & \end{array} \quad (49.21)$$

which is similar to the modification (49.6) of imaginary bivectors after spatial readjustment, but with the

modification factor $\varkappa_{yzrgb}$ in place of $\varkappa_{rgb}$. So adjusted, all the generators of $SU(16)$ commute with all Dirac vectors (49.20), as required by Coleman-Mandula. The nature and necessity for the adjustment (49.21) can be deduced from the following argument. A spinor in its own t -bit-up frame is ϵ_a with the t -bit of a up. In a Lorentz-transformed frame, the spinor becomes a complex linear combination of itself ϵ_a and its all-bit-flip $\epsilon_{\bar{a}}$ with t -bit down. A rotation in the i -plane generated by $\gamma_i^+ \gamma_i^-$ rotates ϵ_a and $\epsilon_{\bar{a}}$ by opposite phases, as if they had opposite i -charges, whereas in fact they have the same charge. The solution is to adjust the generator to $\gamma_i^+ \gamma_i^- \varkappa_{yzrgb}$, which rotates ϵ_a and $\epsilon_{\bar{a}}$ by the same phase. The chirality operator $\varkappa_{yzrgb}$ commutes with all the original generators of $SU(16)$, so leaves the algebra of $SU(16)$ unchanged. An alternative adjustment might be to multiply all imaginary $SU(16)$ generators by the time chiral operator $\varkappa_t$, but the resulting adjusted generators would all anticommute with the Dirac vectors (49.20), rather than commuting as they should.

To summarize, after the adjustment (49.21) is made to the generators of the gauge group $SU(16)$, the gauge algebra and the Dirac algebra are commuting subalgebras of the $Cl(11,1)$ algebra, as required by Coleman-Mandula. In this fashion the 10 compact internal dimensions and the 3+1 large spacetime dimensions accommodate into a spacetime of dimension 11+1.

One of the puzzles that emerged from the analysis of Hamilton and McMaken (2023) concerned the 4 time-weak bivectors $\gamma_t^+ \gamma_i^\pm$, $i = y, z$, which Hamilton and McMaken (2023) argued mediated electroweak symmetry breaking. The problem is that those 4 time-weak bivectors did not commute with the grand Higgs scalar $\langle T \rangle$, equation (50.31), that mediates grand symmetry breaking, so those bivectors could not have generated any gauge symmetry before grand symmetry breaking, because if they did, then after grand symmetry breaking the associated four spin 1 gauge bosons would combine by the Goldstone mechanism (Goldstone, Salam, and Weinberg, 1962) with four spin 0 Higgs bosons having the same symmetry, whereas those 4 scalar bosons apparently persisted to produce electroweak symmetry breaking. The only way that Hamilton and McMaken (2023) were able to make sense of this puzzle was to postulate that the primordial $Spin(11,1)$ symmetry was somehow broken to $Spin(10,1)$ already prior to grand symmetry breaking, so that γ_t^+ did not participate in any gauge symmetry.

The analysis of this section 49.3 provides an elegant solution to the puzzle, which is that the longitudinal dimension γ_t^+ was already part of the 3+1 large spacetime dimensions, equation (49.20d), and was not available to be part of any internal gauge symmetry.

Symmetry breaking and the Higgs mechanism

The previous Chapter 49 established that the standard-model and Dirac algebras are commuting subalgebras of the $Cl(11, 1)$ geometric algebra. The Chan-Paton condition of bosonic string theory, §??, requires that the internal gauge group before grand symmetry breaking was $SU(16)$, rather than $Spin(10)$. Section 49.3 established that the $SU(16)$ and Dirac algebras are also commuting subalgebras of the $Cl(11, 1)$ geometric algebra, although the relation between Dirac and $Cl(11, 1)$ algebras was different before (49.20) versus after (49.14) a spatial readjustment that could have taken place at any time between grand and electroweak symmetry breaking.

The present chapter discovers the minimal symmetry breaking chain (50.1) that breaks $SU(16)$ to today's low-energy gauge group, and the minimal Higgs field $\mathbf{E}$, equation (50.2), that mediates the symmetry breaking.

A gauge field is a Lorentz vector because it arises as a connection in a gauge-covariant spacetime derivative. A Higgs field on the other hand must be a Lorentz scalar, since otherwise it would break the Lorentz symmetry of the vacuum (it would impose a preferred spatial direction and/or rest frame), contrary to observation. For a gauge field, a scalar Lagrangian can be constructed from the commutator of the gauge-covariant derivative. For a Higgs field, any generator that commutes with Lorentz symmetries defines a Lorentz scalar that could play the role of a scalar field in the Lagrangian. Whereas gauge fields must transform according to the adjoint representation of the group, Higgs generators could potentially transform in any representation of the group. However, in the $Cl(11, 1)$ model, the minimal Higgs sector, equation (50.2), turns out to transform in the adjoint representation, just like the gauge fields.

The general principles underlying symmetry breaking by the Higgs mechanism (Englert and Brout, 1964; Higgs, 1964) are: (1) the Higgs field before symmetry breaking is a scalar (spin 0) multiplet of the unbroken symmetry; (2) one component of the Higgs multiplet acquires a nonzero vacuum expectation value, and that component is invariant (uncharged, a singlet) under the symmetries that remain after symmetry breaking; (3) components of the Higgs multiplet whose symmetry is broken are absorbed into longitudinal components of the broken gauge (spin 1) fields by the Goldstone mechanism (Goldstone, Salam, and Weinberg, 1962), giving those gauge fields mass; and (4) unbroken components of the Higgs multiplet persist as scalar fields, potentially available to mediate the next level of symmetry breaking.

There is a substantial literature on possible symmetry breaking chains from $Spin(10)$ (Harvey, Ramond,

and Reiss, 1980; del Aguila and Ib'añez, 1981; Di Luzio, 2011; Dueck and Rodejohann, 2013; Ajaib, Gogoladze, and Shafi, 2013; Altarelli, 2013; Altarelli and Meloni, 2013; Babu and Khan, 2015; Jarkovská et al., 2022; Furey and Hughes, 2022; Krasnov, 2022a). The big difference between the Higgs sector in the present chapter and those in the literature is the presence in the $\text{Cl}(11, 1)$ algebra of a 6th bit, the t -bit, and the associated two extra dimensions, the time dimension $\gamma_0 = -i\gamma_t^-$ and its longitudinal spatial partner γ_t^+ . Specifically, the grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), is proportional to $-i\gamma_t^+ \gamma_t^-$, modified per (49.6), while the electroweak Higgs field $\langle \mathbf{H} \rangle$, equation (50.4), is proportional to $\gamma_t^+ \gamma_y^-$, modified per (49.6). The choices of grand and electroweak Higgs fields $\langle \mathbf{T} \rangle$ and $\langle \mathbf{H} \rangle$ are not ad hoc. They possess the properties required to accomplish grand, §50.4, and electroweak, §50.3, symmetry breaking, and they are the only elements of the $\text{Cl}(11, 1)$ algebra with those properties.

In the standard approach to spontaneous symmetry breaking in $\text{Spin}(10)$ (Croon et al., 2019), the Higgs field must be part of a $\text{Spin}(10)$ multiplet in order that its Lagrangian be invariant under $\text{Spin}(10)$. The standard approach is premised on the assumption that the Dirac and $\text{Spin}(10)$ algebras commute, which is not true in the present construction; rather, the Dirac and standard-model algebras here are commuting subalgebras of the $\text{Cl}(11, 1)$ geometric algebra. In the standard approach the electroweak Higgs field must be an odd $\text{Spin}(10)$ multivector (because it flips only 1 bit, the y -bit), so must be a vector or pseudovector (dimension 10), trivector or pseudotrivector (dimension 120), or pentavector (2 possibilities, a pentavector or a pseudopentavector, each of dimension 126). In $\text{Cl}(11, 1)$, a multiplet $\mathbf{H}$ of fields, equation (50.3), with properties matching the electroweak Higgs multiplet is present without having to be introduced ad hoc.

50.1 The minimal symmetry breaking chain

In the minimal symmetry breaking chain from $\text{SU}(16)$, there prove to be four stages of symmetry breaking:

$$\begin{array}{ccccccc} \text{SU}(16) & \xrightarrow{10^{16} \text{ GeV?}} & \text{SU}(8)_R \times \text{SU}(8)_L & \xrightarrow{10^{15} \text{ GeV}} & \text{Spin}(4)_w \times \text{Spin}(6)_c & \xrightarrow{10^{12} \text{ GeV}} & \\ & & \text{U}(1)_Y \times \text{SU}(2)_L \times \text{SU}(3)_c & \xrightarrow{160 \text{ GeV}} & \text{U}(1)_Q \times \text{SU}(3)_c & . & \end{array} \quad (50.1)$$

The top line is the prediction, the bottom line the standard model. The group $\text{Spin}(4)_w \times \text{Spin}(6)_c$ is the Pati-Salam group (Pati and Salam, 1974). The associated minimal Higgs field that mediates the symmetry breaking chain (50.1) is a scalar field transforming in the adjoint representation of the gauge group. Denote this Higgs multiplet $\mathbf{E}$, after Englert and Brout (1964), who proposed the Higgs mechanism at the same time as Higgs (1964),

$$\mathbf{E} \equiv E^A \gamma_A , \quad (50.2)$$

where γ_A are multivectors of grade $\hat{2} \equiv 2 \pmod{4}$ in $\text{Cl}(11, 1)$, and all imaginary bivectors are to be understood as being multiplied by $\varkappa_{yzrgb}$ or $\varkappa_{rgb}$ per the modifications (49.21) or (49.6) respectively before and after spatial readjustment. The modifications (49.21) or (49.6) are necessary to ensure that all unbroken fields commute with all Lorentz bivectors (49.19) or (49.13) respectively before and after spatial readjustment.

A feature of a Higgs field transforming in the adjoint representation is that every massless gauge boson has

a matching scalar Higgs field available to give the gauge boson mass at symmetry breaking, offering some flexibility in the identity of the grand unified gauge group.

It is possible that there is a more complicated symmetry breaking chain than (50.1), but the chain (50.1) is the one that emerges without contrivance. The chain (50.1) is the minimal symmetry breaking chain, and the Higgs field $\mathbf{E}$ defined by equation (50.2) is the associated minimal Higgs multiplet.

50.2 Constraints on symmetry breaking

Symmetry breaking from SU(16) is constrained by the condition that the spacetime and internal algebras commute at every step of unification, as required by the Coleman-Mandula theorem. For example, although the Dirac vectors γ_m given by either equations (49.14) or (49.20) all commute with all the bivector generators of the $U(1)_Y \times SU(2)_L \times SU(3)_c$ standard model, modified per (49.6) or (49.21), those Dirac vectors γ_m do *not* commute with general SU(5) transformations. As long as Dirac vectors γ_m are given by either of equations (49.14) or (49.20), and spacetime transformations commute with internal rotations, SU(5) cannot be an internal symmetry.

An exhaustive search establishes that the largest subgroup of Spin(10) whose bivector generators, modified per (49.6), all commute with the spacetime vectors (49.14) is the Pati-Salam group $\text{Spin}(4)_w \times \text{Spin}(6)_c$, in which the generators of the weak group $\text{Spin}(4)_w$ are the 6 bivectors formed from antisymmetric products of the four weak vectors γ_i^+ and γ_j^- with i and j running over y, z , while the generators of the colour group $\text{Spin}(6)_c$ are the 15 bivectors formed from antisymmetric products of the six colour vectors γ_i^+ and γ_j^- with i and j running over r, g, b . If the group is enlarged from Spin(10) to SU(16), then there is an additional commuting hexavector generator I_{rgb} , whose modified form is $I_{rgb}\mathcal{Z}_{rgb} = -i$, but that generator is proportional to the scalar multivector 1, and generates no additional symmetry.

If the algebra of bivectors is enlarged to include those of Spin(11, 1), then there are an additional 4 bivectors that commute with the spacetime vectors (49.14), namely the products $\gamma_t^+ \gamma_i^\pm$, $i = y, z$ of the 11th spatial dimension γ_t^+ with the 4 weak vectors $\gamma_i^\pm$, $i = y, z$, modified per (49.6). The 4 bivectors $\gamma_t^+ \gamma_i^\pm$, $i = y, z$, modified per (49.6), call attention to themselves because they match the properties of the Lorentz-scalar electroweak Higgs multiplet $\mathbf{H}$, equation (50.3).

50.3 Electroweak symmetry breaking

The 4 bivector generators $\gamma_t^+ \gamma_i^\pm$ with $i = y, z$ call attention to themselves because they transform spinors by one unit of standard-model charge y or z , whereas the $6 + 15 = 21$ bivector generators of the Pati-Salam (1974) group $\text{Spin}(4)_w \times \text{Spin}(6)_c$ transform spinors by an even number of standard-model charges $yzrgb$. The electroweak Higgs field responsible for breaking y -symmetry carries one unit of y -charge. The electroweak Higgs field gives masses to fermions by flipping their y -bit.

The Weinberg (1967) theory of electroweak symmetry breaking requires the electroweak Higgs field to be part of a multiplet of 4 fields that transform into each other under $U(1)_Y \times SU(2)_L$. Indeed the 4 bivector

generators $\gamma_t^+ \gamma_i^\pm$ with $i = y, z$ provide such a set of fields. The 4-component Higgs field $\mathbf{H}$ may be defined by

$$\mathbf{H} \equiv H^{i\pm} \gamma_t^+ \gamma_i^\pm, \quad i = y, z, \quad (50.3)$$

where the imaginary bivectors $\gamma_t^+ \gamma_i^\pm$, $i = y, z$, are to be understood as being multiplied by $\varkappa_{rgb}$ per the modification (49.6). Electroweak symmetry breaking occurs when the Higgs field acquires a vacuum expectation value $\langle \mathbf{H} \rangle$ proportional to $\gamma_t^+ \gamma_y^-$,

$$\langle \mathbf{H} \rangle = \langle H \rangle \gamma_t^+ \gamma_y^- \varkappa_{rgb}, \quad (50.4)$$

in which the modification factor $\varkappa_{rgb}$ has been included explicitly to avoid possible misunderstanding. When combined with the time axis $-i\gamma_0 \equiv \gamma_t^-$ in a fermion mass term $\bar{\psi} \cdot \langle \mathbf{H} \rangle \psi = -i\psi^\dagger \gamma_0 \langle \mathbf{H} \rangle \psi$, the vacuum Higgs field (50.4) yields a Dirac mass term

$$\gamma_t^- \langle \mathbf{H} \rangle = \langle H \rangle \gamma_t^- \gamma_t^+ \gamma_y^- \varkappa_{rgb}, \quad (50.5)$$

which indeed flips the y -bit as it should. The vacuum Higgs field (50.4) is proportional to γ_y^- not γ_y^+ because the factor γ_y^- in the mass term (50.5) preserves the spinor identity, whereas γ_y^+ flips between spinor and antispinor, in much the same way that in the Dirac algebra the time axis γ_0 maps massive spinors and antispinors to themselves, whereas the time axis' boost partner γ_3 flips between massive spinor and massive antispinor, equations (14.102). Note that whereas the Higgs multiplet $\mathbf{H}$, equation (50.3), commutes with the time axis γ_0 , the pseudoscalar multiplet $J\mathbf{H}$ does not, so it is $\langle \mathbf{H} \rangle$ and not $J\langle \mathbf{H} \rangle$ that generates a Dirac mass.

The electroweak Higgs field (50.3) possesses $\text{Spin}(4)_w$ rotational symmetry in the 4-dimensional space of y, z vectors. The manner in which the $\text{Spin}(4)_w$ symmetry of the electroweak Higgs field is broken to the electroweak symmetry $U(1)_Y \times \text{SU}(2)_L$ is noted in §50.6.2, equation (50.58).

50.3.1 Masses of electroweak gauge bosons

This subsection 50.3.1 is essentially a standard exposition of electroweak symmetry breaking by the Higgs mechanism, adapted to the present notation; see for example Peskin and Schroeder (1995, Ch. 20) for a pedagogical account. The main result is that electroweak symmetry breaking generates masses m_Z and m_W , equations (50.20), for the electroweak Z and $W^\pm$ gauge bosons, and those masses are related, equations (50.14), to the coupling parameters g_Y and g_w of $U(1)_Y$ and $\text{SU}(2)_L$. The experimental confirmation (Arnison et al., 1983) that the mass ratio m_Z/m_W agreed with prediction was one of the great successes of the standard model.

To avoid overloading the notation, all the gauge and Higgs generators in this section 50.3.1 are treated as bivectors, whereas correctly they should be modified per (49.6), by multiplying all imaginary bivectors by the colour chiral operator $\varkappa_{rgb}$. The operator $\varkappa_{rgb}$ commutes with all standard-model bivectors, so its exclusion makes no difference to the algebra.

For a spinor field ψ , the gauge-covariant derivative with respect to $U(1)_Y \times \text{SU}(2)_L$ transformations is

$$D_m \psi = (\partial_m + g_Y \mathbf{B}_m + g_w \mathbf{W}_m) \psi, \quad (50.6)$$

where $\mathbf{B}_m$ and $\mathbf{W}_m$ are the $U(1)_Y$ and $SU(2)_L$ gauge fields

$$\mathbf{B}_m \equiv iB_m Y, \quad \mathbf{W}_m \equiv iW_m^i \tau_i, \quad (50.7)$$

and g_Y and g_w are dimensionless coupling strengths for those fields. Here iY , equation (48.49), is the generator of the hypercharge symmetry $U(1)_Y$, while $i\tau_i$ with τ_i satisfying the Pauli algebra, equations (48.45), are the 3 generators of the left-handed weak group $SU(2)_L$. The weak Pauli matrix τ_3 acting on a spinor has eigenvalue equal to twice the weak isospin $2I_L = z - y$, equation (48.10b). The electromagnetic charge generator iQ , equation (48.50), is related to the hypercharge and weak generators iY and $i\tau_3$ by, equation (48.14),

$$Q = \frac{1}{2}(Y + \tau_3). \quad (50.8)$$

The sum $W_m^i \tau_i$ in the gauge field $\mathbf{W}_m$, equation (50.7), can be expressed with respect to either an orthonormal or a chiral basis,

$$W_m^i \tau_i = W_m^1 \tau_1 + W_m^2 \tau_2 + W_m^3 \tau_3 = W_m^+ \tau_+ + W_m^- \tau_- + W_m^3 \tau_3, \quad (50.9)$$

where $W_m^\pm \equiv (W_m^1 \mp iW_m^2)/\sqrt{2}$, and the chiral Pauli operators $\tau_\pm$ are

$$\tau_+ \equiv \frac{\tau_1 + i\tau_2}{\sqrt{2}} = \frac{\boldsymbol{\gamma}_z \wedge \boldsymbol{\gamma}_y}{\sqrt{2}}, \quad \tau_- \equiv \frac{\tau_1 - i\tau_2}{\sqrt{2}} = \frac{\boldsymbol{\gamma}_y \wedge \boldsymbol{\gamma}_z}{\sqrt{2}}. \quad (50.10)$$

The operator τ_+ increases z -charge by 1 and decreases y -charge by 1, and therefore carries +1 unit of each of electric charge Q and weak isospin I_L . Conversely, τ_- decreases z -charge by 1 and increases y -charge by 1, and therefore carries -1 unit of each of electric charge Q and weak isospin I_L . The operators Y and τ_3 leave y - and z -charge unchanged, so carry zero electric charge Q and weak isospin I_L .

Introduce the weak mixing, or Weinberg, angle θ_w defined by

$$\sin \theta_w \equiv \frac{g_Y}{g}, \quad \cos \theta_w \equiv \frac{g_w}{g}, \quad g \equiv \sqrt{g_Y^2 + g_w^2}. \quad (50.11)$$

Define the electromagnetic and weak fields A_m and Z_m to be the orthogonal linear combinations of B_m and W_m^3 ,

$$\begin{pmatrix} A_m \\ Z_m \end{pmatrix} \equiv \begin{pmatrix} \cos \theta_w & \sin \theta_w \\ -\sin \theta_w & \cos \theta_w \end{pmatrix} \begin{pmatrix} B_m \\ W_m^3 \end{pmatrix}. \quad (50.12)$$

In terms of the rotated electromagnetic and weak fields A_m and Z_m , the electroweak gauge connection is

$$g_Y \mathbf{B}_m + g_w \mathbf{W}_m = i \left(2e A_m Q + 2g Z_m (I_L - \sin^2 \theta_w Q) + g_w (W_m^+ \tau_+ + W_m^- \tau_-) \right), \quad (50.13)$$

where the electromagnetic coupling e is

$$e = \frac{g_Y g_w}{g} = g_Y \cos \theta_w = g_w \sin \theta_w = g \cos \theta_w \sin \theta_w. \quad (50.14)$$

The electromagnetic coupling e is related to the fine-structure constant α by $e^2 = 4\pi\alpha$.

The gauge-covariant derivative of the 4-component Higgs field $\mathbf{H}$ with respect to $U(1)_Y \times SU(2)_L$ transformations is

$$D_m \mathbf{H} = \partial_m \mathbf{H} + g_Y [\mathbf{B}_m, \mathbf{H}] + g_w [\mathbf{W}_m, \mathbf{H}]. \quad (50.15)$$

Whereas in the covariant derivative (50.6) of a spinor ψ , the fields $\mathbf{B}_m$ and $\mathbf{W}_m$ act directly on the spinor, in the covariant derivative (50.15) of the Higgs field $\mathbf{H}$, the fields act as a commutator, because whereas a spinor ψ transforms as $\psi \rightarrow R\psi$ under a rotor R (an element of the group $U(1)_Y \times SU(2)_L$), a multivector such as the Higgs field $\mathbf{H}$ transforms as $\mathbf{H} \rightarrow R\mathbf{H}\tilde{R}$, with $\tilde{R} = R^{-1}$ the reverse of R . The covariant derivative of the expectation value (50.4) of the Higgs field is

$$D_m \langle \mathbf{H} \rangle = \langle H \rangle (g_Y B_m [iY, \gamma_t^+ \gamma_y^-] + g_w W_m^i [i\tau_i, \gamma_t^+ \gamma_y^-]) . \quad (50.16)$$

The relevant commutators of the generators iY of $U(1)_Y$, equation (48.49), and $i\tau_i$ of $SU(2)_L$, equations (48.45), with the electroweak Higgs field $\gamma_t^+ \gamma_y^-$ are

$$[iY, \gamma_t^+ \gamma_y^-] = \gamma_y^+ \gamma_y^- \gamma_t^+ \gamma_y^- = \gamma_t^+ \gamma_y^+ , \quad (50.17a)$$

$$[i\tau_1, \gamma_t^+ \gamma_y^-] = -\gamma_y^- \gamma_z^+ \gamma_t^+ \gamma_y^- = \gamma_t^+ \gamma_z^+ , \quad (50.17b)$$

$$[i\tau_2, \gamma_t^+ \gamma_y^-] = -\gamma_y^- \gamma_z^- \gamma_t^+ \gamma_y^- = \gamma_t^+ \gamma_z^- , \quad (50.17c)$$

$$[i\tau_3, \gamma_t^+ \gamma_y^-] = -\gamma_y^+ \gamma_y^- \gamma_t^+ \gamma_y^- = -\gamma_t^+ \gamma_y^+ . \quad (50.17d)$$

With the commutators (50.17), the covariant derivative (50.16) of the expectation value of the Higgs field becomes

$$\begin{aligned} D_m \langle \mathbf{H} \rangle &= \langle H \rangle \left((g_Y B_m - g_w W_m^3) \gamma_t^+ \gamma_y^+ + g_w (W_m^1 \gamma_t^+ \gamma_z^+ + W_m^2 \gamma_t^+ \gamma_z^-) \right) \\ &= \langle H \rangle \left(-g Z_m \gamma_t^+ \gamma_y^+ + g_w (W_m^+ \gamma_t^+ \gamma_z + W_m^- \gamma_t^+ \gamma_{\bar{z}}) \right) . \end{aligned} \quad (50.18)$$

Notice that the expression (50.18) for the covariant derivative $D_m \langle \mathbf{H} \rangle$ does not depend on the electromagnetic field A_m , which stems from the fact that $\langle \mathbf{H} \rangle$ commutes with the electric charge operator Q . The kinetic term in the scalar Higgs Lagrangian is $-\frac{1}{2} (D^m \tilde{\mathbf{H}}) \cdot (D_m \mathbf{H})$, where $\tilde{\mathbf{H}}$ is the reverse of $\mathbf{H}$. When the Higgs field acquires a nonzero expectation value $\langle \mathbf{H} \rangle$, it contributes to the Lagrangian a kinetic term proportional to (abbreviating $Z^m Z_m = (Z_m)^2$ and so forth)

$$(D^m \tilde{\langle \mathbf{H} \rangle}) \cdot (D_m \langle \mathbf{H} \rangle) = \langle H \rangle^2 \left(g^2 (Z_m)^2 + g_w^2 ((W_m^+)^2 + (W_m^-)^2) \right) . \quad (50.19)$$

The contribution (50.19) to the Lagrangian has the form of mass-squared terms for the Z_m and $W_m^\pm$ electroweak fields. The Higgs field thus generates masses m_Z and m_W for the Z_m and $W_m^\pm$ fields,

$$m_Z \equiv \langle H \rangle g , \quad m_W \equiv \langle H \rangle g_w . \quad (50.20)$$

The electromagnetic field A_m remains massless. In accordance with the remarks after equations (50.10), the electromagnetic field A_m and weak field Z_m both carry zero electric charge and weak isospin, while the weak fields $W_m^\pm$ carry respectively ± 1 unit of each of electric charge Q and weak isospin I_L .

50.3.2 Electroweak Higgs Lagrangian

The Lagrangian L_H of the 4-component Higgs field $\mathbf{H}$ is

$$L_H = -\frac{1}{2} (D^m \tilde{\mathbf{H}}) \cdot (D_m \mathbf{H}) - V(\tilde{\mathbf{H}} \cdot \mathbf{H}) , \quad (50.21)$$

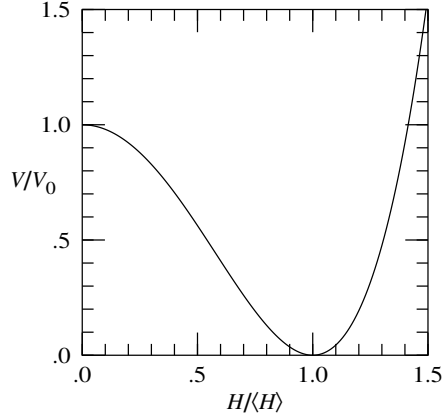


Figure 50.1 Mexican hat quartic potential V of a Higgs field of magnitude H .

where $D_m \mathbf{H}$ is the gauge-covariant derivative (50.15) of the Higgs field, and $V(\widetilde{\mathbf{H}} \cdot \mathbf{H})$ is a potential energy, a function of the scalar product $H^2 \equiv \widetilde{\mathbf{H}} \cdot \mathbf{H}$ of the Higgs field $\mathbf{H}$ and its reverse $\widetilde{\mathbf{H}}$. The potential energy V is postulated to have a minimum at a nonzero value of H^2 , which serves to make it energetically favourable for the Higgs field to acquire a nonzero expectation value. A commonly adopted potential, with the virtue of yielding a renormalizable quantum field theory, is a “Mexican hat” quartic, illustrated in Figure 50.1,

$$V(H^2) = \rho_H - \frac{1}{4}m_H^2 H^2 + \frac{1}{8}\lambda H^4. \quad (50.22)$$

The Higgs field $\mathbf{H}$ has units of mass, and the potential V has units of mass⁴, or energy density. The constant $V(0) = \rho_H$ looks like a vacuum density that could play the role of a cosmological constant before electroweak symmetry breaking, when $H = 0$. The quantity m_H proves to be the mass of the Higgs boson, equation (50.30). The minimum of the potential V defines the vacuum expectation value $\langle H \rangle$ of the magnitude of the Higgs field,

$$\langle H \rangle = \sqrt{\frac{m_H^2}{\lambda}}. \quad (50.23)$$

Having acquired a nonzero expectation value, the reconfigured 4-component Higgs field generates a single massive spin 0 particle, the Higgs boson. As emphasized above, central to the behaviour of the Higgs field is that its components rotate into each other under $U(1)_Y \times SU(2)_L$ transformations, equations (50.17). Therefore the Higgs field $\mathbf{H}$ can be written as a product of a unitary rotation U in $U(1)_Y \times SU(2)_L$ and a Higgs field $\mathbf{H}_0$ pointed in a certain direction, which can be taken to be the broken direction (50.4),

$$\mathbf{H} \equiv U \mathbf{H}_0, \quad \mathbf{H}_0 \equiv H \boldsymbol{\gamma}_y^- \boldsymbol{\gamma}_t^+. \quad (50.24)$$

By definition, the gauge-covariant derivative of the Higgs field transforms under the unitary rotation U as

$$D_m \mathbf{H} = U D_m \mathbf{H}_0. \quad (50.25)$$

The Higgs Lagrangian (50.21) is, by construction, invariant under gauge transformations, so in terms of the (un)rotated Higgs fields $\mathbf{H}_0$ is

$$L_H = -\frac{1}{2}(D^m \widetilde{\mathbf{H}}_0) \cdot (D_m \mathbf{H}_0) - V(\widetilde{\mathbf{H}}_0 \cdot \mathbf{H}_0) . \quad (50.26)$$

Define the perturbation h of the magnitude H of the Higgs field by

$$h \equiv H - \langle H \rangle . \quad (50.27)$$

In terms of h , the potential $V(H^2)$, equation (50.22), in the Higgs Lagrangian (50.26) is (note that $\lambda = m_H^2/\langle H \rangle^2$ from equation (50.23))

$$V = \frac{1}{2}m_H^2 h^2 \left(1 + \frac{h}{2\langle H \rangle}\right)^2 . \quad (50.28)$$

The potential vanishes at $h = 0$ provided that the constant term in equation (50.22) is $\rho_H = \frac{1}{8}m_H^2\langle H \rangle^2$. The Higgs Lagrangian (50.26) in terms of the perturbation h is

$$L_H = -\frac{1}{2} \left((\partial^m h)(\partial_m h) + \left(m_Z^2 (Z_m)^2 + m_W^2 ((W_m^+)^2 + (W_m^-)^2) \right) \left(1 + \frac{h}{\langle H \rangle}\right)^2 + m_H^2 h^2 \left(1 + \frac{h}{2\langle H \rangle}\right)^2 \right) . \quad (50.29)$$

If the Z_m and $W_m^\pm$ mass terms are set aside, then to lowest order in h the Lagrangian (50.29) looks like the Lagrangian of a free scalar field of mass m_H ,

$$L_H = -\frac{1}{2} \left((\partial^m h)(\partial_m h) + m_H^2 h^2 \right) . \quad (50.30)$$

The interpretation is that h describes a spin 0 field of mass m_H , the Higgs boson. Other terms proportional to powers of h in the Higgs Lagrangian (50.29) describe interactions between the Higgs boson h and the weak gauge fields, and self interactions of the Higgs boson.

50.4 Grand symmetry breaking

50.4.1 Grand Higgs field

As remarked in §50.2, the largest subgroup of $\text{Spin}(10)$ (or of $\text{SU}(16)$) that commutes with all the Dirac vectors (49.14), and can therefore be on the path to grand unification while the Dirac vectors are given by (49.14), is the Pati-Salam group $\text{Spin}(4)_w \times \text{Spin}(6)_c$.

According to the general rules governing symmetry breaking by a Higgs field, the grand Higgs field must acquire a nonzero expectation value that (1) is a Lorentz scalar, commuting with all Lorentz bivectors (49.13), (2) commutes with all generators of the unbroken group, and (3) fails to commute with other generators. There is just one multivector of the $\text{Cl}(11,1)$ algebra that fits the aspiration to be the grand Higgs field, namely the time bivector $\boldsymbol{\gamma}_t^+ \boldsymbol{\gamma}_t^-$ multiplied by the colour chiral operator $\boldsymbol{\varkappa}_{rgb}$,

$$\langle \mathbf{T} \rangle = -i \langle T \rangle \boldsymbol{\gamma}_t^+ \boldsymbol{\gamma}_t^- \boldsymbol{\varkappa}_{rgb} , \quad (50.31)$$

the imaginary i coming from the time vector being timelike, $\gamma_0 = i\gamma_t^-$. The grand Higgs field (50.31) (1) commutes with all Lorentz bivectors (49.13) and is therefore a Lorentz scalar, (2) commutes with the 21 Pati-Salam $\text{Spin}(4)_w \times \text{Spin}(6)_c$ bivectors modified per (49.6), and (3) fails to commute with the 24 modified bivectors of $\text{Spin}(10)$ that are not Pati-Salam bivectors. Although $\langle \mathbf{T} \rangle$ is a Lorentz scalar, commuting with all Lorentz bivectors (49.13), it anticommutes with all Dirac vectors (49.14), but that is fine because, unlike a gauge field, a Higgs field need only be a Lorentz scalar, not a Dirac scalar. In particular, $\langle \mathbf{T} \rangle$ fails to commute with the time axis $\gamma_0 = i\gamma_t^-$, which has the happy side effect of generating a Majorana mass for the right-handed neutrino, §50.4.4.

The primordial internal gauge group prior to grand symmetry breaking is not $\text{Spin}(10)$, but rather the subgroup $\text{SU}(16)$ of $G^2(10)$. The grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), preserves the generators of $\text{SU}(16)$ that commute with I_{rgb} , and breaks the generators of $\text{SU}(16)$ that fail to commute with I_{rgb} , breaking $\text{SU}(16)$ to

$$\text{SU}(16) \rightarrow \text{U}(1)_{rgb} \times \text{SU}(8)_R \times \text{SU}(8)_L . \quad (50.32)$$

where $\text{U}(1)_{rgb}$ is the group generated by I_{rgb} . The generators of $\text{SU}(8)_R \times \text{SU}(8)_L$ do not commute with the Dirac vectors (49.14), but they do commute with the Dirac vectors (49.20). To ensure compliance with the Coleman-Mandula theorem, the Dirac vectors in the $\text{SU}(8)_R \times \text{SU}(8)_L$ regime must be given by equations (49.20). The grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), commutes with all Lorentz generators both before and after spatial readjustment, equations (49.19) and (49.13), so is a Lorentz scalar in all regimes, as it should be.

The grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), is a bivector modified per (49.6), just like the electroweak Higgs field (50.4), and just like the gauge bivectors of the standard model discussed in §49.2.1. According to the general rules, the grand Higgs field $\langle \mathbf{T} \rangle$ must be a component of a multiplet of the unbroken symmetry, and the gauge (vector) and Higgs (scalar) fields corresponding to the symmetries that are broken by the grand Higgs field must transform consistently, so that the scalars and vectors can combine into massive gauge bosons. These requirements point to the conclusion that the Higgs field is a scalar field transforming in the adjoint representation, as anticipated in equation (50.2).

The $6 + 15 = 21$ bivectors of $\text{Spin}(4)_w \times \text{Spin}(6)_c$, namely $\gamma_i^\pm \gamma_j^\pm$ with i and j both in y, z , or i and j both in r, g, b modified per (49.6), commute with the grand Higgs field (50.31) and with the Dirac vectors (49.14), and therefore yield unbroken scalar and gauge fields. $\text{Spin}(4)_w \times \text{Spin}(6)_c$ preserves time, weak, and colour chiralities $\varkappa_t$, $\varkappa_{yz}$, and $\varkappa_{rgb}$, equations (49.8), so $\text{Spin}(4)_w$ and $\text{Spin}(6)_c$ gauge bosons allow transformations between fermions of the same boost ($\varkappa_{tyz}$) and spin ($\varkappa_{rgb}$),

$$\text{SU}(2)_R : \quad \nu_{\uparrow\uparrow} \leftrightarrow e_{\uparrow\uparrow} , \quad u_{\uparrow\uparrow}^c \leftrightarrow d_{\uparrow\uparrow}^c , \quad (50.33a)$$

$$\text{Spin}(6)_c/\text{SU}(3)_c : \quad \nu_{\uparrow\uparrow} \leftrightarrow u_{\uparrow\uparrow}^c , \quad e_{\uparrow\uparrow} \leftrightarrow d_{\uparrow\uparrow}^c . \quad (50.33b)$$

The $\text{SU}(2)_R$ transitions (50.33a) are right-handed versions of analogous $\text{SU}(2)_L$ left-handed transitions in the standard model. The $\text{Spin}(6)_c/\text{SU}(3)_c$ transitions (50.33b) and their left-handed equivalents transform between leptons and quarks, for which reason the associated gauge bosons are called leptiquarks.

The group $\text{SU}(8)_R \times \text{SU}(8)_L$ has $2 \times 63 = 126$ generators, Appendix 50.8, of which the $126 - 21 = 105$

generators fail to commute with the grand Higgs field (50.31). These 105 ... The 26 scalar bivectors combine with the 24 vector bivectors by the Goldstone mechanism to yield 24 massive charged gauge bosons,

The $4 \times 6 = 24$ bivectors of $\text{Spin}(10)/(\text{Spin}(4)_w \times \text{Spin}(6)_c)$, namely $\gamma_i^\pm \gamma_j^\pm$ with i in y, z , and j in r, g, b , modified per (49.6), fail to commute with the grand Higgs field (50.31), and they also fail to commute with some Lorentz bivectors (49.13). The transitions preserve time and $\text{Spin}(10)$ chiralities $\varkappa_t$ and $\varkappa_{yrgb}$, but break the individual weak and colour chiralities:

$$\begin{array}{ccccccc} \bar{\nu}_{\downarrow\downarrow} & \leftrightarrow & u_{\uparrow\uparrow}^c & \leftrightarrow & \bar{u}_{\downarrow\downarrow}^c & \leftrightarrow & \nu_{\uparrow\uparrow} \\ \bar{e}_{\downarrow\downarrow} & & d_{\uparrow\uparrow}^c & & \bar{d}_{\downarrow\downarrow}^c & & e_{\uparrow\uparrow} \end{array} . \quad (50.34)$$

The 24 scalar bivectors combine with the 24 vector bivectors by the Goldstone mechanism to yield 24 massive charged gauge bosons, which include 12 bivectors in $\text{SU}(5)$, with charges $i\bar{j}$ or $\bar{i}j$, and 12 bivectors not in $\text{SU}(5)$, with charges ij or $\bar{i}\bar{j}$. Observations and experiment show no sign of Lorentz violation (Mattingly, 2005). To avoid a violation of Lorentz invariance, the gauge bosons must presumably acquire a mass of the order of the grand unified scale.

50.4.2 Cosmological inflation

A leading idea of the standard model of cosmology is that inflation in the early universe was driven by the energy associated with grand unification, Chapter 38.

The grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), is available to drive inflation. The mechanism of grand symmetry breaking by a single Higgs field is consistent with the simplest conventional models of inflation.

50.4.3 Gauge boson masses at grand symmetry breaking

Subsection 50.4.1 argued that grand symmetry breaking breaks 24 of the 45 bivectors of $\text{Spin}(10)$, generating 24 massive gauge bosons.

The $\text{Spin}(10)$ covariant derivative of the expectation value $\langle \mathbf{T} \rangle$, equation (50.31), of the grand Higgs field is

$$D_m \langle \mathbf{T} \rangle = g \sum_A [C_m, \langle \mathbf{T} \rangle] = -i \langle \mathbf{T} \rangle g \sum_A C_m^A [\gamma_A, \gamma_t^+ \gamma_t^- \varkappa_{rgb}] , \quad (50.35)$$

where g is a grand coupling parameter, and $C_m = C_m^A \gamma_A$ are the gauge fields associated with the 45 $\text{Spin}(10)$ bivectors γ_A modified per (49.6). The nonvanishing commutators of the $\text{Spin}(10)$ bivectors with the grand Higgs generator $\gamma_t^+ \gamma_t^- \varkappa_{rgb}$ are the 24 commutators

$$[\gamma_i^\pm \gamma_j^\pm, \gamma_t^+ \gamma_t^- \varkappa_{rgb}] = 2 \gamma_i^\pm \gamma_j^\pm \gamma_t^+ \gamma_t^- \varkappa_{rgb} \quad (i \text{ in } y, z, j \text{ in } r, g, b) . \quad (50.36)$$

The bivector factors $\gamma_i^\pm \gamma_j^\pm$ should strictly be modified per (49.6), but this modification can be omitted here because $\gamma_t^+ \gamma_t^-$ commutes with $\varkappa_{rgb}$, so the commutation rules (50.36) are unaffected by the possible extra factor of $\varkappa_{rgb}$ on $\gamma_i^\pm \gamma_j^\pm$. When the grand Higgs field acquires a nonzero expectation value $\langle \mathbf{T} \rangle$, it contributes

to the Lagrangian a kinetic term proportional to

$$(D^m \langle \tilde{\mathbf{T}} \rangle) \cdot (D_m \langle \mathbf{T} \rangle) = 4 \langle T \rangle^2 g^2 \sum_A (C_m^A)^2, \quad (50.37)$$

in which the sum over A is over noncommuting gauge fields $\mathbf{C}_m$. The contribution (50.37) takes the form of a mass-squared term for each of the noncommuting fields, so the grand Higgs field $\langle \mathbf{T} \rangle$ generates a mass

$$m_T = 2 \langle T \rangle g \quad (50.38)$$

for each of the 24 noncommuting gauge fields.

It was noted in §50.4.1 that the 24 bivectors $\gamma_i^\pm \gamma_j^\pm$ with i in y, z , and j in r, g, b , modified per (49.6), fail to commute not only with the grand Higgs field $\langle \mathbf{T} \rangle$, but also with the Lorentz bivectors (49.13), so the 24 bivectors do not form Lorentz-covariant multiplets, and the covariant derivative D_m of these bivectors in equation (50.35) cannot be the 3+1 dimensional Lorentz-covariant derivative associated with the Dirac vectors γ_m of equations (49.14). The resolution of this issue is that the spacetime vectors before grand symmetry breaking were given by equations (49.20) and not by (49.14). Before grand symmetry breaking, the 24 bivectors $\gamma_i^\pm \gamma_j^\pm$, modified per (49.21), commute with all the Lorentz bivectors (49.19), and the covariant derivative D_m in equation (50.35) should be interpreted as the Lorentz-covariant derivative associated with the vectors (49.20) before grand symmetry breaking.

50.4.4 The grand Higgs field generates a Majorana mass for the right-handed neutrino

The grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), has the felicitous property that it is able to generate a Majorana mass for the right-handed neutrino, which allows the left-handed neutrino to acquire a small mass via the see-saw mechanism, §48.4.1. The Majorana mass term is

$$\bar{\nu} \cdot \langle \mathbf{T} \rangle \nu = -i \nu^\dagger \gamma_0 \langle \mathbf{T} \rangle \nu = \nu^\dagger \gamma_t^- \langle \mathbf{T} \rangle \nu, \quad (50.39)$$

which flips the right-handed neutrino's t -bit. The right-handed neutrino, the all-Spin(10)-bits-up entry in the Spin(11,1) chart (49.7), has two components, a right-handed neutrino $\nu_{\uparrow\uparrow}$ with t -bit up, and a left-handed antineutrino partner $\bar{\nu}_{\downarrow\uparrow}$ with t -bit down. The two components have opposite boost, but the same spin, so flipping the t -bit generates a mass for the right-handed neutrino. Only the right-handed neutrino can acquire mass this way, because only the right-handed neutrino has zero couplings to the gauge groups $U(1)_Y \times SU(2)_L \times SU(3)_c$ of the standard model. For other fermions, conservation of standard-model charges prohibits flipping the t -bit.

Although the grand Higgs field $\langle \mathbf{T} \rangle$ acquires its nonzero expectation value at grand symmetry breaking, it is only later, at Pati-Salam symmetry breaking, §50.6, that the grand Higgs field can generate a Majorana mass for the right-handed neutrino. Before Pati-Salam symmetry breaking, Spin(4)_w × Spin(6)_c transitions (50.33) unite right-handed neutrinos with other right-handed leptons and quarks. Spin(4)_w × Spin(6)_c transitions preserve chirality, and likewise Lorentz transformations preserve chirality, so before Pati-Salam symmetry

breaking all fermions are massless chiral fermions, including the right-handed neutrino, which is part of the 16-component right-handed lepton-quark multiplet $\{\nu_R, e_R, u_R^c, d_R^c\}$.

After Pati-Salam symmetry breaking into the standard model, the right-handed $\text{Spin}(4)_w \times \text{Spin}(6)_c$ transitions (50.33) no longer occur. The right-handed neutrino becomes isolated into a singlet of the standard model, and standard-model symmetries do not prohibit the Majorana mass term (50.39) from flipping the right-handed neutrino's t -bit.

The nonvanishing charges of the right-handed neutrino $\nu_{\uparrow\uparrow}$ and of its left-handed antineutrino partner $\bar{\nu}_{\downarrow\downarrow}$ are $\rho_3 = +1$ and -1 respectively, and $B-L = -1$ and $+1$ respectively. Thus the Majorana mass term (50.39) violates baryon-minus-lepton $B-L$ symmetry. According to the standard picture of leptogenesis (Buchmüller, Di Bari, and Plümacher, 2005), out-of-equilibrium CP -violating decay of an initially thermal distribution of massive right-handed Majorana neutrinos can generate a nonzero $B-L$ number, which sphaleron processes then transform into a nonzero baryon number.

In the usual $\text{Spin}(10)$ model, the standard Higgs used to break $B-L$ so as to generate a Majorana mass is a pentavector (Harvey, Ramond, and Reiss, 1980; del Aguila and Ib'añez, 1981; King et al., 2021), belonging to the dimension $\binom{10}{5} = 126 + \overline{126}$ representation of $\text{Spin}(10)$. Specifically, the pentavector

$$\gamma_y \gamma_z \gamma_r \gamma_g \gamma_b + \gamma_{\bar{y}} \gamma_{\bar{z}} \gamma_{\bar{r}} \gamma_{\bar{g}} \gamma_{\bar{b}} \quad (50.40)$$

has the property that it flips between the all-bit-up right-handed neutrino and its all-bit-down left-handed antineutrino partner in the $\text{Spin}(10)$ chart (48.15), $\uparrow\uparrow\uparrow\uparrow \leftrightarrow \downarrow\downarrow\downarrow\downarrow$, while yielding zero acting on any other spinor. The pentavector (50.40) commutes with $\text{SU}(3)_c$ but not with $B-L$, so explicitly breaks $B-L$ symmetry.

The Higgs pentavector (50.40) that generates a Majorana mass in $\text{Spin}(10)$ does not work for the $\text{Spin}(11, 1)$ model because the modification (49.6) that allows $\text{Spin}(10)$ and $\text{Spin}(3, 1)$ to unify into $\text{Spin}(11, 1)$ puts the left-handed antineutrino partner $\bar{\nu}_{\downarrow\downarrow}$ of the right-handed neutrino $\nu_{\uparrow\uparrow}$ into the same $\text{SU}(5)$ column of the $\text{Spin}(11, 1)$ chart (49.7), instead of into the opposite $\text{SU}(5)$ columns of the $\text{Spin}(10)$ chart (48.15). In $\text{Spin}(11, 1)$, the multivector $\gamma_t^- \langle \mathbf{T} \rangle$ that transforms between $\nu_{\uparrow\uparrow}$ and $\bar{\nu}_{\downarrow\downarrow}$ has $\text{Spin}(10)$ grade 0, not 5.

50.5 $\text{SU}(8)$ symmetry breaking

As seen in §50.4.1, $\text{SU}(8)$ symmetry breaking from $\text{SU}(8)_R \times \text{SU}(8)_L$ to the Pati-Salam group is not mediated by a Higgs field. Rather, $\text{SU}(8)$ symmetry breaking is associated with a reconfiguration of the 3+1 large spacetime dimensions. Specifically, the longitudinal spatial Dirac companion γ_3 of the time dimension γ_0 changes from being a plain grade 1 vector γ_t^+ of $\text{Cl}(11, 1)$, equation (49.20d) to a grade 5 vector of $\text{Cl}(11, 1)$, comprising γ_t^+ multiplied by the weak pseudoscalar I_{yz} , equation (49.14d).

50.6 Pati-Salam symmetry breaking

50.6.1 The Higgs field that breaks Pati-Salam symmetry

According to the general rules, the Higgs field that breaks the Pati-Salam $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry must be part of a multiplet of Lorentz-scalar fields that rotate into each other under $\text{Spin}(4)_w \times \text{Spin}(6)_c$. Now the grand Higgs field $\langle \mathbf{T} \rangle$, equation (50.31), that mediates grand symmetry breaking is part of a 66-component scalar multiplet $\mathbf{E}$, equation (50.2), of $\text{Spin}(11,1)$ bivectors modified per (49.6). After grand symmetry breaking to the Pati-Salam group, the scalar multiplet $\mathbf{E}$ contains an unbroken 21-component Pati-Salam multiplet generated by the $6 + 15 = 21$ modified bivectors of $\text{Spin}(4)_w \times \text{Spin}(6)_c$. This scalar Pati-Salam multiplet could potentially contain a component $\mathbf{P}$ that, when it acquires an expectation value $\langle \mathbf{P} \rangle$, breaks the Pati-Salam symmetry. Indeed it does.

The Higgs field $\mathbf{P}$ that breaks the Pati-Salam symmetry must leave the standard-model group unbroken, so must commute with standard-model bivectors modified per (49.6). The weak subgroup $\text{SU}(2)_L$ permutes yz charges while preserving the total charge $y+z$, while the colour subgroup $\text{SU}(3)_c$ permutes rgb charges while preserving the total charge $r+g+b$. Therefore the Higgs field $\mathbf{P}$ must be a linear combination of generators that are invariant under permutations of each of yz and rgb . There are two such basis bivectors in $\text{Spin}(4)_w \times \text{Spin}(6)_c$, namely the third Pauli generator $i\rho_3$ that generates the $\text{U}(1)_R$ subgroup of the right-handed weak group $\text{SU}(2)_R$, equation (48.46c); and the baryon minus lepton operator $i(B-L)$ that generates the $\text{U}(1)_c$ subgroup of the colour group $\text{Spin}(6)_c$, equation (48.48). The corresponding conserved charges are (the expression for $B-L$ repeats equation (48.12))

$$\rho_3 = y + z, \quad B - L = -\frac{2}{3}(r + g + b). \quad (50.41)$$

If the Higgs field $\mathbf{P}$ lies inside the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ submultiplet of $\mathbf{E}$, then $\mathbf{P}$ must be a linear combination of the bivectors $i\rho_3$ and $i(B-L)$ modified per (49.6),

$$\mathbf{P} = i(P^3 \rho_3 + P_m^{B-L}(B-L)) \kappa_{rgb}. \quad (50.42)$$

The factor κ_{rgb} is from the modification (49.6).

When the Higgs doublet $\mathbf{P}$, equation (50.42), acquires a vacuum expectation value $\langle \mathbf{P} \rangle$, it must leave unbroken the hypercharge group $\text{U}(1)_Y$ of the standard model, which is generated by the bivector iY , equation (48.49). The hypercharge Y is a linear combination of the charges ρ_3 and $B-L$, equation (48.10a),

$$Y = \rho_3 + B - L. \quad (50.43)$$

When the Higgs doublet $\mathbf{P}$ acquires its vacuum expectation value, it condenses the two phases R and $B-L$ into one,

$$\text{U}(1)_R \times \text{U}(1)_{B-L} \rightarrow \text{U}(1)_Y. \quad (50.44)$$

This is similar to the way that electroweak symmetry breaking involves condensing the two phases of hypercharge and weak isospin into one, $\text{U}(1)_Y \times \text{U}(1)_L \rightarrow \text{U}(1)_Q$. The symmetry breaking (50.44) breaks the Pati-Salam group $\text{Spin}(4)_w \times \text{Spin}(6)_c$ to the standard-model group $\text{U}(1)_Y \times \text{SU}(2)_L \times \text{SU}(3)_c$.

The linear combination of $i\rho_3$ and $iB-L$ to which $\mathbf{P}$ breaks can be deduced from the condition that the unbroken symmetry is $U(1)_Y$. The $\text{Spin}(4)_w \times \text{Spin}(6)_c$ gauge connection is

$$g_r \mathbf{R}_m + g_w \mathbf{W}_m + g_c \mathbf{C}_m , \quad (50.45)$$

where $\mathbf{R}_m$, $\mathbf{W}_m$, and $\mathbf{C}_m$ are respectively gauge fields of $SU(2)_R$, $SU(2)_L$, and $\text{Spin}(6)_c$, and g_r , g_w , and g_c are the respective dimensionless coupling parameters. The part of the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ connection (50.45) associated with the $U(1)_R \times U(1)_{B-L}$ symmetry is

$$i (g_r R_m^3 \rho_3 + g_c C_m^{B-L} (B-L)) \varkappa_{rgb} , \quad (50.46)$$

where R_m^3 and C_m^{B-L} are the corresponding connection coefficients. Analogously to the weak mixing angle θ_w , equations (50.11), define the $SU(2)_R \times \text{Spin}(6)_c$ mixing angle θ_r by

$$\sin \theta_r \equiv \frac{g_r}{g} , \quad \cos \theta_r \equiv \frac{g_c}{g} , \quad g \equiv \sqrt{g_r^2 + g_c^2} . \quad (50.47)$$

And analogously to the rotated fields A_m and Z_m of electroweak theory, equation (50.12), define gauge fields B_m and P_m to be orthogonal linear combinations of R_m^3 and C_m^{B-L} ,

$$\begin{pmatrix} B_m \\ P_m \end{pmatrix} \equiv \begin{pmatrix} \cos \theta_r & \sin \theta_r \\ -\sin \theta_r & \cos \theta_r \end{pmatrix} \begin{pmatrix} R_m^3 \\ C_m^{B-L} \end{pmatrix} . \quad (50.48)$$

In terms of the rotated fields B_m and P_m , the $U(1)_R \times U(1)_{B-L}$ connection (50.46) is (a commuting factor of $\varkappa_{rgb}$ from the modification (49.6) has been dropped for brevity from both sides of equation (50.49))

$$i (g_r R_m^3 \rho_3 + g_c C_m^{B-L} (B-L)) = i (g_Y B_m Y - g P_m (\sin^2 \theta_r \rho_3 - \cos^2 \theta_r (B-L))) , \quad (50.49)$$

where the hypercharge coupling parameter g_Y is

$$g_Y \equiv g_r \cos \theta_r = g_c \sin \theta_r = g \cos \theta_r \sin \theta_r . \quad (50.50)$$

The term proportional to $B_m Y$ in the connection (50.49) has the correct form for the unbroken $U(1)_Y$ hypercharge connection. The remaining term on the right hand side of equation (50.49) represents the broken direction, the direction along which the Higgs doublet $\mathbf{P}$ acquires an expectation value $\langle \mathbf{P} \rangle$,

$$\langle \mathbf{P} \rangle = i \langle P \rangle (\sin^2 \theta_r \rho_3 - \cos^2 \theta_r (B-L)) \varkappa_{rgb} , \quad (50.51)$$

where the factor $\varkappa_{rgb}$ from the modification (49.6) has been restored.

50.6.2 Masses of Pati-Salam gauge bosons

For simplicity, factors of $\varkappa_{rgb}$ from the modification (49.6) are omitted throughout this subsection (50.6.2). The factor $\varkappa_{rgb}$ commutes with all the generators, so has no effect on the predictions (50.56) of masses of Pati-Salam gauge bosons.

The $\text{Spin}(4)_w \times \text{Spin}(6)_c$ covariant derivative of the expectation value $\langle \mathbf{P} \rangle$ of the Higgs doublet $\mathbf{P}$, equation (50.51), is

$$\begin{aligned} D_m \langle \mathbf{P} \rangle &= \langle P \rangle (g_r \sin^2 \theta_r [\mathbf{R}_m, i\rho_3] - g_c \cos^2 \theta_r [\mathbf{C}_m, i(B-L)]) \\ &= g \langle P \rangle (\sin^3 \theta_r [\mathbf{R}_m, i\rho_3] - \cos^3 \theta_r [\mathbf{C}_m, i(B-L)]) , \end{aligned} \quad (50.52)$$

where $\mathbf{R}_m$ and $\mathbf{C}_m$ are the gauge fields of respectively $\text{SU}(2)_R$ and $\text{Spin}(6)_c$,

$$\mathbf{R}_m = iR_m^i \rho_i , \quad \mathbf{C}_m = C_m^{[i\pm j\pm]} \boldsymbol{\gamma}_i^\pm \boldsymbol{\gamma}_j^\pm \quad (i, j \text{ in } r, g, b) . \quad (50.53)$$

The nonvanishing commutators of the Pati-Salam $\text{Spin}(4)_w \times \text{Spin}(6)_c$ gauge fields with the component generators ρ_3 and $B-L$ of the expectation value $\langle \mathbf{P} \rangle$ of the Higgs doublet are

$$[\rho_2, \rho_3] = 2i\rho_1 , \quad (50.54a)$$

$$[\rho_1, \rho_3] = -2i\rho_2 , \quad (50.54b)$$

$$[\tfrac{1}{2}(1 + \varkappa_{ij}) \boldsymbol{\gamma}_i^+ \boldsymbol{\gamma}_j^+, B-L] = -\tfrac{2}{3}(1 + \varkappa_{ij}) \boldsymbol{\gamma}_i^+ \boldsymbol{\gamma}_j^- \quad (i, j \text{ in } r, g, b) , \quad (50.54c)$$

$$[\tfrac{1}{2}(1 + \varkappa_{ij}) \boldsymbol{\gamma}_i^+ \boldsymbol{\gamma}_j^-, B-L] = \tfrac{2}{3}(1 + \varkappa_{ij}) \boldsymbol{\gamma}_i^+ \boldsymbol{\gamma}_j^+ \quad (i, j \text{ in } r, g, b) . \quad (50.54d)$$

The right-handed Pauli matrices ρ_i are given by equations (48.46). The leptoquark bivectors of $\text{Spin}(6)_c$ in equations (50.54c) and (50.54d) are given in expanded form by equations (48.43).

The scalar product of each commutator of equations (50.54) with its reverse equals 2 for the top two lines, and $2 \times (\frac{2}{3})^2$ for the bottom two lines. The square of the covariant derivative (50.52) is then

$$(D^m \langle \tilde{\mathbf{P}} \rangle) \cdot (D_m \langle \mathbf{P} \rangle) = g^2 \langle P \rangle^2 \left(2 \sin^6 \theta_r (R_m^i)^2 + 2 \left(\tfrac{2}{3}\right)^2 \cos^6 \theta_r (C_m^{[i\pm j\pm]})^2 \right) , \quad (50.55)$$

where the fields on the right hand side are the subset of the weak and colour fields that fail to commute with the Higgs field $\langle \mathbf{P} \rangle$. The two right-handed weak fields R_m^i , $i = 1, 2$, carry two units of y, z charge and zero colour charge r, g, b ; they transform right-handed leptons and quarks into their right-handed weak partners, transformations (50.33a) and their antiparticle equivalents. The six leptoquark fields $C^{[i\pm j\pm]}$ carry zero y or z charge, and two units of r, g, b charge; they are called leptoquarks because they transform between leptons and quarks, transformations (50.33b) and their antiparticle equivalents. Equation (50.55) shows that the Higgs field $\langle \mathbf{P} \rangle$ generates masses for the noncommuting fields,

$$m_R = \sqrt{2} g \langle P \rangle \sin^3 \theta_r , \quad m_C = \sqrt{2} \tfrac{2}{3} g \langle P \rangle \cos^3 \theta_r . \quad (50.56)$$

The mass m_R is the mass of each of the two right-handed weak gauge bosons after $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking; the mass is different from the mass m_W of the two left-handed charged weak gauge bosons. The mass m_C is the mass of each of the six leptoquark gauge bosons after $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking. A prediction of the model is that the masses m_R and m_C are related by, from equations (50.56),

$$\frac{m_R}{m_C} = \tfrac{3}{2} \tan^3 \theta_r = 1.1 , \quad (50.57)$$

with $\tan \theta_r = 0.90$ from equation (50.73).

Besides not commuting with right-handed weak and leptoquark fields, the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ Higgs field $\langle \mathbf{P} \rangle$ also fails to commute with the electroweak Higgs multiplet $\mathbf{H}$ (50.3),

$$[\gamma_t^+ \gamma_i^\pm, \rho_3] = \pm \gamma_t^+ \gamma_i^\mp \quad i = y, z. \quad (50.58)$$

Since the Higgs multiplet $\mathbf{H}$ does not generate a gauge symmetry, the covariant derivative $D_m \langle \mathbf{P} \rangle$, equation (50.52), goes not generate any mass term for the electroweak Higgs multiplet. $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking breaks the original $\text{Spin}(4)_w$ symmetry of the electroweak Higgs multiplet down to the electroweak symmetry $\text{U}(1)_Y \times \text{SU}(2)_L$, but the number of Higgs components, 4, remains unchanged.

50.6.3 Scalar fields after electroweak symmetry breaking

After Pati-Salam symmetry breaking, the components of the original scalar multiplet $\mathbf{E}$, equation (50.2), that survive as potentially light scalar fields comprise the 4 components of the electroweak Higgs multiplet (50.3), and a standard-model $\text{U}(1)_Y \times \text{SU}(2)_L \times \text{SU}(3)_c$ component of dimension $1+3+8 = 12$. What happens to these fields subsequently, at electroweak symmetry breaking to today's low energy symmetry $\text{U}(1)_Q \times \text{SU}(3)_c$?

The fate of the 4-component electroweak Higgs multiplet (50.3) at electroweak symmetry breaking is of course known: three of the four degrees of freedom of the Higgs multiplet generate masses for the $W^\pm$ and Z electroweak gauge bosons, while the fourth yields a single electroweak Higgs scalar, of mass 125 GeV, that was detected at the Large Hadron Collider (LHC) in 2012 (Aad et al., 2012; Chatrchyan et al., 2012).

The 12-component standard-model $\text{U}(1)_Y \times \text{SU}(2)_L \times \text{SU}(3)_c$ remnant of the original $\text{Cl}(11, 1)$ $\mathbf{E}$ scalar field, equation (50.2), persists after electroweak symmetry breaking. It is the 4 components of the electroweak Higgs multiplet that are rearranged by electroweak symmetry breaking to $\text{U}(1)_Q \times \text{SU}(3)_c$, not the corresponding components of the $\mathbf{E}$ field. This is just as well: the world would be a very different place if the electroweak Higgs field were not there to couple right- and left-handed fermions and thereby make them massive.

Electroweak symmetry breaking breaks the symmetry of the 12-component standard-model remnant of $\mathbf{E}$, but the number of scalar components, 12, remains the same. These 12 scalar fields are scalar counterparts of standard-model gauge bosons, with identical standard-model charges. The masses of the scalar fields are unknown, because they depend on the unknown potential energies of the scalar fields. Presumably, to have escaped detection, the masses of most of these fields must be greater than the electroweak scale.

Of the 12 scalar fields, two, namely the scalar counterparts of the photon or Z -boson, carry zero standard-model charge, and could potentially constitute the cosmological dark matter. A crucial property of a dark matter particle is that it must be stable against annihilation into lighter particles. One possibility is that a scalar particle might be stabilized by a parity symmetry (Kadastik, Kannike, and Raidal, 2009; Kadastik, Kannike, and Raidal, 2010), but no such symmetry is apparent in the present model. An alternative is that the scalar particle is so light, $m \sim 10^{-22}$ eV, that its decay rate is tiny, a hypothesis called Fuzzy Dark Matter (Hui et al., 2017). Event Horizon Telescope observations of the M87 black hole disfavor ultralight scalar dark matter with mass in the range $\sim 3-5 \times 10^{-21}$ eV (Davoudias and Denton, 2019).

A perhaps more promising candidate for cold dark matter is the dilaton, a particle predicted by string theory, §??.

50.6.4 Fermion masses

In the standard model, after electroweak symmetry breaking, fermions acquire nonzero rest masses by interacting with the electroweak Higgs field. The electroweak Higgs field $\langle \mathbf{H} \rangle$, equation (50.4), carries weak charge y , and flips fermions between right- and left-handed chiralities. In the present picture, the Higgs field $\langle \mathbf{H} \rangle$ flips fermions between Dirac boosts $\uparrow \leftrightarrow \downarrow$ while preserving spin, in accordance with the Spin(11,1) chart (49.7) of spinors. It is the flipping between boosts that allows a fermion to be a superposition of opposite boosts, and therefore to have rest mass.

Unlike the electroweak Higgs field, the Pati-Salam Higgs field $\langle \mathbf{P} \rangle$, equation (50.51), carries zero standard model charge, and does not flip between chiralities. The $\langle \mathbf{P} \rangle$ Higgs field does not flip boosts, and cannot give fermions rest masses.

The conclusion that fermions can acquire their mass only after electroweak symmetry breaking is intriguing. In the Cl(11,1) model, each generation of fermions belongs to a spinor representation of Spin(11,1). The masses of the three known generations of fermions (excluding neutrinos) show a striking progression of masses, with all e -generation fermions being less massive than all μ -generation masses, which are in turn all less massive than τ -generation masses. The most massive fermion known is the τ -generation top, with mass 173 GeV, Table 48.2, close to the 160 GeV (D'Onofrio and Rummukainen, 2016) electroweak scale. If the progression of masses continues, then there is no fourth generation. The known fermions are all there is. This is consistent with the evidence that there are only 3 neutrino types with masses less than half the mass of the Z neutral weak gauge boson, equation (48.22).

The question of why the rest masses of fermions have the pattern observed remains one of the deepest mysteries of the standard model, §48.4.

50.7 Energy scales of unification

This section 50.7 shows from the running of standard-model coupling parameters, Figure 50.2, that the predicted energy scale of Spin(4)_w × Spin(6)_c unification is 10¹² GeV, while the energy scale where the weak and colour couplings meet is 10¹⁵ GeV, equations (50.70).

The relations (50.50) predict that the hypercharge, right-handed weak, and colour coupling parameters are related by

$$\frac{gg_Y}{g_r g_c} = 1 . \quad (50.59)$$

In renormalization theory the coupling parameters vary with the energy at which they are probed. The condition (50.59) can be interpreted as determining the energy scale of Spin(4)_w × Spin(6)_c symmetry breaking.

Spin(4)_w is isomorphic, Exercise 39.2, to the product SU(2)_R × SU(2)_L of right- and left-handed weak groups, and there could potentially be two distinct coupling parameters, g_r and g_w , associated with the two groups. This section starts out treating the two coupling parameters as distinct, but then focuses on the special case of equal right- and left-handed coupling parameters, corresponding to exact chiral symmetry.

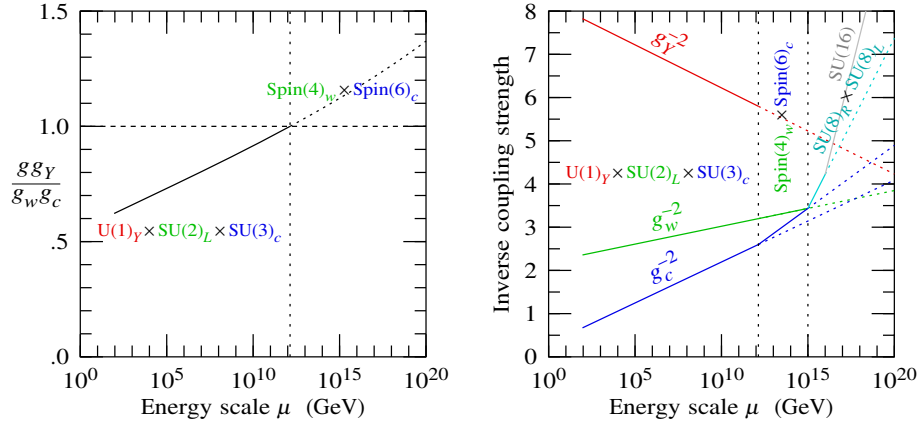


Figure 50.2 (Left) Symmetry breaking of the Pati-Salam group $\text{Spin}(4)_w \times \text{Spin}(6)_c$ to the standard model should occur where $gg_Y/(g_w g_c) = 1$, which happens at energy scale $\mu = 10^{12}$ GeV, equation (50.70). (Right) Running of the standard-model coupling parameters g_Y , g_w , and g_c with renormalization energy scale μ , equation (50.60). The solid blue line (g_c^{-2}) tracks the colour force $\text{SU}(3)_c$ below the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ transition energy, then $\text{Spin}(6)_c$ above the transition energy. The solid green line (g_w^{-2}) tracks the weak force $\text{SU}(2)_L$ below the transition energy, then $\text{Spin}(4)_w$ above the transition energy. The weak and colour couplings g_w and g_c merge at $\mu \approx 10^{15}$ GeV. The solid cyan line tracks the gauge group $\text{SU}(8)_R \times \text{SU}(8)_L$ above the weak-colour merger scale $\mu \approx 10^{15}$ GeV, equation (50.70). Grand symmetry breaking from $\text{SU}(16)$, mediated by the grand Higgs field $\mathbf{T}$, equation (50.31), could occur at any energy at or above the merger scale $\mu \approx 10^{15}$ GeV. The grey line tracks the grand unified group $\text{SU}(16)$ with grand symmetry breaking taken at $\mu = 10^{16}$ GeV. The lines are straight lines when g^{-2} is plotted versus log energy.

Equal coupling parameters would be expected if $\text{Spin}(4)_w$ is the broken remnant of a higher spin group, as in the scenario considered here.

According to renormalization theory, to leading (one-loop) order, the coupling parameter g associated with a gauge group G varies with the log of the cutoff energy μ as (e.g. Peskin, 1997, eq. (39), Kazakov, 2000, eq. (2.6), Schienbein et al., 2019, eq. (5.15), or King et al., 2021, eq. (3.3))

$$\frac{dg^{-2}}{d \ln \mu} = \frac{1}{8\pi^2} \left(\frac{11}{3} S_2(G, \text{adjoint}) - \frac{2}{3} S_2(G, \text{spinor}) n_f - \frac{1}{6} S_2(G, \text{scalar}) n_s \right), \quad (50.60)$$

where $S_2(G, \text{rep})$ is the Dynkin index, Table 39.1, of the representation of the group, and n_f and n_s are respectively the number of fermion and real scalar multiplets that couple to the gauge group G . For $\text{U}(1)_Y$, the formula is (Peskin, 1997, eq. (40))

$$\frac{dg_Y^{-2}}{d \ln \mu} = \frac{1}{8\pi^2} \left(-\frac{2}{3} \sum_f \left(\frac{1}{2} Y \right)^2 - \frac{1}{6} \sum_s \left(\frac{1}{2} Y \right)^2 \right). \quad (50.61)$$

The coupling parameters of the standard model, evaluated at the energy $\mu = m_Z = 91.1884$ GeV of the Z

boson, are, from (Tanabashi et al., 2018),

$$g_Y^2 = \frac{e^2}{\cos^2 \theta_w} = 0.1278, \quad g_w^2 = \frac{e^2}{\sin^2 \theta_w} = 0.4242, \quad g_c^2 = 4\pi\alpha_s = 1.492, \quad (50.62)$$

where the weak mixing angle satisfies $\sin^2 \theta_w = 0.2315$, the electromagnetic coupling is $e^2 = 4\pi\alpha$, and the fine structure and strong coupling constants α and α_s are

$$\alpha(\mu = m_Z) = 1/127.955, \quad \alpha_s(\mu = m_Z) = 0.1187. \quad (50.63)$$

The Dynkin index S_2 of a multiplet in the representations relevant here is, Table 39.1,

$$S_2(\text{SU}(N), \text{adjoint}) = N, \quad S_2(\text{SU}(N), \text{spinor}) = \frac{1}{2}, \quad (50.64a)$$

$$S_2(\text{Spin}(N), \text{grade } p) = \binom{N-2}{p-1}, \quad S_2(\text{Spin}(N), \text{spinor}) = 2^{[(N-1)/2]-3}. \quad (50.64b)$$

The adjoint representation is the special representation where the fields upon which the group acts are the group generators themselves. For $\text{Spin}(N)$, the adjoint representation is the bivector representation, the multivector representation of grade $p = 2$. $\text{Spin}(N)$ has a unique spinor representation, but $\text{SU}(N)$ has spinor representations of spinor grades $p \leq N/2$ (for example the $\text{SU}(5)$ representations given by the columns of the $\text{Spin}(10)$ chart (48.15)). The spinor representation of $\text{SU}(N)$ given by equation (50.64a) is that for spinor grade $p = 1$, which is the only nontrivial spinor grade for the groups $\text{SU}(2)$ and $\text{SU}(3)_c$ relevant here.

The numbers of fermion, electroweak scalar, and adjoint scalar multiplets in each of the standard-model, Pati-Salam, $\text{SU}(8)_R \times \text{SU}(8)_L$, and $\text{SU}(16)$ regimes are as follows.

- Fermions in the standard-model regime. In the standard model, the left-handed weak group $\text{SU}(2)_L$ acts on 4 fermion multiplets, namely the (ν_L, e_L) left-handed lepton multiplet, and the three (u_L^c, d_L^c) left-handed quark multiplets of colours $c = r, g, b$. The colour group $\text{SU}(3)_c$ acts on 4 fermion multiplets, namely the right- and left-handed up and down quark multiplets u_R, u_L, d_R , and d_L . Each fermion multiplet comes in 3 generations, so the number of fermion multiplets in equation (50.60) is $n_f = 4 \times 3 = 12$ for each of $\text{SU}(2)_L$ and $\text{SU}(3)_c$.
- Fermions in the Pati-Salam regime. In the Pati-Salam regime, the weak right-handed group $\text{SU}(2)_R$ is similar to the left-handed group $\text{SU}(2)_L$, but acting on right-handed in place of left-handed multiplets. The number of $\text{SU}(2)_R$ fermion multiplets is the same as for $\text{SU}(2)_L$, $n_f = 4 \times 3 = 12$. The colour group $\text{Spin}(6)_c$ acts on enlarged fermion multiplets that contain leptons as well as quarks, (ν_L, u_L) , (ν_R, u_R) , (e_L, d_L) , and (e_R, d_R) , but the number of fermion multiplets remains the same, $n_f = 4 \times 3 = 12$.
- Fermions in the $\text{SU}(8)_R \times \text{SU}(8)_L$ or $\text{SU}(16)$ regimes. Each $\text{SU}(8)$ group transforms just 1 fermion multiplet in each of 3 generations, for $n_f = 3$, and the same is true for $\text{SU}(16)$.
- Electroweak scalars in the standard-model regime. In the standard model, the electroweak multiplet transforms as a pair of 2-component $\text{SU}(2)_L$ spinors with opposite hypercharges $Y = \pm 1$, so $n_s = 2$ for each of $\text{U}(1)_Y$ and $\text{SU}(2)_L$, while $n_s = 0$ for $\text{SU}(3)_c$. This is the same as in the “minimal” standard model of the electroweak Higgs field.
- Electroweak scalars in the Pati-Salam regime. In the Pati-Salam regime, the electroweak Higgs multiplet (50.3) transforms as a 4-component vector under $\text{Spin}(4)_w$. But $\text{Spin}(4)_w$ is isomorphic to $\text{SU}(2)_R \times$

$SU(2)_L$, and the electroweak multiplet effectively transforms as a pair of 2-component spinors under each of $SU(2)_R$ and $SU(2)_L$, so $n_s = 2$ for each of $SU(2)_R$ and $SU(2)_L$. The scalar Dynkin index is the spinor index $S_2 = \frac{1}{2}$.

- Adjoint Higgs scalars in the standard-model regime. In the standard model regime, the Higgs multiplet $\mathbf{E}$ that breaks the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry, equation (50.2), transforms as a bivector under each of $SU(2)_L$ and $\text{Spin}(6)_c$. The bivectors carry zero Y charge. Thus the $\mathbf{E}$ multiplet contributes $n_s = 1$ for each of $SU(2)_L$ and $\text{Spin}(6)_c$, with adjoint Dynkin index $S_2 = 2$ for $SU(2)_L$, and adjoint index $S_2 = 3$ for $\text{Spin}(6)_c$.
- Adjoint Higgs scalars in the Pati-Salam regime. In the Pati-Salam regime, the Higgs multiplet $\mathbf{E}$ that breaks the $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry transforms as a bivector under each of $SU(2)_R$, $SU(2)_L$, and $\text{Spin}(6)_c$. Thus the $\mathbf{E}$ multiplet contributes $n_s = 1$ for each of $SU(2)_R$, $SU(2)_L$, and $\text{Spin}(6)_c$, with adjoint Dynkin index $S_2 = 2$ for $SU(2)_R$ and $SU(2)_L$, and adjoint index $S_2 = 4$ for $\text{Spin}(6)_c$.
- Adjoint Higgs scalars in the $SU(8)_R \times SU(8)_L$ or $SU(16)$ regimes. Each $SU(8)$ group transforms just 1 scalar multiplet, and the same is true for $SU(16)$.

The above counting of scalars includes those from the adjoint Higgs scalar multiplet $\mathbf{E}$, equation (50.2). Correctly, only scalars whose masses are less than the running energy scale μ should be included. In the literature, in calculating the running of coupling parameters it is standard to ignore Higgs scalars other than the known electroweak Higgs scalar, on the grounds that any additional Higgs scalars may well be massive. In the present context it is more natural to assume that adjoint Higgs scalars $\mathbf{E}$ are massless before symmetry breaking merges them with their massless gauge boson partners to form massive gauge bosons. For generality, let n_E denote the effective number n_s of massless scalars contributed by the $\mathbf{E}$ multiplet before symmetry breaking. If all unbroken $\mathbf{E}$ scalars are massless, as assumed by default here, then $n_E = 1$. If all unbroken $\mathbf{E}$ scalars are massive, the standard assumption, then $n_E = 0$. The calculation shown in Figure 50.2 assumes that adjoint Higgs scalars $\mathbf{E}$ are massless before symmetry breaking, $n_E = 1$.

To summarize, in the standard model, the factor in parentheses on the right hand side of equation (50.60) for the running of coupling parameters is

$$U(1)_Y : \quad -\frac{2}{3} \times \frac{5}{6} \times 12 - \frac{1}{6} \times \frac{1}{2} \times 2 = -\frac{41}{6}, \quad (50.65a)$$

$$SU(2)_L : \frac{11}{3} \times 2 - \frac{2}{3} \times \frac{1}{2} \times 12 - \frac{1}{6} \times (\frac{1}{2} \times 2 + 2 \times n_E) = \frac{19-2n_E}{6} = \frac{17}{6}, \quad (50.65b)$$

$$SU(3)_c : \frac{11}{3} \times 3 - \frac{2}{3} \times \frac{1}{2} \times 12 - \frac{1}{6} \times 3 \times n_E = \frac{14-n_E}{2} = \frac{13}{2}, \quad (50.65c)$$

where the final value assumes the default $n_E = 1$. The sum over fermions of squared hypercharge in equation (50.61) is $\sum_f (\frac{1}{2}Y)^2 = \frac{10}{3}$ per generation, or an average of $\frac{5}{6}$ per $SU(2)_L$ or $SU(3)_c$ fermion multiplet, which accounts for the factor $\frac{5}{6}$ in the $U(1)_Y$ expression (50.65a) (a common practice, not followed here, is to multiply the hypercharge coupling parameter g_Y^2 by $\frac{5}{3}$ so that the fermionic factor $\frac{5}{6}$ in (50.65a) becomes $\frac{1}{2}$, the same as the fermionic factors in $SU(2)_L$ and $SU(3)_c$; the scalar factor in (50.65a) would then change from $\frac{1}{6}$ to $\frac{1}{10}$, and the overall value would change from $-\frac{41}{6}$ to $-\frac{41}{10}$).

In the Pati-Salam regime before $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking, the factor in the running of

coupling parameters is

$$\text{SU}(2)_R, \text{SU}(2)_L : \frac{11}{3} \times 2 - \frac{2}{3} \times \frac{1}{2} \times 12 - \frac{1}{6} \times (\frac{1}{2} \times 2 + 2 \times n_E) = \frac{19-2n_E}{6} = \frac{17}{6} , \quad (50.66a)$$

$$\text{Spin}(6)_c : \frac{11}{3} \times 4 - \frac{2}{3} \times \frac{1}{2} \times 12 - \frac{1}{6} \times 4 \times n_E = \frac{32-2n_E}{3} = 10 , \quad (50.66b)$$

where again the final value assumes the default $n_E = 1$. The factors for $\text{SU}(2)_R$ and $\text{SU}(2)_L$ are the factors for each of the groups individually (and are, naturally, the same); the factors do not add.

In the $\text{SU}(8)_R \times \text{SU}(8)_L$ regime, the factor in the running of the common coupling parameter

$$\text{SU}(8)_R, \text{SU}(8)_L : \frac{11}{3} \times 8 - \frac{2}{3} \times \frac{1}{2} \times 3 - \frac{1}{6} \times 8 \times n_E = \frac{85-4n_E}{3} = 27 , \quad (50.67)$$

where again the final value assumes the default $n_E = 1$. The factors for $\text{SU}(8)_R$ and $\text{SU}(8)_L$ are the factors for each of the groups individually; the factors do not add. In the $\text{SU}(16)$ regime, the factor in the running of the coupling parameter

$$\text{SU}(16) : \frac{11}{3} \times 16 - \frac{2}{3} \times \frac{1}{2} \times 3 - \frac{1}{6} \times 16 \times n_E = \frac{173-8n_E}{3} = 55 , \quad (50.68)$$

where again the final value assumes the default $n_E = 1$.

If the right-handed weak coupling parameter g_r is treated as unknown, then the condition (50.59) leads to no prediction for the energy scale of $\text{Spin}(4)_w \times \text{Spin}(6)_c$ unification. However, in a scenario where $\text{Spin}(4)_w$ is the broken remnant of a larger gauge group, as here, the right- and left-handed coupling parameters g_r and g_w of $\text{Spin}(4)_w$ are same,

$$g_r = g_w , \quad (50.69)$$

in which case the condition (50.59) leads to a definite prediction.

The left panel of Figure 50.2 shows the combination $gg_Y/(g_w g_c)$, equation (50.59), which is predicted to be 1 at $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking. The right panel of Figure 50.2 shows the running of the hypercharge, weak, and colour coupling parameters g_Y , g_w , and g_c as a function of the renormalization cutoff energy μ . The energy scale of $\text{Spin}(4)_w \times \text{Spin}(6)_c$ symmetry breaking, where $gg_Y/(g_w g_c) = 1$, and the energy scale of grand unification, where $g_w = g_c$, are predicted to be

$$\mu_{\frac{gg_Y}{g_w g_c}=1} = 1.4 \times 10^{12} \text{ GeV} , \quad \mu_{g_w=g_c} = 1.0 \times 10^{15} \text{ GeV} . \quad (50.70)$$

The energy scales (50.70) assume that the adjoint Higgs scalars are all light before symmetry breaking, $n_E = 1$. If instead the adjoint Higgs scalars are all heavy before symmetry breaking, $n_E = 0$, then

$$\mu_{\frac{gg_Y}{g_w g_c}=1} = 4.3 \times 10^{11} \text{ GeV} , \quad \mu_{g_w=g_c} = 2.5 \times 10^{14} \text{ GeV} . \quad (50.71)$$

The predicted energy $\sim 1.4 \times 10^{12} \text{ GeV}$ of Pati-Salam unification is slightly greater than that of the most energetic cosmic rays observed (Telescope Array Collaboration, 2014; Alves Batista et al., 2019).

The predicted energy $\sim 10^{15} \text{ GeV}$ of grand unification falls below the lower limit

$$\mu_{\text{Spin}(10)} \gtrsim 4 \times 10^{15} \text{ GeV} \quad (50.72)$$

inferred for $\text{Spin}(10)$ models by King et al. (2021, Fig. 4) from the Super-Kamiokande lower limit of $\gtrsim$

1.6×10^{34} yr on the proton lifetime (Takenaka et. al., 2020). (Note that King et al. (2021, Table 2)'s model I4, where $\text{SO}(10)$ breaks to the standard model via the Pati-Salam group, predicts a symmetry breaking scale of 3.5×10^{16} GeV, but that breaking is mediated by a different Higgs field than considered here.)

Proton decay becomes possible not when the weak and colour parameters meet, but rather when the weak and colour forces unify (King et al., 2021). The predicted energy $\approx 10^{15}$ GeV at which the weak and colour parameters meet is less than the lower limit of $\approx 4 \times 10^{15}$ GeV on the scale of grand unification obtained by (King et al., 2021) from the Super-Kamiokande lower limit on proton lifetime (Takenaka et. al., 2020). The presence of a $\text{SU}(8)_R \times \text{SU}(8)_L$ phase would allow grand unification to occur at any energy above 10^{15} GeV, alleviating the possible tension with the absence of proton decay.

Hamilton and McMaken (2023) emphasized that the value $\sim 10^{15}$ GeV from equations (50.70) falls below the lower limit $\sim 4 \times 10^{15}$ GeV, which might already rule out the $\text{Spin}(11,1)$ model, although that conclusion was not definitive because the prediction of 10^{15} GeV is based on the simplifying assumptions of 1-loop renormalization and an abrupt transition at the Pati-Salam energy, and moreover the $\text{Spin}(11,1)$ symmetry breaking chain differs from the $\text{Spin}(10)$ chains considered by King et al. (2021).

However, string theory that grand unification does not happen when the weak and colour parameters meet, $g_w = g_c$.

However, the tension over the energy scale of grand unification is relieved by string theory, Chapter ?? . The string theory described there predicts that the gauge group was $\text{SU}(16)$ and not just $\text{Spin}(10)$, and the grand Higgs field $\langle T \rangle$, equation (50.31), breaks $\text{SU}(16)$ to $\text{SU}(8)_R \text{SU}(8)_L$.

The $\text{Spin}(4)_w \times \text{Spin}(6)_c$ mixing angle θ_r defined by equations (50.47) is predicted to be

$$\theta_r = 0.73 = 42^\circ, \quad \tan \theta_r = 0.90, \quad (50.73)$$

insensitive to whether adjoint scalars are taken to be light or massive in the running of coupling parameters.

50.8 Appendix: Embedding the Pati-Salam group in $\text{SU}(8)_R \times \text{SU}(8)_L$

The pseudoscalar $I_{i\dots j}$ over p charges $i\dots j$ is defined to be the grade $2p$ product of p diagonal bivectors $I_i \equiv \gamma_i^+ \gamma_i^- = i \gamma_i \gamma_i$, and is related to the chiral operator $\varkappa_{i\dots j}$ whose square is one, $\varkappa_{i\dots j}^2 = 1$, by

$$I_{i\dots j} \equiv I_i \dots I_j \equiv \gamma_i^+ \gamma_i^- \dots \gamma_j^+ \gamma_j^- = i^p \varkappa_{i\dots j} = i^p \gamma_i \gamma_i \dots \gamma_j \gamma_j. \quad (50.74)$$

Acting on a spinor, the chiral operator $\varkappa_{i\dots j}$ is ± 1 depending on whether the number of $i\dots j$ up-bits of the spinor is even or odd, with all-bits-up having $\varkappa_{i\dots j} = +1$.

The group $\text{SU}(8)_R \times \text{SU}(8)_L$ has rank 14, its 14 diagonal generators comprising the 5 bivector generators I_i and the 9 hexavector generators I_{ijk} excluding I_{rgb} . The group $\text{SU}(8)_R \times \text{SU}(8)_L$ has dimension $2 \times 63 = 126$, comprising 21 bivector and 105 hexavector generators in $\text{Cl}(10)$. The bivectors are the $6+15 = 21$ generators of the Pati-Salam group $\text{Spin}(4)_w \times \text{Spin}(6)_c$, and consist of $2+3 = 5$ diagonal generators I_i and $4+12 = 16$ nondiagonal generators. The 105 hexavector generators consist of 9 diagonal generators I_{ijk} , $12+36 = 48$ products of a diagonal quadvector I_{ij} and a nondiagonal weak or colour bivector, and $3 \times 16 = 48$ products of a diagonal colour bivector I_i and a nondiagonal quadvector with 2 weak bits and 2 colour bits. The

$4+12 = 16$ nondiagonal bivectors and $12+36 = 48$ hexavector products of I_{ij} and a nondiagonal bivector combine to $16+48 = 64$ weak and colour generators that flip 2 bits. The 48 hexavector products of I_i and a nondiagonal quadvector flip 2 weak bits and 2 colour bits.

The fermions (and gauge bosons) separate into two distinct conjugate representations of $SU(8)_R \times SU(8)_L$. One of the spinor representations contains 8 right-handed fermions $\psi_{\uparrow\uparrow}$ (ψ running over $\nu_e, e, u^c, d^c, c = r, g, b$) transforming under $SU(8)_R$, together with 8 right-handed antifermions $\bar{\psi}_{\downarrow\downarrow}$ transforming under $SU(8)_L$, while the conjugate representation contains 8 left-handed antifermions $\bar{\psi}_{\uparrow\downarrow}$ (conjugates of $\psi_{\uparrow\uparrow}$) transforming under the conjugate $SU(8)_R$, together with 8 left-handed fermions $\psi_{\downarrow\uparrow}$ (conjugates of $\bar{\psi}_{\downarrow\downarrow}$) transforming under the conjugate $SU(8)_L$. Without loss of generality, restrict to the spinor representation of $SU(8)_R \times SU(8)_L$ containing right-handed fermions $\psi_{\uparrow\uparrow}$ and right-handed antifermions $\bar{\psi}_{\downarrow\downarrow}$.

The nondiagonal generators of $SU(8)_R \times SU(8)_L$ split into Pati-Salam generators that flip 2 bits, and non-Pati-Salam generators that flip 4 bits. The $4+12 = 16$ bivector+hexavector weak $SU(8)_R \times SU(8)_L$ generators that flip 2 bits generate $\text{Spin}(4)_w = SU(2)_R \times SU(2)_L$ transformations (note that each $\leftrightarrow$ accounts for 2 transformations $\rightarrow$ and $\leftarrow$)

$$SU(8)_R \supset SU(2)_R : \quad \nu_{\uparrow\uparrow} \leftrightarrow e_{\uparrow\uparrow}, \quad u_{\uparrow\uparrow}^c \leftrightarrow d_{\uparrow\uparrow}^c, \quad (50.75a)$$

$$SU(8)_L \supset SU(2)_L : \quad \bar{\nu}_{\downarrow\downarrow} \leftrightarrow \bar{e}_{\downarrow\downarrow}, \quad \bar{u}_{\downarrow\downarrow}^c \leftrightarrow \bar{d}_{\downarrow\downarrow}^c. \quad (50.75b)$$

The $12+36 = 48$ bivector+hexavector colour $SU(8)_R \times SU(8)_L$ generators that flip 2 bits generate $\text{Spin}(6)_c$ transitions

$$SU(8)_R \supset SU(3)_{c,R} : \quad u_{\uparrow\uparrow}^c \leftrightarrow u_{\uparrow\uparrow}'^c, \quad d_{\uparrow\uparrow}^c \leftrightarrow d_{\uparrow\uparrow}'^c, \quad (50.76a)$$

$$SU(8)_L \supset SU(3)_{c,L} : \quad \bar{u}_{\downarrow\downarrow}^c \leftrightarrow \bar{u}_{\downarrow\downarrow}'^c, \quad \bar{d}_{\downarrow\downarrow}^c \leftrightarrow \bar{d}_{\downarrow\downarrow}'^c, \quad (50.76b)$$

$$SU(8)_R \supset \frac{\text{Spin}(6)_{c,R}}{SU(3)_{c,R}} : \quad \nu_{\uparrow\uparrow} \leftrightarrow u_{\uparrow\uparrow}^c, \quad e_{\uparrow\uparrow} \leftrightarrow d_{\uparrow\uparrow}^c, \quad (50.76c)$$

$$SU(8)_L \supset \frac{\text{Spin}(6)_{c,L}}{SU(3)_{c,L}} : \quad \bar{\nu}_{\downarrow\downarrow} \leftrightarrow \bar{u}_{\downarrow\downarrow}^c, \quad \bar{e}_{\downarrow\downarrow} \leftrightarrow \bar{d}_{\downarrow\downarrow}^c. \quad (50.76d)$$

The diagonal factors I_{ij} that go into hexavector generators of the Pati-Salam transitions (50.75) and (50.76) serve to distinguish the transitions into separate representations, and do not produce generators outside the Pati-Salam group. The 48 hexavector $SU(8)_R \times SU(8)_L$ non-Pati-Salam transformations that simultaneously flip 2 weak bits and 2 colour bits are

$$\frac{SU(8)_R}{SU(2)_R \times \text{Spin}(6)_{c,R}} : \quad \nu_{\uparrow\uparrow} \leftrightarrow d_{\uparrow\uparrow}^c, \quad e_{\uparrow\uparrow} \leftrightarrow u_{\uparrow\uparrow}^c, \quad u_{\uparrow\uparrow}^c \leftrightarrow d_{\uparrow\uparrow}'^c, \quad (50.77a)$$

$$\frac{SU(8)_L}{SU(2)_L \times \text{Spin}(6)_{c,L}} : \quad \bar{\nu}_{\downarrow\downarrow} \leftrightarrow \bar{d}_{\downarrow\downarrow}^c, \quad \bar{e}_{\downarrow\downarrow} \leftrightarrow \bar{u}_{\downarrow\downarrow}^c, \quad \bar{u}_{\downarrow\downarrow}^c \leftrightarrow \bar{d}_{\downarrow\downarrow}'^c. \quad (50.77b)$$

The weak-colour transformations (50.77) could be accomplished by two separate weak (50.75) and colour (50.76) Pati-Salam steps. The $SU(8)$ transitions (50.76c), (50.76d), and (50.77) interconvert between baryons and leptons, permitting proton decay.

50.9 Appendix: Embedding Spin(10) in SU(16)

The group SU(16) has rank 15, meaning it has 15 mutually commuting diagonal generators, which comprise the 5 bivector generators I_i and the 10 hexavector generators I_{ijk} , with indices $i, j, \dots$ running over y, z, r, g, b . The group SU(16) has dimension 255, comprising 45 bivector and 210 hexavector generators in Cl(10). The 45 bivectors, the generators of Spin(10), consist of 5 diagonal generators I_i and 40 nondiagonal generators. The 210 hexavector generators consist of 10 diagonal generators I_{ijk} , $10 \times 12 = 120$ products of a diagonal quadvector I_{ij} and a nondiagonal bivector, and $5 \times 16 = 80$ products of a diagonal bivector I_i and a nondiagonal quadvector. The 40 nondiagonal bivectors and 120 products of I_{ij} and a nondiagonal bivector combine to 160 generators that flip 2 bits. The 80 products of I_i and a nondiagonal quadvector flip 4 bits.

The fermions (and gauge bosons) separate into two distinct conjugate representations of SU(16). One of the spinor representations contains 8 right-handed fermions $\psi_{\uparrow\uparrow}$ together with 8 right-handed antifermions $\psi_{\downarrow\downarrow}$, while the conjugate representation contains 8 left-handed antifermions $\bar{\psi}_{\uparrow\downarrow}$ together with 8 left-handed fermions $\psi_{\downarrow\uparrow}$. Without loss of generality, restrict to the spinor representation of SU(16) containing right-handed fermions $\psi_{\uparrow\uparrow}$ and right-handed antifermions $\bar{\psi}_{\downarrow\downarrow}$.

The nondiagonal generators of SU(16) split into Spin(10) generators that flip 2 bits, and non-Spin(10) generators that flip 4 bits. The subgroup $SU(8)_R \times SU(8)_L$ of SU(16) transitions are given by equations (50.75)–(50.77). The remaining $SU(16)/(SU(8)_R \times SU(8)_L)$ transitions are as follows. The $24+72 = 96$ bivector+hexavector generators that flip 2 bits generate non-Pati-Salam Spin(10) transitions

$$\frac{SU(16)}{SU(8)_R \times SU(8)_L} \supset \frac{Spin(10)}{Spin(4)_w \times Spin(6)_c} : \begin{array}{cccc} \nu_{\uparrow\uparrow} & \leftrightarrow & \bar{u}_{\downarrow\downarrow}^c & \leftrightarrow & u_{\uparrow\uparrow}^{c'} & \leftrightarrow & \bar{\nu}_{\downarrow\downarrow} \\ e_{\uparrow\uparrow} & & \bar{d}_{\downarrow\downarrow}^c & & d_{\uparrow\uparrow}^{c'} & & \bar{e}_{\downarrow\downarrow} \end{array} . \quad (50.78)$$

The 32 hexavector generators that flip 4 bits are

$$\frac{SU(16)}{SU(8)_R \times SU(8)_L} \Big/ \frac{Spin(10)}{Spin(4)_w \times Spin(6)_c} : \begin{array}{cc} \nu_{\uparrow\uparrow} & \leftrightarrow & \bar{\nu}_{\downarrow\downarrow} \\ e_{\uparrow\uparrow} & & \bar{e}_{\downarrow\downarrow} \end{array} , \quad \begin{array}{cc} u_{\uparrow\uparrow}^c & \leftrightarrow & \bar{u}_{\downarrow\downarrow}^c \\ d_{\uparrow\uparrow}^c & & \bar{d}_{\downarrow\downarrow}^c \end{array} . \quad (50.79)$$

The weak-colour transformations (50.79) could be accomplished by three separate weak and colour Spin(10) steps (50.78).

Bibliography

- Aad, G. et al. (ATLAS Collaboration, 2934 authors) (2012). “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. *Phys. Lett. B* 716, pp. 1–29. DOI: [10.1016/j.physletb.2012.08.020](https://doi.org/10.1016/j.physletb.2012.08.020). arXiv: [1207.7214](https://arxiv.org/abs/1207.7214) [[hep-ex](#)] (cit. on pp. 649, 1199, 1244).
- Abbott, B. P. et al. (LIGO and VIRGO Collaborations) (2016). “Observation of Gravitational Waves from a Binary Black Hole Merger”. *Phys. Rev. Lett.* 116, 061102. DOI: [10.1103/PhysRevLett.116.061102](https://doi.org/10.1103/PhysRevLett.116.061102). arXiv: [1602.03837](https://arxiv.org/abs/1602.03837) [[gr-qc](#)] (cit. on p. 130).
- Abramowicz, M. A. and Fragile, P. C. (2013). “Foundations of Black Hole Accretion Disk Theory”. *Living Rev. Rel.* 16.1. DOI: [10.12942/lrr-2013-1](https://doi.org/10.12942/lrr-2013-1). arXiv: [1104.5499](https://arxiv.org/abs/1104.5499) [[astro-ph.HE](#)] (cit. on p. 711).
- Ade, P. A. R. et al. (BICEP2 Collaboration, 47 authors) (2014). “BICEP2 I: Detection Of B-mode Polarization at Degree Angular Scales”. arXiv: [1403.3985](https://arxiv.org/abs/1403.3985) [[astro-ph.CO](#)] (cit. on pp. 812, 956).
- Adhikari, R. et al. (136 authors) (2017). “A White Paper on keV sterile neutrino Dark Matter”. *Journal of Cosmology and Astroparticle Physics* 2017.01, pp. 025–025. DOI: [10.1088/1475-7516/2017/01/025](https://doi.org/10.1088/1475-7516/2017/01/025). arXiv: [1602.04816](https://arxiv.org/abs/1602.04816) [[hep-ph](#)] (cit. on p. 1205).
- Ajaib, M. A., Gogoladze, I., and Shafi, Q. (2013). “Sparticle spectroscopy from SO(10) GUT with a unified Higgs sector”. *Phys. Rev. D* 88, 095019. DOI: [10.1103/physrevd.88.095019](https://doi.org/10.1103/physrevd.88.095019). arXiv: [1307.4882](https://arxiv.org/abs/1307.4882) [[hep-ph](#)] (cit. on p. 1230).
- Aker, M. et al. (KATRIN Collaboration, 209 authors) (2019). “Improved Upper Limit on the Neutrino Mass from a Direct Kinematic Method by KATRIN”. *Phys. Rev. Lett.* 123, 221802. DOI: [10.1103/physrevlett.123.221802](https://doi.org/10.1103/physrevlett.123.221802) (cit. on p. 269).
- Akhmedov, E. K. and Smirnov, A. Y. (2009). “Paradoxes of neutrino oscillations”. *Physics of Atomic Nuclei* 72, pp. 1363–1381. DOI: [10.1134/S1063778809080122](https://doi.org/10.1134/S1063778809080122). arXiv: [0905.1903](https://arxiv.org/abs/0905.1903) [[hep-ph](#)] (cit. on p. 1212).
- Akiyama, K. et al. (Event Horizon Telescope Collaboration, 348 authors) (2019). “First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole”. *Astrophys. J. Lett.* 875, L1. DOI: [10.3847/2041-8213/ab0ec7](https://doi.org/10.3847/2041-8213/ab0ec7). arXiv: [1906.11243](https://arxiv.org/abs/1906.11243) [[astro-ph.GA](#)] (cit. on pp. 128, 129).
- Altarelli, G. (2013). “The Higgs: so simple yet so unnatural”. *Physica Scripta* 2013, 014011. DOI: [10.1088/0031-8949/2013/t158/014011](https://doi.org/10.1088/0031-8949/2013/t158/014011). arXiv: [1308.0545](https://arxiv.org/abs/1308.0545) [[hep-ph](#)] (cit. on p. 1230).

- Altarelli, G. and Meloni, D. (2013). “A non Supersymmetric SO(10) Grand Unified Model for All the Physics below M_{GUT} ”. *JHEP* 08, p. 021. DOI: [10.1007/jhep08\(2013\)021](https://doi.org/10.1007/jhep08(2013)021). arXiv: [1305.1001](https://arxiv.org/abs/1305.1001) [[hep-th](#)] (cit. on p. 1230).
- Alves Batista, R., Biteau, J., Bustamante, M., Dolag, K., Engel, R., Fang, K., Kampert, K.-H., Kostunin, D., Mostafa, M., Murase, K., Sigl, G., Oikonomou, F., Olinto, A. V., Panasyuk, M. I., Taylor, A., and Unger, M. (2019). “Open Questions in Cosmic-Ray Research at Ultrahigh Energies”. *Frontiers in Astronomy and Space Sciences* 6, 23. DOI: [10.3389/fspas.2019.00023](https://doi.org/10.3389/fspas.2019.00023). arXiv: [1903.06714](https://arxiv.org/abs/1903.06714) [[astro-ph.HE](#)] (cit. on p. 1249).
- Amos, D. E. (1986). “A Portable Package for Bessel Functions of a Complex Argument and Nonnegative Order”. *Trans. Math. Software* 12, pp. 265–273 (cit. on p. 948).
- Arnison, G. et al. (UA1 Collaboration, 138 authors) (1983). “Experimental observation of lepton pairs of invariant mass around 95 GeV/ c^2 at the CERN SPS collider”. *Phys. Lett. B* 126.5, pp. 398–410. DOI: [10.1016/0370-2693\(83\)90188-0](https://doi.org/10.1016/0370-2693(83)90188-0) (cit. on pp. 1199, 1232).
- Arnold, P., Moore, G. D., and Yaffe, L. G. (2000). “Transport coefficients in high temperature gauge theories: (I) Leading-log results”. *JHEP* 11, p. 001. arXiv: [hep-ph/0010177](https://arxiv.org/abs/hep-ph/0010177) (cit. on p. 622).
- Arnowitt, R., Deser, S., and Misner, C. W. (1959). “Dynamical Structure and Definition of Energy in General Relativity”. *Phys. Rev.* 116, pp. 1322–1330. DOI: [10.1103/PhysRev.116.1322](https://doi.org/10.1103/PhysRev.116.1322) (cit. on p. 495).
- (1963). “The dynamics of general relativity”. *Gravitation: an introduction to current research*. Ed. by Witten, L. John Wiley & Sons, pp. 227–265 (cit. on pp. 495, 500).
- Atiyah, M. F., Bott, R., and Shapiro, A. (1964). “Clifford modules”. *Topology* 3, pp. 3–38 (cit. on pp. 1066, 1205).
- Aubert, J. J., Becker, U., Biggs, P. J., Burger, J., Chen, M., Everhart, G., Goldhagen, P., Leong, J., McCorriston, T., Rhoades, T. G., Rohde, M., Ting, S. C. C., Wu, S. L., and Lee, Y. Y. (1974). “Experimental Observation of a Heavy Particle J ”. *Phys. Rev. Lett.* 33, pp. 1404–1406. DOI: [10.1103/PhysRevLett.33.1404](https://doi.org/10.1103/PhysRevLett.33.1404) (cit. on p. 1199).
- Augustin, J.-E., Boyarski, A. M., Breidenbach, M., Bulos, F., Dakin, J. T., Feldman, G. J., Fischer, G. E., Fryberger, D., Hanson, G., Jean-Marie, B., Larsen, R. R., Lüth, V., Lynch, H. L., Lyon, D., Morehouse, C. C., Paterson, J. M., Perl, M. L., Richter, B., Rapidis, P., Schwitters, R. F., Tanenbaum, W. M., Vannucci, F., Abrams, G. S., Briggs, D., Chinowsky, W., Friedberg, C. E., Goldhaber, G., Hollebeek, R. J., Kadyk, J. A., Lulu, B., Pierre, F., Trilling, G. H., Whitaker, J. S., Wiss, J., and Zipse, J. E. (1974). “Discovery of a Narrow Resonance in e^+e^- Annihilation”. *Phys. Rev. Lett.* 33, pp. 1406–1408. DOI: [10.1103/PhysRevLett.33.1406](https://doi.org/10.1103/PhysRevLett.33.1406) (cit. on p. 1199).
- Babichev, E., Chernov, S., Dokuchaev, V., and Eroshenko, Y. (2008). “Ultra-hard fluid and scalar field in the Kerr-Newman metric”. *Phys. Rev. D* 78, 104027. DOI: [10.1103/PhysRevD.78.104027](https://doi.org/10.1103/PhysRevD.78.104027). arXiv: [0807.0449](https://arxiv.org/abs/0807.0449) [[gr-qc](#)] (cit. on pp. 621, 650).
- Babu, K. S. and Khan, S. (2015). “A Minimal Non-Supersymmetric SO(10) Model: Gauge Coupling Unification, Proton Decay and Fermion masses”. *Physical Review D* 92, 075018. DOI: [10.1103/physrevd.92.075018](https://doi.org/10.1103/physrevd.92.075018). arXiv: [1507.06712](https://arxiv.org/abs/1507.06712) [[hep-ph](#)] (cit. on p. 1230).
- Baez, J. C. and Huerta, J. (2010). “The Algebra of Grand Unified Theories”. *Bull. Am. Math. Soc.* 47, pp. 483–552. DOI: [10.1090/S0273-0979-10-01294-2](https://doi.org/10.1090/S0273-0979-10-01294-2). arXiv: [0904.1556](https://arxiv.org/abs/0904.1556) [[hep-th](#)] (cit. on pp. 1200, 1201, 1204–1206).

- Baker, J. G., Centrella, J., Choi, D.-I., Koppitz, M., and Meter, J. van (2006a). “Binary black hole merger dynamics and waveforms”. *Phys. Rev. D* 73, 104002. DOI: [10.1103/PhysRevD.73.104002](#). arXiv: [gr-qc/0602026](#) (cit. on p. 496).
- (2006b). “Gravitational wave extraction from an inspiraling configuration of merging black holes”. *Phys. Rev. Lett.* 96, 111102. DOI: [10.1103/PhysRevLett.96.111102](#). arXiv: [gr-qc/0511103](#) (cit. on p. 496).
- Balbus, S. A. and Hawley, J. F. (1998). “Instability, turbulence, and enhanced transport in accretion disks”. *Rev. Mod. Phys.* 70, pp. 1–53 (cit. on pp. 653, 711).
- Balbus, S. A. (2003). “Enhanced Angular Momentum Transport in Accretion Disks”. *Ann. Rev. Astron. Astrophys.* 41, pp. 555–597. DOI: [10.1146/annurev.astro.41.081401.155207](#). arXiv: [astro-ph/0306208](#) (cit. on p. 711).
- Bardeen, J. M. (1970). “Kerr Metric Black Holes”. *Nature* 226, pp. 64–65. DOI: [10.1038/226064a0](#) (cit. on pp. 711, 712).
- Barrabès, C., Israel, W., and Poisson, E. (1990). “Collision of light-like shells and mass inflation in rotating black holes”. *Class. Quant. Grav.* 7, pp. L273–L278. DOI: [10.1088/0264-9381/7/12/002](#) (cit. on p. 738).
- Baumann, D. (2009). *TASI Lectures on Inflation*. arXiv: [0907.5424 \[hep-th\]](#) (cit. on p. 1003).
- (2022). *Cosmology*. Cambridge, England: Cambridge University Press. 463 pp. DOI: [10.1017/9781108937092](#) (cit. on p. 1003).
- Baumgarte, T. W. and Shapiro, S. L. (1998). “On the numerical integration of Einstein’s field equations”. *Phys. Rev. D* 59, 024007. DOI: [10.1103/PhysRevD.59.024007](#). arXiv: [gr-qc/9810065](#) (cit. on pp. 496, 536).
- (2010). *Numerical Relativity: Solving Einstein’s Equations on the Computer*. Cambridge, England: Cambridge University Press (cit. on p. 536).
- Becchi, C., Rouet, A., and Stora, R. (1976). “Renormalization of gauge theories”. *Annals of Physics* 98, pp. 287–321. DOI: [10.1016/0003-4916\(76\)90156-1](#) (cit. on p. 1149).
- Bednyakov, A. V., Kniehl, B. A., Pikelner, A. F., and Veretin, O. L. (2015). “Stability of the Electroweak Vacuum: Gauge Independence and Advanced Precision”. *Phys. Rev. Lett.* 115, 201802. DOI: [10.1103/physrevlett.115.201802](#). arXiv: [1507.08833 \[hep-ph\]](#) (cit. on p. 1006).
- Bekenstein, J. D. (1973). “Black holes and entropy”. *Phys. Rev. D* 7, pp. 2333–2346. DOI: [10.1103/PhysRevD.7.2333](#) (cit. on pp. 631, 651).
- Belinski, V. A. (2014). “On the cosmological singularity”. *Proc. XIII Marcel Grossmann Meeting, Stockholm 2012*. arXiv: [1404.3864 \[gr-qc\]](#) (cit. on pp. 496, 514, 527).
- Belinskii, V. A. and Khalatnikov, I. M. (1971). “General solution of the gravitational equations with a physical oscillatory singularity”. *Sov. Phys. JETP* 32, pp. 169–172 (cit. on pp. 496, 527).
- Belinskii, V. A., Khalatnikov, I. M., and Lifshitz, E. M. (1970). “Oscillatory approach to a singular point in the relativistic cosmology”. *Advances in Physics* 19, pp. 525–573. DOI: [10.1080/00018737000101171](#) (cit. on pp. 496, 527, 535).
- (1972). “Construction of a General Cosmological Solution of the Einstein Equation with a Time Singularity”. *Sov. Phys. JETP* 35, pp. 838–841 (cit. on pp. 496, 527).
- (1982). “A general solution of the Einstein equations with a time singularity”. *Advances in Physics* 31, pp. 639–667. DOI: [10.1080/00018738200101428](#) (cit. on pp. 496, 514, 527–529, 533).

- Berger, B. K. (2002). “Numerical Approaches to Spacetime Singularities”. *Living Rev. Rel.* 5.1. arXiv: [gr-qc/0201056](#). URL: <https://www.livingreviews.org/lrr-2002-1> (cit. on p. 527).
- Bertschinger, E. (1993). *Cosmological dynamics: Course 1, 1993 Les Houches Lectures*. arXiv: [astro-ph/9503125](#) (cit. on p. 768).
- Bezrukov, F., Kalmykov, M. Y., Kniehl, B. A., and Shaposhnikov, M. (2012). “Higgs boson mass and new physics”. *JHEP* 10, p. 140. DOI: [10.1007/jhep10\(2012\)140](#). arXiv: [1205.2893 \[hep-ph\]](#) (cit. on p. 1006).
- Bhabha, H. J. (1936). “The Scattering of Positrons by Electrons with Exchange on Dirac’s Theory of the Positron”. *Proc. Roy. Soc. A* 136, pp. 195–206 (cit. on p. 1190).
- Bianchi, L. (1898). “Sugli spazii a tre dimensioni che ammettono un gruppo continuo di movimenti (On the spaces of three dimensions that admit a continuous group of movements)”. *Soc. Ital. Sci. Mem. di Mat.* 11, pp. 267–352 (cit. on p. 514).
- BICEP/Keck Collaboration (91 authors) (2021). “Improved Constraints on Primordial Gravitational Waves using Planck, WMAP, and BICEP/Keck Observations through the 2018 Observing Season”. *Phys. Rev. Lett.* 127, 151301. DOI: [10.1103/physrevlett.127.151301](#). arXiv: [2110.00483 \[astro-ph\]](#) (cit. on p. 997).
- Biretta, J. A., Sparks, W. B., and Macchetto, F. (1999). “Hubble Space Telescope Observations of Superluminal Motion in the M87 Jet”. *Astrophys. J.* 520, pp. 621–626. DOI: [10.1086/307499](#) (cit. on p. 40).
- Birrell, N. D. and Davies, P. C. W. (1982). *Quantum fields in curved space*. Cambridge, England: Cambridge University Press. 340 pp. (cit. on p. 316).
- Bjorken, J. D. and Drell, S. D. (1964). *Relativistic Quantum Mechanics*. McGraw-Hill. 300 pp. (cit. on pp. 1080, 1129).
- Blagojević, M. and Hehl, F. W. (2013). *Gauge Theories of Gravitation*. Imperial College Press. 635 pp. (cit. on p. 443).
- Blanchet, S. and Di Bari, P. (2012). “The minimal scenario of leptogenesis”. *New Journal of Physics* 14, 125012. DOI: [10.1088/1367-2630/14/12/125012](#). arXiv: [1211.0512 \[hep-ph\]](#) (cit. on p. 1205).
- Blau, S. K., Guendelman, E. I., and Guth, A. H. (1987). “Dynamics of false-vacuum bubbles”. *Phys. Rev. D* 35, pp. 1747–1766. DOI: [10.1103/PhysRevD.35.1747](#) (cit. on pp. 605, 608, 610, 614).
- Bona, C., Ledvinka, T., Palenzuela, C., and Zacek, M. (2003). “General covariant evolution formalism for numerical relativity”. *Phys. Rev. D* 67, 104005. DOI: [10.1103/PhysRevD.67.104005](#). arXiv: [gr-qc/0302083](#) (cit. on p. 541).
- Bonanno, A., Droz, S., Israel, W., and Morsink, S. M. (1994a). “Structure of the inner singularity of a spherical black hole”. *Phys. Rev. D* 50, pp. 7372–7375. DOI: [10.1103/PhysRevD.50.7372](#). arXiv: [gr-qc/9403019](#) (cit. on p. 649).
- (1994b). “Structure of the spherical black hole interior”. *Proc. Roy. Soc. London A* 450, pp. 553–567. arXiv: [gr-qc/9411050](#) (cit. on p. 646).
- Bott, R. (1970). “The Periodicity Theorem For The Classical Groups And Some Of Its Applications”. *Advances in Mathematics* 4, pp. 353–411. DOI: [10.1016/0001-8708\(70\)90030-7](#) (cit. on p. 1219).
- Bousso, R. (2002). “The holographic principle”. *Rev. Mod. Phys.* 74, pp. 825–874. DOI: [10.1103/RevModPhys.74.825](#). arXiv: [hep-th/0203101](#) (cit. on pp. 632, 654, 655).
- Bradley, J. (1728). “An account of a new discovered motion of the fixed stars”. *Phil. Trans.* 35, pp. 637–661 (cit. on p. 44).

- Brady, P. R. (1995). “Selfsimilar scalar field collapse: Naked singularities and critical behavior”. *Phys. Rev. D* 51, pp. 4168–4176. DOI: [10.1103/PhysRevD.51.4168](https://doi.org/10.1103/PhysRevD.51.4168). arXiv: [gr-qc/9409035](https://arxiv.org/abs/gr-qc/9409035) (cit. on p. 649).
- Brady, P. R. and Smith, J. D. (1995). “Black hole singularities: A Numerical approach”. *Phys. Rev. Lett.* 75, pp. 1256–1259. DOI: [10.1103/PhysRevLett.75.1256](https://doi.org/10.1103/PhysRevLett.75.1256). arXiv: [gr-qc/9506067](https://arxiv.org/abs/gr-qc/9506067) (cit. on pp. 646, 649).
- Brauer, R. and Weyl, H. (1935). “Spinors in n Dimensions”. *American Journal of Mathematics* 57, pp. 425–449. DOI: [10.2307/2371218](https://doi.org/10.2307/2371218) (cit. on pp. 1052, 1206, 1216).
- Brillouin, L. (1926). “La m’ecanique ondulatoire de Schrödinger: une m’ethode g’en’erale de resolution par approximations successives”. *Comptes Rendus de l’Academie des Sciences* 187, pp. 24–26 (cit. on p. 900).
- Brown, J. D., Diener, P., Field, S. E., Hesthaven, J. S., Herrmann, F., Mroue, A. H., Sarbach, O., Schnetter, E., Tiglio, M., and Wagman, M. (2012). “Numerical simulations with a first order BSSN formulation of Einstein’s field equations”. *Phys. Rev. D* 85, 084004. DOI: [10.1103/PhysRevD.85.084004](https://doi.org/10.1103/PhysRevD.85.084004). arXiv: [1202.1038 \[gr-qc\]](https://arxiv.org/abs/1202.1038) (cit. on p. 536).
- Buchman, L. T. and Bardeen, J. M. (2005). “Schwarzschild tests of the WEBB tetrad formulation for numerical relativity”. *Phys. Rev. D* 72, 124014. DOI: [10.1103/PhysRevD.72.124014](https://doi.org/10.1103/PhysRevD.72.124014). arXiv: [gr-qc/0508111](https://arxiv.org/abs/gr-qc/0508111) (cit. on p. 481).
- Buchman, L. T. and Bardeen, J. M. (2003). “A Hyperbolic tetrad formulation of the Einstein equations for numerical relativity”. *Phys. Rev. D* 67, 084017. [Erratum: *Phys. Rev. D* 72, 049903(2005)]. DOI: [10.1103/PhysRevD.72.049903](https://doi.org/10.1103/PhysRevD.72.049903), [10.1103/PhysRevD.67.084017](https://doi.org/10.1103/PhysRevD.67.084017). arXiv: [gr-qc/0301072](https://arxiv.org/abs/gr-qc/0301072) (cit. on p. 481).
- Buchmüller, W., Di Bari, P., and Plümacher, M. (2005). “Leptogenesis for pedestrians”. *Annals of Physics* 315.2, pp. 305–351. DOI: [10.1016/j.aop.2004.02.003](https://doi.org/10.1016/j.aop.2004.02.003). arXiv: [hep-ph/0401240 \[hep-ph\]](https://arxiv.org/abs/hep-ph/0401240) (cit. on p. 1240).
- Buchmüller, W., Peccei, R. D., and Yanagida, T. (2005). “Leptogenesis as the origin of matter”. *Annual Review of Nuclear and Particle Science* 55, pp. 311–355. DOI: [10.1146/annurev.nucl.55.090704.151558](https://doi.org/10.1146/annurev.nucl.55.090704.151558). arXiv: [hep-ph/0502169 \[hep-ph\]](https://arxiv.org/abs/hep-ph/0502169) (cit. on p. 1205).
- Burgess, C. P. (2004). “Quantum gravity in everyday life: General relativity as an effective field theory”. *Living Rev. Rel.* 7, pp. 5–56. DOI: [10.12942/lrr-2004-5](https://doi.org/10.12942/lrr-2004-5). arXiv: [gr-qc/0311082 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0311082) (cit. on p. 1097).
- Burko, L. M. (1997). “Structure of the black hole’s Cauchy horizon singularity”. *Phys. Rev. Lett.* 79, pp. 4958–4961. eprint: [gr-qc/9710112](https://arxiv.org/abs/gr-qc/9710112) (cit. on pp. 646, 649).
- (1999). “Singularity deep inside the spherical charged black hole core”. *Phys. Rev. D* 59, 024011. eprint: [gr-qc/9809073](https://arxiv.org/abs/gr-qc/9809073) (cit. on p. 649).
- (2002). “Survival of the black hole’s Cauchy horizon under non-compact perturbations”. *Phys. Rev. D* 66, 024046. eprint: [gr-qc/0206012](https://arxiv.org/abs/gr-qc/0206012) (cit. on pp. 646, 649).
- (2003). “Black hole singularities: A new critical phenomenon”. *Phys. Rev. Lett.* 90, 121101. eprint: [gr-qc/0209084](https://arxiv.org/abs/gr-qc/0209084) (cit. on pp. 646, 649).
- Burko, L. M. and Ori, A. (1998). “Analytic study of the null singularity inside spherical charged black holes”. *Phys. Rev. D* 57, pp. 7084–7088. DOI: [10.1103/PhysRevD.57.7084](https://doi.org/10.1103/PhysRevD.57.7084). arXiv: [gr-qc/9711032](https://arxiv.org/abs/gr-qc/9711032) (cit. on pp. 646, 649).
- Buttazzo, D., Degraasi, G., Giardino, P. P., Giudice, G. F., Sala, F., Salvio, A., and Strumia, A. (2013). “Investigating the near-criticality of the Higgs boson”. *JHEP* 12, p. 089. DOI: [10.1007/jhep12\(2013\)089](https://doi.org/10.1007/jhep12(2013)089). arXiv: [1307.3536 \[hep-ph\]](https://arxiv.org/abs/1307.3536) (cit. on p. 1006).

- Campanelli, M., Lousto, C. O., Marronetti, P., and Zlochower, Y. (2006). “Accurate evolutions of orbiting black-hole binaries without excision”. *Phys. Rev. Lett.* 96, 111101. DOI: [10.1103/PhysRevLett.96.111101](https://doi.org/10.1103/PhysRevLett.96.111101). arXiv: [gr-qc/0511048](https://arxiv.org/abs/gr-qc/0511048) (cit. on p. 496).
- Campanelli, M., Lousto, C. O., and Zlochower, Y. (2006). “The Last orbit of binary black holes”. *Phys. Rev. D* 73, 061501. DOI: [10.1103/PhysRevD.73.061501](https://doi.org/10.1103/PhysRevD.73.061501). arXiv: [gr-qc/0601091](https://arxiv.org/abs/gr-qc/0601091) (cit. on p. 496).
- Carroll, S. M., Press, W. H., and Turner, E. L. (1992). “The Cosmological constant”. *Ann. Rev. Astron. Astrophys.* 30, pp. 499–542. DOI: [10.1146/annurev.aa.30.090192.002435](https://doi.org/10.1146/annurev.aa.30.090192.002435) (cit. on p. 853).
- Cartan, É. (1904). “Sur la structure des groupes infinis de transformations”. *Annales Scientifiques de l’École Normale Supérieure* 21, pp. 153–206 (cit. on pp. 313, 418, 460).
- (1913). “Les groupes projectifs qui ne laissent invariante aucune multiplicité plane”. *Bulletin Société Mathématique de France* 41, pp. 53–96. DOI: [10.24033/bsmf.916](https://doi.org/10.24033/bsmf.916) (cit. on p. 1035).
- (1938). *Leçons sur la théorie des spineurs*. Paris, France: Hermann & Cie (cit. on p. 1035).
- (1945). *Les Systèmes Différentiels Extérieurs et leurs Applications Géométriques*. Paris, France: Hermann & Cie (cit. on p. 395).
- Carter, B. (1968a). “Global structure of the Kerr family of gravitational fields”. *Phys. Rev.* 174, pp. 1559–1571 (cit. on p. 218).
- (1968b). “Hamilton-Jacobi and Schrödinger separable solutions of Einstein’s equations”. *Commun. Math. Phys.* 10, pp. 280–310. DOI: [10.1007/BF03399503](https://doi.org/10.1007/BF03399503) (cit. on pp. 657, 660, 663, 739).
- Chandrasekhar, S. (1983). *The mathematical theory of black holes*. Oxford, England: Clarendon Press. 646 pp. (cit. on p. 326).
- Chatrchyan, S. et al. (CMS Collaboration, 2885 authors) (2012). “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. *Phys. Lett. B* 716, pp. 30–61. DOI: [10.1016/j.physletb.2012.08.021](https://doi.org/10.1016/j.physletb.2012.08.021). arXiv: [1207.7235 \[hep-ex\]](https://arxiv.org/abs/1207.7235) (cit. on pp. 649, 1199, 1244).
- Chluba, J. and Thomas, R. M. (2011). “Towards a complete treatment of the cosmological recombination problem”. *Mon. Not. Roy. Astron. Soc.* 412, pp. 748–764. DOI: [10.1111/j.1365-2966.2010.17940.x](https://doi.org/10.1111/j.1365-2966.2010.17940.x). arXiv: [1010.3631 \[astro-ph.CO\]](https://arxiv.org/abs/1010.3631) (cit. on p. 884).
- Choptuik, M. W. (1993). “Universality and scaling in gravitational collapse of a massless scalar field”. *Phys. Rev. Lett.* 70, pp. 9–12. DOI: [10.1103/PhysRevLett.70.9](https://doi.org/10.1103/PhysRevLett.70.9) (cit. on pp. 632, 653).
- Christodoulou, D. (1984). “Violation of cosmic censorship in the gravitational collapse of a dust cloud”. *Commun. Math. Phys.* 93, pp. 171–195. DOI: [10.1007/BF01223743](https://doi.org/10.1007/BF01223743) (cit. on p. 601).
- (1986). “The problem of a self-gravitating scalar field”. *Commun. Math. Phys.* 105, pp. 337–361 (cit. on p. 649).
- Clifford, W. K. (1878). “Applications of Grassmann’s extensive algebra”. *Am. J. Math.* 1, pp. 350–358. DOI: [10.2307/2369379](https://doi.org/10.2307/2369379) (cit. on pp. 330, 332, 1206).
- Cline, J. M. (2018). *TASI Lectures on Early Universe Cosmology: Inflation, Baryogenesis and Dark Matter*. arXiv: [1807.08749 \[hep-ph\]](https://arxiv.org/abs/1807.08749) (cit. on p. 1205).
- Coleman, S. and Mandula, J. (1967). “All Possible Symmetries of the S Matrix”. *Phys. Rev.* 159, pp. 1251–1256. DOI: [10.1103/PhysRev.159.1251](https://doi.org/10.1103/PhysRev.159.1251) (cit. on pp. 1216, 1225).
- Conway, J. H. and Sloane, N. J. A. (1982). “On the enumeration of lattices of determinant one”. *Journal of Number Theory* 15, pp. 83–94. DOI: [10.1016/0022-314X\(82\)90084-1](https://doi.org/10.1016/0022-314X(82)90084-1) (cit. on pp. 1030, 1033).

- Coquereaux, R. (1982). “Modulo 8 periodicity of real Clifford algebras and particle physics”. *Phys. Lett. B* 115, pp. 389–395. ISSN: 0370-2693. DOI: [10.1016/0370-2693\(82\)90524-X](https://doi.org/10.1016/0370-2693(82)90524-X) (cit. on p. 1219).
- Cornwell, J. F. (1997). *Group Theory in Physics—An Introduction*. San Diego, CA, USA: Academic Press. DOI: [10.1016/B978-0-12-189800-7.X5000-6](https://doi.org/10.1016/B978-0-12-189800-7.X5000-6) (cit. on p. 1016).
- Corson, E. M. (1953). *Introduction to Tensors, Spinors, and Relativistic Wave-Equations*. London and Glasgow: Blackie & Son (cit. on pp. 445, 493).
- Croon, D., Gonzalo, T. E., Graf, L., Košnik, N., and White, G. (2019). “GUT Physics in the Era of the LHC”. *Frontiers in Physics* 7. DOI: [10.3389/fphy.2019.00076](https://doi.org/10.3389/fphy.2019.00076). arXiv: [1903.04977 \[hep-ph\]](https://arxiv.org/abs/1903.04977) (cit. on p. 1230).
- Cyburt, R. H., Fields, B. D., Olive, K. A., and Yeh, T.-H. (2016). “Big Bang Nucleosynthesis: 2015”. *Rev. Mod. Phys.* 88, 015004. DOI: [10.1103/RevModPhys.88.015004](https://doi.org/10.1103/RevModPhys.88.015004). arXiv: [1505.01076 \[astro-ph.CO\]](https://arxiv.org/abs/1505.01076) (cit. on pp. 867, 898).
- D’Onofrio, M. and Rummukainen, K. (2016). “Standard model cross-over on the lattice”. *Phys. Rev. D* 93, 025003. DOI: [10.1103/physrevd.93.025003](https://doi.org/10.1103/physrevd.93.025003). arXiv: [1508.07161 \[hep-ph\]](https://arxiv.org/abs/1508.07161) (cit. on p. 1245).
- Dafermos, M. (2005). “The interior of charged black holes and the problem of uniqueness in general relativity”. *Commun. Pure Appl. Math.* 58, pp. 445–504. arXiv: [gr-qc/0307013](https://arxiv.org/abs/gr-qc/0307013) (cit. on pp. 645, 649).
- Dafermos, M. and Rodnianski, I. (2005). “A proof of Price’s law for the collapse of a self-gravitating scalar field”. *Invent. Math.* 162, pp. 381–457. DOI: [10.1007/s00222-005-0450-3](https://doi.org/10.1007/s00222-005-0450-3). arXiv: [gr-qc/0309115](https://arxiv.org/abs/gr-qc/0309115) (cit. on pp. 645, 649).
- Das, Sudeep and Sherwin, Blake D. et al. (ACT Collaboration, 40 authors) (2011). “Detection of the Power Spectrum of Cosmic Microwave Background Lensing by the Atacama Cosmology Telescope”. *Phys. Rev. Lett.* 107, 021301. DOI: [10.1103/PhysRevLett.107.021301](https://doi.org/10.1103/PhysRevLett.107.021301). arXiv: [1103.2124 \[astro-ph.CO\]](https://arxiv.org/abs/1103.2124) (cit. on p. 233).
- Davidson, S., Nardi, E., and Nir, Y. (2008). “Leptogenesis”. *Phys. Rept.* 466, pp. 105–177. arXiv: [0802.2962 \[hep-ph\]](https://arxiv.org/abs/0802.2962) (cit. on p. 1205).
- Davoudias, H. and Denton, P. B. (2019). “Ultralight Boson Dark Matter and Event Horizon Telescope Observations of M87*”. *Phys. Rev. Lett.* 123, 021102. DOI: [10.1103/PhysRevLett.123.021102](https://doi.org/10.1103/PhysRevLett.123.021102). arXiv: [1904.09242 \[astro-ph.CO\]](https://arxiv.org/abs/1904.09242) (cit. on p. 1244).
- de Bernardis, P., Ade, P. A. R., Bock, J. J., Bond, J. R., Borrill, J., Boscaleri, A., Coble, K., Crill, B. P., De Gasperis, G., Farese, P. C., Ferreira, P. G., Ganga, K., Giacometti, M., Hivon, E., Hristov, V. V., Iacoangeli, A., Jaffe, A. H., Lange, A. E., Martinis, L., Masi, S., Mason, P. V., Mauskopf, P. D., Melchiorri, A., Miglio, L., Montroy, T., Netterfield, C. B., Pascale, E., Piacentini, F., Pogosyan, D., Prunet, S., Rao, S., Romeo, G., Ruhl, J. E., Scaramuzzi, F., Sforna, D., and Vittorio, N. (2000). “A flat Universe from high-resolution maps of the cosmic microwave background radiation”. *Nature* 404, pp. 955–959. eprint: [astro-ph/0004404](https://arxiv.org/abs/astro-ph/0004404) (cit. on pp. 851, 933).
- de Sitter, W. (1916). “On Einstein’s theory of gravitation and its astronomical consequences. Second paper”. *Mon. Not. Roy. Astron. Soc.* 77, pp. 155–184 (cit. on p. 780).
- Degrassi, G., Di Vita, S., Elias-Miró, J., Espinosa, J. R., Giudice, G. F., Isidori, G., and Strumia, A. (2012). “Higgs mass and vacuum stability in the Standard Model at NNLO”. *JHEP* 08, p. 098. DOI: [10.1007/jhep08\(2012\)098](https://doi.org/10.1007/jhep08(2012)098). arXiv: [1205.6497 \[hep-ph\]](https://arxiv.org/abs/1205.6497) (cit. on p. 1006).

- del Aguila, F. and Ib'añez, L. E. (1981). “Higgs bosons in SO(10) and partial unification”. *Nucl. Phys. B* 177, pp. 60–86. DOI: [10.1016/0550-3213\(81\)90266-2](https://doi.org/10.1016/0550-3213(81)90266-2) (cit. on pp. 1230, 1240).
- Di Luzio, L. (2011). *Aspects of symmetry breaking in Grand Unified Theories*. arXiv: [1110.3210](https://arxiv.org/abs/1110.3210) [hep-ph] (cit. on p. 1230).
- Dicke, R. H., Peebles, P. J. E., Roll, P. G., and Wilkinson, D. T. (1965). “Cosmic Black-Body Radiation”. *Astrophys. J. Lett.* 142, pp. 414–419. DOI: [10.1086/148306](https://doi.org/10.1086/148306) (cit. on p. 231).
- Diener, P., Herrmann, F., Pollney, D., Schnetter, E., Seidel, E., et al. (2006). “Accurate evolution of orbiting binary black holes”. *Phys. Rev. Lett.* 96, 121101. DOI: [10.1103/PhysRevLett.96.121101](https://doi.org/10.1103/PhysRevLett.96.121101). arXiv: [gr-qc/0512108](https://arxiv.org/abs/gr-qc/0512108) (cit. on p. 496).
- Dirac, P. A. M. (1964). *Lectures on Quantum Mechanics*. New York: Belfer Graduate School of Science (cit. on pp. 428, 434, 1105).
- Donoghue, J. F. (2019). “Gravitons and Pions”. *Eur. Phys. J.* arXiv: [1908.11003](https://arxiv.org/abs/1908.11003) [nucl-th] (cit. on p. 1097).
- Doran, C. (2000). “A new form of the Kerr solution”. *Phys. Rev. D* 61, 067503. DOI: [10.1103/PhysRevD.61.067503](https://doi.org/10.1103/PhysRevD.61.067503). arXiv: [gr-qc/9910099](https://arxiv.org/abs/gr-qc/9910099) (cit. on pp. 222, 576).
- Drewes, M. (2013). “The phenomenology of right handed neutrinos”. *International Journal of Modern Physics E* 22, 1330019. DOI: [10.1142/s0218301313300191](https://doi.org/10.1142/s0218301313300191). arXiv: [1303.6912](https://arxiv.org/abs/1303.6912) [hep-ph] (cit. on p. 1205).
- Droste, J. (1916). “Het veld van een enkel centrum in Einstein’s theorie der zwaarekracht end de beweging van een stoffelijk punt in dat veld”. *Verslagen van de Gewone Vergaderingen der Wis- en Natuurkundige Afdeeling, Koninklijke Akademie van Wetenschappen te Amsterdam* 25, pp. 163–180 (cit. on pp. 134, 154).
- Dueck, A. and Rodejohann, W. (2013). “Fits to SO(10) grand unified models”. *JHEP* 09, p. 024. DOI: [10.1007/jhep09\(2013\)024](https://doi.org/10.1007/jhep09(2013)024). arXiv: [1306.4468](https://arxiv.org/abs/1306.4468) [hep-th] (cit. on p. 1230).
- Dynkin, E. B. (1957a). “Maximal subgroups of the classical groups”. *Amer. Math. Soc. Translations, Series 2* 6, pp. 245–378. DOI: [10.1090/trans2/006](https://doi.org/10.1090/trans2/006) (cit. on p. 1015).
- (1957b). “Semisimple subalgebras of semisimple Lie algebras”. *Amer. Math. Soc. Translations, Series 2* 6, pp. 111–244. DOI: [10.1090/trans2/006](https://doi.org/10.1090/trans2/006) (cit. on p. 1015).
- Eddington, A. S. (1924). “A comparison of Whitehead’s and Einstein’s formulæ”. *Nature* 113, p. 192 (cit. on p. 155).
- Einstein, A. (1917). “Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie”. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 142–152 (cit. on p. 83).
- Einstein, A. (1905). “Zur Elektrodynamik bewegter Körper”. *Annalen der Physik* 322, pp. 891–921. DOI: [10.1002/andp.19053221004](https://doi.org/10.1002/andp.19053221004) (cit. on p. 11).
- (1907). “Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen”. *Jahrbuch der Radioaktivität und Elektronik* 4 (cit. on p. 52).
- (1915). “Die Feldgleichungen der Gravitation”. *Königlich Preussische Akademie der Wissenschaften (Berlin), Sitzungsberichte*, pp. 844–847 (cit. on p. 53).
- Eisenhauer, F., Genzel, R., Alexander, T., Abuter, R., Paumard, T., Ott, T., Gilbert, A., Gillessen, S., Horrobin, M., Trippe, S., Bonnet, H., Dumas, C., Hubin, N., Kaufer, A., Kissler-Patig, M., Monnet, G., Stroebele, S., Szeifert, T., Eckart, A., Schoedel, R., and Zucker, S. (2005). “SINFONI in the Galactic Center: young stars and IR flares in the central light month”. *Astrophys. J.* 628, pp. 246–259. DOI: [10.1086/430667](https://doi.org/10.1086/430667). arXiv: [astro-ph/0502129](https://arxiv.org/abs/astro-ph/0502129) (cit. on p. 170).

- Elias-Miró, J., Espinosa, J. R., Giudice, G. F., Isidori, G., Riotto, A., and Strumia, A. (2012). “Higgs mass implications on the stability of the electroweak vacuum”. *Phys. Lett. B* 709, pp. 222–228. DOI: [10.1016/j.physletb.2012.02.013](https://doi.org/10.1016/j.physletb.2012.02.013). arXiv: [1112.3022](https://arxiv.org/abs/1112.3022) [[hep-ph](#)] (cit. on p. [1006](#)).
- Englert, F. and Brout, R. (1964). “Broken Symmetry and the Mass of Gauge Vector Mesons”. *Phys. Rev. Lett.* 13, pp. 321–323. DOI: [10.1103/PhysRevLett.13.321](https://doi.org/10.1103/PhysRevLett.13.321) (cit. on pp. [1199](#), [1229](#), [1230](#)).
- Estabrook, F. B., Robinson, R. S., and Wahlquist, H. D. (1997). “Hyperbolic equations for vacuum gravity using special orthonormal frames”. *Class. Quant. Grav.* 14, pp. 1237–1247. DOI: [10.1088/0264-9381/14/5/025](https://doi.org/10.1088/0264-9381/14/5/025). arXiv: [gr-qc/9703072](https://arxiv.org/abs/gr-qc/9703072) (cit. on p. [481](#)).
- Feynman, R. P. (1949). “Space-Time Approach to Quantum Electrodynamics”. *Phys. Rev.* 76, pp. 769–789. DOI: [10.1103/PhysRev.76.769](https://doi.org/10.1103/PhysRev.76.769) (cit. on p. [1129](#)).
- Finkelstein, D. (1958). “Past-Future Asymmetry of the Gravitational Field of a Point Particle”. *Phys. Rev.* 110, pp. 965–967. DOI: [10.1103/PhysRev.110.965](https://doi.org/10.1103/PhysRev.110.965) (cit. on p. [155](#)).
- Fixsen, D. J. (2009). “The Temperature of the Cosmic Microwave Background”. *Astrophys. J.* 707, pp. 916–920. DOI: [10.1088/0004-637X/707/2/916](https://doi.org/10.1088/0004-637X/707/2/916). arXiv: [0911.1955](https://arxiv.org/abs/0911.1955) [[astro-ph.CO](#)] (cit. on pp. [231](#), [897](#)).
- Fock, V. (1957). “Three Lectures on Relativity Theory”. *Rev. Mod. Phys.* 29, pp. 325–333. DOI: [10.1103/RevModPhys.29.325](https://doi.org/10.1103/RevModPhys.29.325) (cit. on pp. [541](#), [772](#)).
- Fong, C. S., Nardi, E., and Riotto, A. (2012). “Leptogenesis in the Universe”. *Advances in High Energy Physics* 2012, pp. 1–59. DOI: [10.1155/2012/158303](https://doi.org/10.1155/2012/158303). arXiv: [1301.3062](https://arxiv.org/abs/1301.3062) [[hep-ph](#)] (cit. on p. [1205](#)).
- Forero, D. V., Tortola, M., and Valle, J. W. F. (2012). “Global status of neutrino oscillation parameters after Neutrino-2012”. *Phys. Rev. D* 86, 073012. DOI: [10.1103/PhysRevD.86.073012](https://doi.org/10.1103/PhysRevD.86.073012). arXiv: [1205.4018](https://arxiv.org/abs/1205.4018) [[hep-ph](#)] (cit. on p. [268](#)).
- Friedmann, A. (1922). “Über die Krümmung des Raumes”. *Zeitschrift für Physik* A10, pp. 377–386 (cit. on p. [236](#)).
- (1924). “Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes”. *Zeitschrift für Physik* A21, pp. 326–332 (cit. on p. [236](#)).
- Fritzsch, H. and Minkowski, P. (1975). “Unified Interactions of Leptons and Hadrons”. *Annals of Physics* 93, pp. 193–266. DOI: [10.1016/0003-4916\(75\)90211-0](https://doi.org/10.1016/0003-4916(75)90211-0) (cit. on p. [1202](#)).
- Frolov, A. V., Kristjansson, K. R., and Thorlacius, L. (2006). “Global geometry of two-dimensional charged black holes”. *Phys. Rev. D* 73, 124036. DOI: [10.1103/PhysRevD.73.124036](https://doi.org/10.1103/PhysRevD.73.124036). arXiv: [hep-th/0604041](https://arxiv.org/abs/hep-th/0604041) (cit. on pp. [640](#), [646](#)).
- Fukugita M. and Yanagida, T. (1986). “Baryogenesis without grand unification”. *Phys. Lett. B* 174, pp. 45–47. DOI: [10.1016/0370-2693\(86\)91126-3](https://doi.org/10.1016/0370-2693(86)91126-3) (cit. on p. [1205](#)).
- Furey, N. and Hughes, M. J. (2022). “Division algebraic symmetry breaking”. *Phys. Lett. B* 831, 137186. DOI: [10.1016/j.physletb.2022.137186](https://doi.org/10.1016/j.physletb.2022.137186). arXiv: [2210.10126](https://arxiv.org/abs/2210.10126) [[hep-ph](#)] (cit. on p. [1230](#)).
- Gebhardt, K., Adams, J., Richstone, D., Lauer, T. R., Faber, S. M., Gültekin, K., Murphy, J., and Tremaine, S. (2011). “The Black-Hole Mass in M87 from Gemini/NIFS Adaptive Optics Observations”. *Astrophys. J.* 729, 119. DOI: [10.1088/0004-637X/729/2/119](https://doi.org/10.1088/0004-637X/729/2/119). arXiv: [1101.1954](https://arxiv.org/abs/1101.1954) [[astro-ph.CO](#)] (cit. on p. [128](#)).
- Gell-Mann, M., Ramond, P., and Slansky, R. (1979). “Complex Spinors and Unified Theories”. *Supergravity*. Ed. by Freedman, D. Z. and Nieuwenhuizen, P. van. North Holland, Amsterdam, pp. 315–321. arXiv: [1306.4669](https://arxiv.org/abs/1306.4669) [[hep-th](#)] (cit. on pp. [1209](#), [1210](#)).

- Georgi, H. (1975). *Particles and Fields – 1974, Proc. Meeting of the APS Division of Particles and Fields, Williamsburg, Virginia*. Ed. by Carlson, C. E. AIP, New York, p. 575 (cit. on p. 1202).
- Georgi, H. and Glashow, S. (1974). “Unity of all elementary-particle forces”. *Phys. Rev. Lett.* 32, pp. 438–441. DOI: [10.1103/PhysRevLett.32.438](https://doi.org/10.1103/PhysRevLett.32.438) (cit. on p. 1201).
- Geroch, R., Held, A., and Penrose, R. (1973). “A space-time calculus based on pairs of null directions”. *J. Math. Phys.* 14, pp. 874–881. DOI: [10.1063/1.1666410](https://doi.org/10.1063/1.1666410) (cit. on p. 981).
- Ghez, A. M., Salim, S., Hornstein, S. D., Tanner, A., Lu, J. R., Morris, M., Becklin, E. E., and Duchene, G. (2005). “Stellar Orbits Around the Galactic Center Black Hole”. *Astrophys. J.* 620, pp. 744–757. DOI: [10.1086/427175](https://doi.org/10.1086/427175). arXiv: [astro-ph/0306130](https://arxiv.org/abs/astro-ph/0306130) (cit. on p. 170).
- Ghez, A. M., Salim, S., Weinberg, N. N., Lu, J. R., Do, T., Dunn, J. K., Matthews, K., Morris, M., Yelda, S., Becklin, E. E., Kremenek, T., Milosavljevic, M., and Naiman, J. (2008). “Measuring Distance and Properties of the Milky Way’s Central Supermassive Black Hole with Stellar Orbits”. *Astrophys. J.* 689, pp. 1044–1062. DOI: [10.1086/592738](https://doi.org/10.1086/592738). arXiv: [0808.2870](https://arxiv.org/abs/0808.2870) [[astro-ph](https://arxiv.org/abs/astro-ph)] (cit. on p. 128).
- Gibbons, G. W. and Hawking, S. W. (1977). “Action Integrals and Partition Functions in Quantum Gravity”. *Phys. Rev. D* 15, pp. 2752–2756. DOI: [10.1103/PhysRevD.15.2752](https://doi.org/10.1103/PhysRevD.15.2752) (cit. on p. 466).
- Gil-Marín, Héctor et al. (SDSS Collaboration, 40 authors) (2020). “The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic power spectrum between redshifts 0.6 and 1.0”. *Mon. Not. Roy. Astron. Soc.* 498, pp. 2492–2531. DOI: [10.1093/mnras/staa2455](https://doi.org/10.1093/mnras/staa2455). arXiv: [2007.08994](https://arxiv.org/abs/2007.08994) [[astro-ph](https://arxiv.org/abs/astro-ph).[v1](https://arxiv.org/abs/2007.08994v1)] (cit. on pp. 235, 857, 858, 894, 913).
- Gillessen, S., Eisenhauer, F., Trippe, S., Alexander, T., Genzel, R., Martins, F., and Ott, T. (2009). “Monitoring stellar orbits around the Massive Black Hole in the Galactic Center”. *Astrophys. J.* 692, pp. 1075–1109. DOI: [10.1088/0004-637X/692/2/1075](https://doi.org/10.1088/0004-637X/692/2/1075). arXiv: [0810.4674](https://arxiv.org/abs/0810.4674) [[astro-ph](https://arxiv.org/abs/astro-ph)] (cit. on p. 128).
- Gnedin, M. L. and Gnedin, N. Y. (1993). “Destruction of the Cauchy horizon in the Reissner-Nordström black hole”. *Class. Quant. Grav.* 10, pp. 1083–1102 (cit. on p. 649).
- Goldberg, J. N., Macfarlane, A. J., Newman, E. T., Rohrlich, F., and Sudarshan, E. C. G. (1967). “Spin-spherical harmonics and $\tilde{\sigma}$ ”. *J. Math. Phys.* 8, pp. 2155–61. DOI: [10.1063/1.1705135](https://doi.org/10.1063/1.1705135) (cit. on p. 981).
- Goldman, M. (1984). *The Demon in the Aether: The Story of James Clerk Maxwell*. Adam Hilger (cit. on p. 10).
- Goldstone, J., Salam, A., and Weinberg, S. (1962). “Broken Symmetries”. *Phys. Rev.* 127, pp. 965–970. DOI: [10.1103/PhysRev.127.965](https://doi.org/10.1103/PhysRev.127.965) (cit. on pp. 1228, 1229).
- Goldwirth, D. S. and Tsvi, P. (1987). “Gravitational collapse of massless scalar field and cosmic censorship”. *Phys. Rev. D* 36, pp. 3575–3581 (cit. on p. 649).
- Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Table of Integrals, Series, and Products*. London: Academic Press (cit. on pp. 1112, 1113).
- Grassmann, H. (1862). *Die Ausdehnungslehre. Vollständig und in strenger Form begründet*. Berlin: Enslin (cit. on pp. 330, 331).
- (1877). “Der Ort der Hamilton’schen Quaternionen in der Audehnungslehre”. *Math. Ann.* 12, pp. 375–386. DOI: [10.1007/BF01444648](https://doi.org/10.1007/BF01444648) (cit. on pp. 330, 332, 1206).

- Graves, J. C. and Brill, D. R. (1960). “Oscillatory Character of Reissner-Nordström Metric for an Ideal Charged Wormhole”. *Phys. Rev.* 120, pp. 1507–1513. DOI: [10.1103/PhysRev.120.1507](https://doi.org/10.1103/PhysRev.120.1507) (cit. on p. 192).
- Gribov, V. N. and Pontecorvo, B. (1969). “Neutrino astronomy and lepton charge”. *Phys. Lett. B* 28, pp. 493–496 (cit. on p. 1211).
- Groenewold, H. J. (1946). “On the principles of elementary quantum mechanics”. *Physica* 12, pp. 405–460. DOI: [10.1016/S0031-8914\(46\)80059-4](https://doi.org/10.1016/S0031-8914(46)80059-4) (cit. on p. 121).
- Gross, D. J., Harvey, J. A., Martinec, E., and Rohm, R. (1985). “Heterotic string theory (I). The free heterotic string”. *Nucl. Phys. B* 256, pp. 253–284. DOI: [10.1016/0550-3213\(85\)90394-3](https://doi.org/10.1016/0550-3213(85)90394-3) (cit. on pp. 1028, 1030, 1033, 1034).
- Gross, D. J. and Wilczek, F. (1973a). “Asymptotically Free Gauge Theories. I”. *Phys. Rev. D* 8, pp. 3633–3652. DOI: [10.1103/PhysRevD.8.3633](https://doi.org/10.1103/PhysRevD.8.3633) (cit. on p. 1096).
- (1973b). “Ultraviolet behavior of non-abelian gauge theories”. *Phys. Rev. Lett.* 30, pp. 1343–1346. DOI: [10.1103/PhysRevLett.30.1343](https://doi.org/10.1103/PhysRevLett.30.1343) (cit. on p. 1199).
- Grumiller, D., Kummer, W., and Vassilevich, D. V. (2002). “Dilaton gravity in two-dimensions”. *Phys. Rept.* 369, pp. 327–430. DOI: [10.1016/S0370-1573\(02\)00267-3](https://doi.org/10.1016/S0370-1573(02)00267-3). arXiv: [hep-th/0204253](https://arxiv.org/abs/hep-th/0204253) (cit. on p. 314).
- Gullstrand, A. (1922). “Allgemeine Lösung des statischen Einkörperproblems in der Einsteinschen Gravitationstheorie”. *Arkiv. Mat. Astron. Fys.* 16(8), pp. 1–15 (cit. on pp. 143, 568).
- Guth, A. H. (1981). “Inflationary universe: A possible solution to the horizon and flatness problems”. *Phys. Rev. D* 23, pp. 347–356. DOI: [10.1103/PhysRevD.23.347](https://doi.org/10.1103/PhysRevD.23.347) (cit. on p. 260).
- Hahn, O., Angulo, R. E., and Abel, T. (2015). “The properties of cosmic velocity fields”. *Mon. Not. Roy. Astron. Soc.* 454, pp. 3920–3937. DOI: [10.1093/mnras/stv2179](https://doi.org/10.1093/mnras/stv2179). arXiv: [1404.2280](https://arxiv.org/abs/1404.2280) [[astro-ph.CO](#)] (cit. on p. 976).
- Hamilton, A. J. S. (2011). *Towards a general description of the interior structure of rotating black holes*. DOI: [10.48550/arXiv.hep-th/0509097](https://doi.org/10.48550/arXiv.hep-th/0509097). arXiv: [1108.3512](https://arxiv.org/abs/1108.3512) [[gr-qc](#)] (cit. on p. 641).
- (2023). “The Supergeometric Algebra”. *Adv. App. Clifford Algebras* 33, 12. DOI: [10.1007/s00006-022-01257-5](https://doi.org/10.1007/s00006-022-01257-5). arXiv: [2212.11998](https://arxiv.org/abs/2212.11998) [[math-ph](#)] (cit. on p. 1203).
- Hamilton, A. J. S. and Avelino, P. P. (2010). “The physics of the relativistic counter-streaming instability that drives mass inflation inside black holes”. *Phys. Rept.* 495, pp. 1–32. DOI: [10.1016/j.physrep.2010.06.002](https://doi.org/10.1016/j.physrep.2010.06.002). arXiv: [0811.1926](https://arxiv.org/abs/0811.1926) [[gr-qc](#)] (cit. on pp. 631, 632, 639–641, 644–646).
- Hamilton, A. J. S. and Lisle, J. P. (2008). “The river model of black holes”. *Am. J. Phys.* 76, pp. 519–532. DOI: [10.1119/1.2830526](https://doi.org/10.1119/1.2830526). arXiv: [gr-qc/0411060](https://arxiv.org/abs/gr-qc/0411060) (cit. on p. 131).
- Hamilton, A. J. S. and McMaken, T. (2023). “Unification of the four forces in the Spin(11,1) geometric algebra”. *Physica Scripta* 98, 085306. DOI: [10.1088/1402-4896/acdaff](https://doi.org/10.1088/1402-4896/acdaff). arXiv: [2307.01243](https://arxiv.org/abs/2307.01243) [[hep-th](#)] (cit. on pp. 1205, 1216, 1228, 1250).
- Hamilton, A. J. S. and Polhemus, G. (2010). “Stereoscopic visualization in curved spacetime: seeing deep inside a black hole”. *New J. Phys.* 12, 123027. DOI: [10.1088/1367-2630/12/12/123027](https://doi.org/10.1088/1367-2630/12/12/123027). arXiv: [1012.4043](https://arxiv.org/abs/1012.4043) [[gr-qc](#)] (cit. on pp. 75, 171, 717).
- Hamilton, A. J. S. and Pollack, S. E. (2005a). “Inside charged black holes. I: Baryons”. *Phys. Rev. D* 71, 084031. DOI: [10.1103/PhysRevD.71.084031](https://doi.org/10.1103/PhysRevD.71.084031). arXiv: [gr-qc/0411061](https://arxiv.org/abs/gr-qc/0411061) (cit. on pp. 632, 637).

- Hamilton, A. J. S. and Pollack, S. E. (2005b). “Inside charged black holes. II: Baryons plus dark matter”. *Phys. Rev. D* 71, 084032. DOI: [10.1103/PhysRevD.71.084032](https://doi.org/10.1103/PhysRevD.71.084032). arXiv: [gr-qc/0411062](https://arxiv.org/abs/gr-qc/0411062) (cit. on p. 632).
- Hammond, R. T. (2002). “Torsion gravity”. *Rept. Prog. Phys.* 65, pp. 599–649. DOI: [10.1088/0034-4885/65/5/201](https://doi.org/10.1088/0034-4885/65/5/201) (cit. on p. 443).
- Hansen, J., Khokhlov, A., and Novikov, I. (2005). “Physics of the interior of a spherical, charged black hole with a scalar field”. *Phys. Rev. D* 71, 064013. DOI: [10.1103/PhysRevD.71.064013](https://doi.org/10.1103/PhysRevD.71.064013). arXiv: [gr-qc/0501015](https://arxiv.org/abs/gr-qc/0501015) (cit. on pp. 646, 649).
- Harrison, E. R. (1970). “Fluctuations at the Threshold of Classical Cosmology”. *Phys. Rev. D* 1, pp. 2726–2730. DOI: [10.1103/PhysRevD.1.2726](https://doi.org/10.1103/PhysRevD.1.2726) (cit. on p. 855).
- Harvey, J. A., Ramond, P., and Reiss, D. B. (1980). “ CP violation and mass relations in SO_{10} ”. *Phys. Lett. B* 92, pp. 309–311. DOI: [10.1016/0370-2693\(80\)90270-1](https://doi.org/10.1016/0370-2693(80)90270-1) (cit. on pp. 1229, 1240).
- Hasert, F. J. et al. (Gargamelle Collaboration, 55 authors) (1974). “Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment”. *Nuc. Phys. B* 73, pp. 1–22. DOI: [10.1016/0550-3213\(74\)90038-8](https://doi.org/10.1016/0550-3213(74)90038-8) (cit. on p. 1199).
- Hawking, S. W. (1974). “Black hole explosions?” *Nature* 248, pp. 30–31 (cit. on pp. 631, 651).
- Hawking, S. W. and Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge, England: Cambridge University Press (cit. on pp. 165, 171, 258, 545, 563).
- Hawkins, T. (2000). *Emergence of the theory of Lie groups: an essay in the history of mathematics 1869–1926*. New York: Springer-Verlag. ISBN: 978-0-387-98963-1 (cit. on p. 1015).
- Hehl, F. W. (2012). “Gauge Theory of Gravity and Spacetime”. arXiv: [1204.3672 \[gr-qc\]](https://arxiv.org/abs/1204.3672) (cit. on p. 443).
- Hehl, F. W., Heyde, P. von der, and Kerlick, G. D. (1976). “General relativity with spin and torsion: Foundations and prospects”. *Rev. Mod. Phys.* 48, pp. 393–416 (cit. on p. 313).
- Hehl, F. W., McCrea, J. D., Mielke, E. W., and Ne’eman, Y. (1995). “Metric affine gauge theory of gravity: Field equations, Noether identities, world spinors, and breaking of dilation invariance”. *Phys. Rept.* 258, pp. 1–171. DOI: [10.1016/0370-1573\(94\)00111-F](https://doi.org/10.1016/0370-1573(94)00111-F). arXiv: [gr-qc/9402012](https://arxiv.org/abs/gr-qc/9402012) (cit. on p. 446).
- Held, A. (1974). “A formalism for the investigation of algebraically special metrics. I”. *Commun. Math. Phys.* 37, pp. 311–326 (cit. on p. 322).
- Hestenes, D. (1966). *Space-Time Algebra*. Gordon & Breach. DOI: [10.1007/978-3-319-18413-5](https://doi.org/10.1007/978-3-319-18413-5) (cit. on pp. 330, 1206).
- Hestenes, D. and Sobczyk, G. (1987). *Clifford Algebra to Geometric Calculus*. D. Reidel Publishing Company. DOI: [10.1007/978-94-009-6292-7](https://doi.org/10.1007/978-94-009-6292-7) (cit. on pp. 330, 1206).
- Higgs, P. W. (1964). “Broken Symmetries and the Masses of Gauge Bosons”. *Phys. Rev. Lett.* 13, pp. 508–509. DOI: [10.1103/PhysRevLett.13.508](https://doi.org/10.1103/PhysRevLett.13.508) (cit. on pp. 1199, 1229, 1230).
- Hilbert, D. (1915). “Die Grundlagen der Physik”. *Konigl. Gesell. d. Wiss. Göttingen, Nachr., Math.-Phys. Kl.*, pp. 395–407 (cit. on pp. 53, 312, 418, 436).
- Hilditch, D. (2013). “An Introduction to Well-posedness and Free-evolution”. *Int. J. Mod. Phys. A* 28, 1340015. DOI: [10.1142/S0217751X13400150](https://doi.org/10.1142/S0217751X13400150). arXiv: [1309.2012 \[gr-qc\]](https://arxiv.org/abs/1309.2012) (cit. on p. 536).
- Hiley, B. J. (2021). “The Moyal-Dirac controversy revisited”. *Journal and Proceedings of the Royal Society of New South Wales* 154, pp. 139–160. DOI: [doi/10.3316/informit.307829650593040](https://doi.org/10.3316/informit.307829650593040) (cit. on p. 121).

- Hinshaw, G., Larson, D., Komatsu, E., Spergel, D. N., Bennett, C. L., Dunkley, J., Nolta, M. R., Halpern, M., Hill, R. S., Odegard, N., Page, L., Smith, K. M., Weiland, J. L., Gold, B., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., Wollack, E., and Wright, E. L. (2012). “Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results”. arXiv: [1212.5226 \[astro-ph.CO\]](#) (cit. on pp. [233](#), [246](#), [258](#)).
- Hod, S. and Piran, T. (1997). “Critical behaviour and universality in gravitational collapse of a charged scalar field”. *Phys. Rev. D* 55, pp. 3485–3496. DOI: [10.1103/PhysRevD.55.3485](#). arXiv: [gr-qc/9606093](#) (cit. on p. [649](#)).
- (1998a). “Mass-inflation in dynamical gravitational collapse of a charged scalar-field”. *Phys. Rev. Lett.* 81, pp. 1554–1557. DOI: [10.1103/PhysRevLett.81.1554](#). arXiv: [gr-qc/9803004](#) (cit. on pp. [646](#), [649](#)).
- (1998b). “The inner structure of black holes”. *Gen. Rel. Grav.* 30, 1555. DOI: [10.1023/A:1026654519980](#). arXiv: [gr-qc/9902008](#) (cit. on pp. [646](#), [649](#)).
- Hu, W., Seljak, U., White, M., and Zaldarriaga, M. (1998). “Complete treatment of CMB anisotropies in a FRW universe”. *Phys. Rev. D* 57, pp. 3290–3301. DOI: [10.1103/physrevd.57.3290](#). arXiv: [astro-ph/9709066](#) (cit. on p. [796](#)).
- Hu, W. and White, M. J. (1997). “CMB anisotropies: Total angular momentum method”. *Phys. Rev. D* 56, pp. 596–615. DOI: [10.1103/PhysRevD.56.596](#). arXiv: [astro-ph/9702170](#) (cit. on pp. [964](#), [965](#)).
- Hubble, E. P. (1929). “A relation between distance and radial velocity among extra-galactic nebulae”. *Proc. Nat. Acad. Sci.* 15, pp. 168–173 (cit. on p. [229](#)).
- Hui, L., Ostriker, J. P., Tremaine, S., and Witten, E. (2017). “Ultralight scalars as cosmological dark matter”. *Phys. Rev. D* 95, 043541. DOI: [10.1103/physrevd.95.043541](#). arXiv: [1610.08297 \[astro-ph.CO\]](#) (cit. on p. [1244](#)).
- Hummer, D. G. (1994). “Total Recombination and Energy Loss Coefficients for Hydrogenic Ions at Low Density for $10 \leq T_e/Z^2 \leq 10^7$ K”. *Mon. Not. Roy. Astron. Soc.* 268, pp. 109–112 (cit. on p. [885](#)).
- Hummer, D. G. and Storey, P. J. (1998). “Recombination of helium-like ions — I. Photoionization cross-sections and total recombination and cooling coefficients for atomic helium”. *Mon. Not. Roy. Astron. Soc.* 297, pp. 1073–1078. DOI: [10.1046/j.1365-8711.1998.2970041073.x](#) (cit. on p. [885](#)).
- Husain, V. and Olivier, M. (2001). “Scalar field collapse in three-dimensional AdS spacetime”. *Class. Quant. Grav.* 18, pp. L1–L10. arXiv: [gr-qc/0008060](#) (cit. on p. [649](#)).
- Ivezić, Ž. et al. (LSST Collaboration, 313 authors) (2019). “LSST: from Science Drivers to Reference Design and Anticipated Data Products”. *Astrophys. J.* 873, 11. DOI: [10.3847/1538-4357/ab042c](#). arXiv: [0805.2366 \[astro-ph\]](#) (cit. on p. [1206](#)).
- Jarkovská, K., Malinský, M., Mede, T., and Susič, V. (2022). “Quantum nature of the minimal potentially realistic SO(10) Higgs model”. *Phys. Rev. D* 105, 095003. DOI: [10.1103/physrevd.105.095003](#). arXiv: [2109.06784 \[hep-ph\]](#) (cit. on p. [1230](#)).
- Jarosik, N., Bennett, C. L., Dunkley, J., Gold, B., Greason, M. R., Halpern, M., Hill, R. S., Hinshaw, G., Kogut, A., Komatsu, E., Larson, D., Limon, M., Meyer, S. S., Nolta, M. R., Odegard, N., Page, L., Smith, K. M., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E., and Wright, E. L. (2011). “Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic

- Results”. *Astrophys. J. Suppl.* 192, 14. DOI: [10.1088/0067-0049/192/2/14](https://doi.org/10.1088/0067-0049/192/2/14). arXiv: [1001.4744](https://arxiv.org/abs/1001.4744) [[astro-ph.CO](#)] (cit. on p. [232](#)).
- Jones, P. (2008). “The general relativistic infinite plane”. *Am. J. Phys.* 76, pp. 73–78. DOI: [10.1119/1.2800354](https://doi.org/10.1119/1.2800354). arXiv: [0708.2906](https://arxiv.org/abs/0708.2906) (cit. on p. [630](#)).
- Kadastik, M., Kannike, K., and Raidal, M. (2009). “Dark Matter as the signal of Grand Unification”. *Phys. Rev. D* 80, 085020. DOI: [10.1103/PhysRevD.80.085020](https://doi.org/10.1103/PhysRevD.80.085020). arXiv: [0907.1894](https://arxiv.org/abs/0907.1894) [[hep-ph](#)] (cit. on p. [1244](#)).
- (2010). “Matter parity as the origin of scalar dark matter”. *Phys. Rev. D* 81, 015002. DOI: [10.1103/physrevd.81.015002](https://doi.org/10.1103/physrevd.81.015002). arXiv: [0903.2475](https://arxiv.org/abs/0903.2475) [[hep-ph](#)] (cit. on p. [1244](#)).
- Kagramanova, V., Kunz, J., Hackmann, E., and Lammerzahl, C. (2010). “Analytic treatment of complete and incomplete geodesics in Taub-NUT space-times”. *Phys. Rev. D* 81, 124044. DOI: [10.1103/PhysRevD.81.124044](https://doi.org/10.1103/PhysRevD.81.124044). arXiv: [1002.4342](https://arxiv.org/abs/1002.4342) [[gr-qc](#)] (cit. on pp. [667](#), [672](#)).
- Kaiser, N. (1984). “On the spatial correlations of Abell clusters”. *Astrophys. J. Lett.* 284, pp. L9–L12. DOI: [10.1086/184341](https://doi.org/10.1086/184341) (cit. on p. [857](#)).
- Kasner, E. (1921). “Geometrical Theorems on Einstein’s Cosmological Equations”. *Am. J. Math.* 43, pp. 217–221. DOI: [10.2307/2370192](https://doi.org/10.2307/2370192) (cit. on pp. [528](#), [530](#), [630](#)).
- Katz, V. J. (1979). “The History of Stokes’ Theorem”. *Mathematics Magazine* 52, pp. 146–156. DOI: [10.2307/2690275](https://doi.org/10.2307/2690275) (cit. on p. [402](#)).
- Kayser, B. (1981). “On the quantum mechanics of neutrino oscillation”. *Phys. Rev. D* 24, pp. 110–116. DOI: [10.1103/PhysRevD.24.110](https://doi.org/10.1103/PhysRevD.24.110) (cit. on p. [1212](#)).
- Kazakov, D. I. (2000). “Beyond the standard model: In search of supersymmetry”. *2000 European School of high-energy physics, Caramulo, Portugal, 20 Aug-2 Sep 2000: Proceedings*, pp. 125–199. DOI: [10.48550/arXiv.hep-ph/0012288](https://doi.org/10.48550/arXiv.hep-ph/0012288). arXiv: [hep-ph/0012288](https://arxiv.org/abs/hep-ph/0012288) (cit. on p. [1246](#)).
- Keisler, R. et al. (SPT Collaboration, 49 authors) (2011). “A Measurement of the Damping Tail of the Cosmic Microwave Background Power Spectrum with the South Pole Telescope”. *Astrophys. J.* 743, 28. DOI: [10.1088/0004-637X/743/1/28](https://doi.org/10.1088/0004-637X/743/1/28). arXiv: [1105.3182](https://arxiv.org/abs/1105.3182) [[astro-ph.CO](#)] (cit. on p. [233](#)).
- Kerr, R. P. (1963). “Gravitational field of a spinning mass as an example of algebraically special metrics”. *Phys. Rev. Lett.* 11, pp. 237–238. DOI: [10.1103/PhysRevLett.11.237](https://doi.org/10.1103/PhysRevLett.11.237) (cit. on p. [212](#)).
- (2009). “Discovering the Kerr and Kerr-Schild metrics”. *The Kerr spacetime: rotating black holes in general relativity*. Ed. by Wiltshire, D. L., Visser, M., and Scott, S. Cambridge, England: Cambridge University Press, pp. 38–72. arXiv: [0706.1109](https://arxiv.org/abs/0706.1109) [[gr-qc](#)] (cit. on p. [212](#)).
- King, S. F., Pascoli, S., Zhou, Y.-L., and Turner, J. (2021). “Confronting SO(10) GUTs with proton decay and gravitational waves”. *JHEP* 10, p. 225. DOI: [10.1007/jhep10\(2021\)225](https://doi.org/10.1007/jhep10(2021)225). arXiv: [2106.15634](https://arxiv.org/abs/2106.15634) [[hep-ph](#)] (cit. on pp. [998](#), [1240](#), [1246](#), [1249](#), [1250](#)).
- Klein, O. and Nishina, Y. (1929). “Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac”. *Zeitschrift für Physik* 52, pp. 853–868 (cit. on p. [1171](#)).
- Kolb, E. W. and Turner, M. S. (1990). “The Early universe”. *Front. Phys.* 69. DOI: [10.1201/9780429492860](https://doi.org/10.1201/9780429492860) (cit. on p. [267](#)).
- Kormendy, J. and Gebhardt, K. (2001). “Supermassive Black Holes in Nuclei of Galaxies”. *AIP Conf. Proc.* 586, pp. 363–381. DOI: [10.1063/1.1419581](https://doi.org/10.1063/1.1419581). arXiv: [astro-ph/0105230](https://arxiv.org/abs/astro-ph/0105230) (cit. on p. [128](#)).

- Kramers, H. A. (1926). “Wellenmechanik und halbzahlige Quantisierung”. *Zeitschrift für Physik* 39, pp. 828–840. DOI: [10.1007/BF01451751](https://doi.org/10.1007/BF01451751) (cit. on p. 900).
- Krasnov, K. (2022a). *Geometry of Spin(10) Symmetry Breaking*. arXiv: [2209.05088](https://arxiv.org/abs/2209.05088) [[hep-th](#)] (cit. on p. 1230).
- (2022b). “Spin(11,3), particles, and octonions”. *J. Math. Phys.* 63, 031701. DOI: [10.1063/5.0070058](https://doi.org/10.1063/5.0070058). arXiv: [2104.01786](https://arxiv.org/abs/2104.01786) [[hep-th](#)] (cit. on p. 1219).
- Kreiss, H. O. and Ortiz, O. E. (2002). “Some mathematical and numerical questions connected with first and second order time dependent systems of partial differential equations”. *Lect. Notes Phys.* 604, pp. 359–370. arXiv: [gr-qc/0106085](https://arxiv.org/abs/gr-qc/0106085) (cit. on p. 536).
- Kronz, F. and Lupher, T. (2019). “Quantum Theory and Mathematical Rigor”. *The Stanford Encyclopedia of Philosophy*. Ed. by Zalta, E. N. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/fall2019/entries/qt-nvd/> (cit. on p. 1097).
- Kruskal, M. D. (1960). “Maximal extension of Schwarzschild metric”. *Phys. Rev.* 119, pp. 1743–1745 (cit. on p. 156).
- Landau, L. D. and Lifshitz, E. M. (1975). *The Classical Theory of Fields, Fourth revised English Edition*. Pergamon Press (cit. on p. 374).
- Langacker, P. (2004). “Electroweak physics”. *Intersections of particle and nuclear physics. Proceedings, 8th Conference, CIPANP 2003, New York, USA, May 19-24, 2003*. Vol. 698, pp. 1–12. DOI: [10.1063/1.1664192](https://doi.org/10.1063/1.1664192). arXiv: [hep-ph/0308145](https://arxiv.org/abs/hep-ph/0308145) (cit. on p. 1096).
- Leavitt, H. S. and Pickering, E. C. (1912). “Periods of 25 Variable Stars in the Small Magellanic Cloud”. *Harvard College Observatory Circular* 173, pp. 1–3 (cit. on p. 229).
- Lemaître, G. (1927). “Un univers homogène de masse constante et de rayon croissant, rendant compte de la vitesse radiale des nébuleuses extra-galactiques”. *Annales de la Société Scientifique de Bruxelles* 47, pp. 49–59 (cit. on p. 229, 236).
- (1931). “Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae”. *Mon. Not. Roy. Astron. Soc.* 91, pp. 483–490 (cit. on p. 236).
- Lense, J. and Thirring, H. (1918). “Über den Einfluß der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie”. *Phys. Z.* 19, pp. 156–163 (cit. on p. 780).
- Lorentz, H. A. (1904). “Electromagnetic phenomena in a system moving with any velocity smaller than that of light”. *Proceedings of the Royal Netherlands Academy of Arts and Sciences* 6, pp. 809–831 (cit. on p. 11).
- Luzum, B., Capitaine, N., Fienga, A., Folkner, W., Fukushima, T., Hilton, J., Hohenkerk, C., Krasinsky, G., Petit, G., Pitjeva, E., Soffel, M., and Wallace, P. (2009). “The IAU 2009 system of astronomical constants: the report of the IAU working group on numerical standards for Fundamental Astronomy”. *Celestial Mechanics and Dynamical Astronomy* 110, pp. 293–304 (cit. on p. 781).
- Lyth, D. H. and Riotto, A. (1999). “Particle physics models of inflation and the cosmological density perturbation”. *Physics Reports* 314, pp. 1–146. DOI: [10.1016/s0370-1573\(98\)00128-8](https://doi.org/10.1016/s0370-1573(98)00128-8). arXiv: [hep-ph/9807278](https://arxiv.org/abs/hep-ph/9807278) (cit. on p. 1003).

- Ma, C.-P. and Bertschinger, E. (1995). “Cosmological perturbation theory in the synchronous and conformal Newtonian gauges”. *Astrophys. J.* 455, pp. 7–25. DOI: [10.1086/176550](https://doi.org/10.1086/176550). arXiv: [astro-ph/9506072](https://arxiv.org/abs/astro-ph/9506072) (cit. on p. 953).
- Majorana, E. (1937). “Teoria simmetrica dell’elettrone e del positrone”. *Il Nuovo Cimento* 14, pp. 171–184 (cit. on p. 1209).
- Maki, Z., Nakagawa, M., and Sakata, S. (1962). “Remarks on the Unified Model of Elementary Particles”. *Prog. Theor. Phys.* 28, 870–880. DOI: [10.1143/PTP.28.870](https://doi.org/10.1143/PTP.28.870) (cit. on p. 1211).
- Mandula, J. E. (2015). “Coleman-Mandula theorem”. *Scholarpedia* 10(6), 7476. DOI: [10.4249/scholarpedia.7476](https://doi.org/10.4249/scholarpedia.7476) (cit. on pp. 1216, 1225).
- Mangano, G., Miele, G., Pastor, S., and Peloso, M. (2002). “A Precision calculation of the effective number of cosmological neutrinos”. *Phys. Lett. B* 534, pp. 8–16. DOI: [10.1016/S0370-2693\(02\)01622-2](https://doi.org/10.1016/S0370-2693(02)01622-2). arXiv: [astro-ph/0111408](https://arxiv.org/abs/astro-ph/0111408) (cit. on p. 897).
- Martin-Garcia, J. M. and Gundlach, C. (2003). “Global structure of Choptuik’s critical solution in scalar field collapse”. *Phys. Rev. D* 68, 024011. DOI: [10.1103/PhysRevD.68.024011](https://doi.org/10.1103/PhysRevD.68.024011). arXiv: [gr-qc/0304070](https://arxiv.org/abs/gr-qc/0304070) (cit. on p. 649).
- Mattingly, D. (2005). “Modern Tests of Lorentz Invariance”. *Living Reviews in Relativity* 8.1, 5. DOI: [10.12942/lrr-2005-5](https://doi.org/10.12942/lrr-2005-5). arXiv: [gr-qc/0502097](https://arxiv.org/abs/gr-qc/0502097) (cit. on p. 1238).
- McClintock, J. E., Narayan, R., Davis, S. W., Gou, L., Kulkarni, A., Orosz, J. A., Penna, R. F., Remillard, R. A., and Steiner, J. F. (2011). “Measuring the Spins of Accreting Black Holes”. *Class. Quant. Grav.* 28, 114009. DOI: [10.1088/0264-9381/28/11/114009](https://doi.org/10.1088/0264-9381/28/11/114009). arXiv: [1101.0811](https://arxiv.org/abs/1101.0811) [[astro-ph](https://arxiv.org/abs/astro-ph).HE] (cit. on p. 128).
- Mele, S. (2006). *Measurements of the Running of the Electromagnetic Coupling at LEP*. arXiv: [hep-ex/0610037](https://arxiv.org/abs/hep-ex/0610037) [[hep-ex](https://arxiv.org/abs/hep-ex)] (cit. on p. 1127).
- Mellinger, A. (2009). “A Color All-Sky Panorama Image of the Milky Way”. *Pub. Astron. Soc. Pacific* 121, pp. 1180–1187 (cit. on p. 170).
- Merali, Z. (2017). *A Big Bang in a Little Room: The Quest to Create New Universes*. New York: Basic Books (cit. on p. 608).
- Mermin, D. N. (1989). “What’s Wrong with this Pillow?” *Physics Today* 42(4), pp. 9–11. DOI: [10.1063/1.2810963](https://doi.org/10.1063/1.2810963) (cit. on p. 1095).
- Meyer, E. T., Sparks, W. B., Biretta, J. A., Anderson, J., Sohn, S. T., Marel, R. P. van der, Norman, C., and Nakamura, M. (2013). “Optical proper motion measurements of the M87 jet: New results from the Hubble Space Telescope”. *Astrophys. J.* 774, L21. DOI: [10.1088/2041-8205/774/2/L21](https://doi.org/10.1088/2041-8205/774/2/L21). arXiv: [1308.4633](https://arxiv.org/abs/1308.4633) (cit. on p. 40).
- Michell, J. (1784). “On the Means of Discovering the Distance, Magnitude, &c. of the Fixed Stars, in Consequence of the Diminution of the Velocity of Their Light, in Case Such a Diminution Should be Found to Take Place in any of Them, and Such Other Data Should be Procured from Observations, as Would be Farther Necessary for That Purpose”. *Phil. Trans. Roy. Soc. London* 74, pp. 35–57 (cit. on p. 131).
- Milne, A. (1935). *Relativity, Gravitation and World Structure*. Clarendon Press, Oxford (cit. on p. 236).
- Minkowski, H. (1909). “Raum und Zeit”. *Physikalische Zeitschrift* 10, pp. 75–88 (cit. on p. 13).
- Minkowski, P. (1977). “ $\mu \rightarrow e\gamma$ at a rate of one out of 10^9 muon decays?” *Phys. Lett. B* 67, pp. 421–428. DOI: [https://doi.org/10.1016/0370-2693\(77\)90435-X](https://doi.org/https://doi.org/10.1016/0370-2693(77)90435-X) (cit. on p. 1209).

- Misner, C. W. (1963). “The Flatter Regions of Newman, Unti, and Tamburino’s Generalized Schwarzschild Space”. *J. Math. Phys.* 4, pp. 924–937 (cit. on p. 672).
- Misner, C. W. and Sharp, D. H. (1964). “Relativistic equations for adiabatic, spherically symmetric gravitational collapse”. *Phys. Rev.* 136, B571–B576. DOI: [10.1103/PhysRev.136.B571](https://doi.org/10.1103/PhysRev.136.B571) (cit. on p. 584).
- Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation*. New York: W. H. Freeman and Co. (cit. on pp. 109, 317).
- Moyal, J. E. (1949). “Quantum mechanics as a statistical theory”. *Proceedings of the Cambridge Philosophical Society* 45, pp. 99–124. DOI: [10.1017/S0305004100000487](https://doi.org/10.1017/S0305004100000487) (cit. on p. 121).
- Nagy, G., Ortiz, O. E., and Reula, O. A. (2004). “Strongly hyperbolic second order Einstein’s evolution equations”. *Phys. Rev. D* 70, 044012. DOI: [10.1103/PhysRevD.70.044012](https://doi.org/10.1103/PhysRevD.70.044012). arXiv: [gr-qc/0402123](https://arxiv.org/abs/gr-qc/0402123) (cit. on p. 536).
- Nesti, F. and Percacci, R. (2008). “Graviweak Unification”. *J. Phys.* A41, 075405. DOI: [10.1088/1751-8113/41/7/075405](https://doi.org/10.1088/1751-8113/41/7/075405). arXiv: [0706.3307](https://arxiv.org/abs/0706.3307) [[hep-th](#)] (cit. on p. 1219).
- (2010). “Chirality in unified theories of gravity”. *Phys. Rev. D* 81, p. 025010. DOI: [10.1103/PhysRevD.81.025010](https://doi.org/10.1103/PhysRevD.81.025010). arXiv: [0909.4537](https://arxiv.org/abs/0909.4537) [[hep-th](#)] (cit. on p. 1219).
- Newman, E. T., Couch, E., Chinnapared, K., Exton, A., Prakash, A., and Torrence, R. (1965). “Metric of a rotating, charged mass”. *J. Math. Phys.* 6, pp. 918–919. DOI: [10.1063/1.1704351](https://doi.org/10.1063/1.1704351) (cit. on p. 212).
- Newman, E. T. and Penrose, R. (1962). “An Approach to Gravitational Radiation by a Method of Spin Coefficients”. *J. Math. Phys.* 3, pp. 566–579. DOI: [10.1063/1.1724257](https://doi.org/10.1063/1.1724257) (cit. on pp. 322, 981, 1067).
- (2009). “Spin-coefficient formalism”. *Scholarpedia* 4(6), 7445. URL: https://www.scholarpedia.org/article/Newman_Penrose_formalism (cit. on p. 322).
- Newman, E. T., Tamburino, L. A., and Unti, T. (1963). “Empty-space generalization of the Schwarzschild metric”. *J. Math. Phys.* 4, pp. 915–923 (cit. on pp. 667, 672).
- NIST (2022). “CODATA Internationally recommended 2018 values of the Fundamental Physical Constants”. URL: <https://physics.nist.gov/cuu/Constants/> (cit. on pp. 1127, 1208).
- Noether, E. (1918). “Invariante Variationsprobleme”. *Nachr. D. König. Gesellsch. D. Wiss. Zu Göttingen, Math-phys. Klasse* 1918, pp. 235–257 (cit. on pp. 421, 427).
- Nolte, D. D. (2010). “The tangled tale of phase space”. *Physics Today* 63 (4), pp. 33–38 (cit. on p. 123).
- Nordström, G. (1918). “On the energy of the gravitational field in Einstein’s theory”. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen* 20, pp. 1238–1245 (cit. on p. 192).
- O’Donnell, S. (1983). *Portrait of a Prodigy*. Dublin: Boole Press (cit. on p. 348).
- Oppenheimer, J. R. and Snyder, H. (1939). “On continued gravitational contraction”. *Phys. Rev.* 56, pp. 455–459. DOI: [10.1103/PhysRev.56.455](https://doi.org/10.1103/PhysRev.56.455) (cit. on pp. 165, 166, 600, 601).
- Oren, Y. and Piran, T. (2003). “On the collapse of charged scalar fields”. *Phys. Rev. D* 68, 044013. DOI: [10.1103/PhysRevD.68.044013](https://doi.org/10.1103/PhysRevD.68.044013). arXiv: [gr-qc/0306078](https://arxiv.org/abs/gr-qc/0306078) (cit. on p. 649).
- Ori, A. (1991). “Inner structure of a charged black hole: an exact mass-inflation solution”. *Phys. Rev. Lett.* 67, pp. 789–792 (cit. on pp. 645, 646).
- (1999). “Oscillatory null singularity inside realistic spinning black holes”. *Phys. Rev. Lett.* 83, pp. 5423–5426. DOI: [10.1103/PhysRevLett.83.5423](https://doi.org/10.1103/PhysRevLett.83.5423). arXiv: [gr-qc/0103012](https://arxiv.org/abs/gr-qc/0103012) (cit. on p. 646).

- Padmanabhan, T. (2010). *Gravitation: Foundations and Frontiers*. Cambridge, England: Cambridge University Press. DOI: [10.1017/CBO9780511807787](https://doi.org/10.1017/CBO9780511807787) (cit. on p. 451).
- Page, D. N. (1983). “Comment on ‘Entropy Evaporated by a Black Hole’”. *Phys. Rev. Lett.* 50, p. 1013. DOI: [10.1103/PhysRevLett.50.1013](https://doi.org/10.1103/PhysRevLett.50.1013) (cit. on p. 747).
- Painlevé, P. (1921). “La mécanique classique et la théorie de la relativité”. *Comptes Rendus de l’Académie des sciences (Paris)* 173, pp. 677–680 (cit. on pp. 143, 568).
- Palatini, A. (1919). “Deduzione invariante delle equazioni gravitazionali dal principio di Hamilton”. *Rend. Circ. Mat. Palermo* 43, pp. 203–212 (cit. on p. 439).
- Particle Data Group, 244 authors (2024). “Review of Particle Physics”. *Phys. Rev. D* 110, 030001. DOI: [10.1103/PhysRevD.110.030001](https://doi.org/10.1103/PhysRevD.110.030001) (cit. on pp. 1006, 1200, 1208, 1209).
- Pati, J. C. and Salam, A. (1974). “Lepton number as the fourth color”. *Phys. Rev. D* 10, pp. 275–289. DOI: [10.1103/PhysRevD.10.275](https://doi.org/10.1103/PhysRevD.10.275) (cit. on pp. 1202, 1230, 1231).
- Peebles, P. J. E. (1968). “Recombination of the Primeval Plasma”. *Astrophys. J.* 153, pp. 1–11. DOI: [10.1086/149628](https://doi.org/10.1086/149628) (cit. on pp. 867, 878–880, 884).
- Pelc, O. and Horwitz, L. P. (1997). “Generalization of the Coleman-Mandula theorem to higher dimension”. *J. Math. Phys.* 38, pp. 139–172. DOI: [10.1063/1.531846](https://doi.org/10.1063/1.531846). arXiv: [hep-th/9605147](https://arxiv.org/abs/hep-th/9605147) (cit. on p. 1216).
- Penrose, R. (1959). “The apparent shape of a relativistically moving sphere”. *Proc. Cambridge Phil. Soc.* 55, pp. 137–139. DOI: [10.1017/S0305004100033776](https://doi.org/10.1017/S0305004100033776) (cit. on p. 43).
- (1965). “Gravitational collapse and spacetime singularities”. *Phys. Rev. Lett.* 14, pp. 57–59 (cit. on pp. 545, 562, 737).
- (1968). “Structure of space-time”. *Battelle Rencontres: 1967 lectures in Mathematics and Physics*. Ed. by de Witt-Morette, C. and Wheeler, J. A. W. A. Benjamin, New York, pp. 121–235 (cit. on p. 643).
- (2011). “Conformal treatment of infinity”. *Gen. Rel. Grav.* 43, pp. 901–922. DOI: [10.1007/s10714-010-1110-5](https://doi.org/10.1007/s10714-010-1110-5) (cit. on p. 161).
- Penzias, A. A. and Wilson, R. W. (1965). “A Measurement of Excess Antenna Temperature at 4080 Mc/s”. *Astrophys. J. Lett.* 142, pp. 419–421. DOI: [10.1086/148307](https://doi.org/10.1086/148307) (cit. on p. 231).
- Percacci, R. (1991). “The Higgs phenomenon in quantum gravity”. *Nucl. Phys.* B353, pp. 271–290. DOI: [10.1016/0550-3213\(91\)90510-5](https://doi.org/10.1016/0550-3213(91)90510-5). arXiv: [0712.3545](https://arxiv.org/abs/0712.3545) [[hep-th](https://arxiv.org/abs/hep-th)] (cit. on p. 1219).
- Perlmutter, S. et al. (Supernova Cosmology Project, 32 authors) (1999). “Measurements of Omega and Lambda from 42 high redshift supernovae”. *Astrophys. J.* 517, pp. 565–586. DOI: [10.1086/307221](https://doi.org/10.1086/307221). arXiv: [astro-ph/9812133](https://arxiv.org/abs/astro-ph/9812133) (cit. on p. 231).
- Peskin, M. E. (1997). *Beyond the Standard Model*. arXiv: [hep-ph/9705479](https://arxiv.org/abs/hep-ph/9705479) (cit. on p. 1246).
- Peskin, M. E. and Schroeder, D. V. (1995). *An Introduction to Quantum Field Theory*. Reading, MA: Perseus Books. DOI: [10.1201/9780429503559](https://doi.org/10.1201/9780429503559) (cit. on pp. 1096, 1156, 1196, 1232).
- Planck Collaboration (160 authors) (2020a). “Planck 2018 results. IX. Constraints on primordial non-Gaussianity”. *Astronomy & Astrophysics* 641, A9. DOI: [10.1051/0004-6361/201935891](https://doi.org/10.1051/0004-6361/201935891). arXiv: [1905.05697](https://arxiv.org/abs/1905.05697) [[astro-ph](https://arxiv.org/abs/astro-ph).C0] (cit. on p. 996).
- Planck Collaboration (175 authors) (2020b). “Planck 2018 results. X. Constraints on Inflation”. *Astronomy & Astrophysics* 641, A10. DOI: [10.1051/0004-6361/201833887](https://doi.org/10.1051/0004-6361/201833887). arXiv: [1807.06211](https://arxiv.org/abs/1807.06211) [[astro-ph](https://arxiv.org/abs/astro-ph).C0] (cit. on pp. 996, 997, 1005).

- Planck Collaboration (177 authors) (2018). “Planck 2018 results. VI. Cosmological parameters”. *Astronomy & Astrophysics* 641, A6. DOI: [10.1051/0004-6361/201833910](https://doi.org/10.1051/0004-6361/201833910). arXiv: [1807.06209](https://arxiv.org/abs/1807.06209) [[astro-ph.CO](#)] (cit. on pp. 244, 246, 247, 251, 257–259, 263, 266, 267, 269, 826, 832, 855, 858, 896, 1209).
- Planck Collaboration (194 authors) (2020c). “Planck 2018 results: I. Overview and the cosmological legacy of Planck”. *Astronomy & Astrophysics* 641, A1. DOI: [10.1051/0004-6361/201833880](https://doi.org/10.1051/0004-6361/201833880). arXiv: [1807.06205](https://arxiv.org/abs/1807.06205) [[astro-ph.CO](#)] (cit. on pp. 232, 234).
- Planck Collaboration (260 authors) (2015). “Planck 2015 results. XIII. Cosmological parameters”. arXiv: [1502.01589](https://arxiv.org/abs/1502.01589) [[astro-ph.CO](#)] (cit. on pp. 812, 956).
- Planck Collaboration (276 authors) (2013). “Planck 2013 results. I. Overview of products and scientific results”. arXiv: [1303.5062](https://arxiv.org/abs/1303.5062) [[astro-ph.CO](#)] (cit. on p. 233).
- Poincaré, H. (1905). “On the Dynamics of the Electron”. *Comptes Rendus* 140, pp. 1504–1508 (cit. on p. 11).
- Poisson, E. and Israel, W. (1990). “Internal structure of black holes”. *Phys. Rev. D* 41, pp. 1796–1809. DOI: [10.1103/PhysRevD.41.1796](https://doi.org/10.1103/PhysRevD.41.1796) (cit. on pp. 566, 631, 649, 737, 738).
- Politzer, H. D. (1973). “Reliable perturbative results for strong interactions”. *Phys. Rev. Lett.* 30, pp. 1346–1349. DOI: [10.1103/PhysRevLett.30.1346](https://doi.org/10.1103/PhysRevLett.30.1346) (cit. on p. 1199).
- Polyakov, A. M. (1974). “Particle spectrum in quantum field theory”. *JETP Lett.* 20, pp. 194–195 (cit. on p. 612).
- (1981). “Quantum geometry of bosonic strings”. *Phys. Lett. B* 103, pp. 207–210. DOI: [10.1016/0370-2693\(81\)90743-7](https://doi.org/10.1016/0370-2693(81)90743-7) (cit. on p. 413).
- Pontecorvo, B. (1958). “Inverse beta processes and nonconservation of lepton charge”. *JETP Lett.* 7, pp. 172–173 (cit. on p. 1211).
- Pretorius, F. (2005a). “Evolution of binary black hole spacetimes”. *Phys. Rev. Lett.* 95, 121101. DOI: [10.1103/PhysRevLett.95.121101](https://doi.org/10.1103/PhysRevLett.95.121101). arXiv: [gr-qc/0507014](https://arxiv.org/abs/gr-qc/0507014) (cit. on p. 496).
- (2005b). “Numerical relativity using a generalized harmonic decomposition”. *Class. Quant. Grav.* 22, pp. 425–452. DOI: [10.1088/0264-9381/22/2/014](https://doi.org/10.1088/0264-9381/22/2/014). arXiv: [gr-qc/0407110](https://arxiv.org/abs/gr-qc/0407110) (cit. on pp. 496, 540, 541).
- (2006). “Simulation of binary black hole spacetimes with a harmonic evolution scheme”. *Class. Quant. Grav.* 23, S529–S552. DOI: [10.1088/0264-9381/23/16/S13](https://doi.org/10.1088/0264-9381/23/16/S13). arXiv: [gr-qc/0602115](https://arxiv.org/abs/gr-qc/0602115) (cit. on p. 496).
- Pretorius, F. and Israel, W. (1998). “Quasispherical light cones of the Kerr geometry”. *Class. Quant. Grav.* 15, pp. 2289–2301. DOI: [10.1088/0264-9381/15/8/012](https://doi.org/10.1088/0264-9381/15/8/012). arXiv: [gr-qc/9803080](https://arxiv.org/abs/gr-qc/9803080) (cit. on pp. 735–737).
- Price, R. H. (1972). “Nonspherical perturbations of relativistic gravitational collapse. I. Scalar and gravitational perturbations”. *Phys. Rev.* 5, pp. 2419–2438 (cit. on p. 649).
- Quigg, C. (2007). “Spontaneous Symmetry Breaking as a Basis of Particle Mass”. *Rept. Prog. Phys.* 70, pp. 1019–1054. DOI: [10.1088/0034-4885/70/7/R01](https://doi.org/10.1088/0034-4885/70/7/R01). arXiv: [0704.2232](https://arxiv.org/abs/0704.2232) [[hep-ph](#)] (cit. on p. 1207).
- Raychaudhuri, A. (1955). “Relativistic cosmology. I”. *Phys. Rev.* 98, pp. 1123–1126. DOI: [10.1103/PhysRev.98.1123](https://doi.org/10.1103/PhysRev.98.1123) (cit. on p. 547).
- Regge, T. and Wheeler, J. A. (1957). “Stability of the Schwarzschild singularity”. *Phys. Rev.* 108, pp. 1063–1069 (cit. on p. 156).
- Reissner, H. (1916). “Über die Eigengravitation des elektrischen Feldes nach der Einsteinschen Theorie”. *Annalen der Physik (Leipzig)* 50, pp. 106–120 (cit. on p. 192).

- Riess, A. G., Casertano, S., Yuan, W., Macri, L., Anderson, J., MacKenty, J. W., Bowers, J. B., Clubb, K. I., Filippenko, A. V., Jones, D. O., and Tucker, B. E. (2018). “New Parallaxes of Galactic Cepheids from Spatially Scanning the Hubble Space Telescope: Implications for the Hubble Constant”. *Astrophys. J.* 855, 136. DOI: [10.3847/1538-4357/aaadb7](https://doi.org/10.3847/1538-4357/aaadb7). arXiv: [1801.01120](https://arxiv.org/abs/1801.01120) [[astro-ph.SR](#)] (cit. on pp. 244, 246).
- Riess, A. G., Macri, L., Casertano, S., Lampeitl, H., Ferguson, H. C., Filippenko, A. V., Jha, S. W., Li, W., and Chornock, R. (2011). “A 3% Solution: Determination of the Hubble Constant with the Hubble Space Telescope and Wide Field Camera 3”. *Astrophys. J.* 730, 119. DOI: [10.1088/0004-637X/732/2/129](https://doi.org/10.1088/0004-637X/732/2/129), [10.1088/0004-637X/730/2/119](https://doi.org/10.1088/0004-637X/730/2/119). arXiv: [1103.2976](https://arxiv.org/abs/1103.2976) [[astro-ph.CO](#)] (cit. on pp. 244, 246).
- Riess, Adam G. and Filippenko, Alexei V. et al. (Supernova Search Team, 20 authors) (1998). “Observational evidence from supernovae for an accelerating universe and a cosmological constant”. *Astron. J.* 116, pp. 1009–1038. DOI: [10.1086/300499](https://doi.org/10.1086/300499). arXiv: [astro-ph/9805201](https://arxiv.org/abs/astro-ph/9805201) (cit. on p. 231).
- Robertson, H. P. (1935). “Kinematics and world structure”. *Astrophys. J.* 82, pp. 284–301 (cit. on p. 236).
- (1936a). “Kinematics and world structure II”. *Astrophys. J.* 83, pp. 187–201 (cit. on p. 236).
- (1936b). “Kinematics and world structure IIF”. *Astrophys. J.* 83, pp. 257–271 (cit. on p. 236).
- Rovelli, C. (2007). *Quantum Gravity*. Cambridge, England: Cambridge University Press (cit. on p. 100).
- Sachs, R. K. (1961). “Gravitational waves in general relativity. 6. The outgoing radiation condition”. *Proc. Roy. Soc. London A* 264, pp. 309–338 (cit. on p. 553).
- Sachs, R. K. and Wolfe, A. M. (1967). “Perturbations of a cosmological model and angular variations of the microwave background”. *Astrophys. J.* 147, pp. 73–90. DOI: [10.1007/s10714-007-0448-9](https://doi.org/10.1007/s10714-007-0448-9) (cit. on p. 950).
- Sakai, N., Nakao, K.-I., Ishihara, H., and Kobayashi, M. (2006). “Is it possible to create a universe out of a monopole in the laboratory?” *Phys. Rev. D* 74, 024026. DOI: [10.1103/PhysRevD.74.024026](https://doi.org/10.1103/PhysRevD.74.024026). arXiv: [gr-qc/0602084](https://arxiv.org/abs/gr-qc/0602084) (cit. on p. 612).
- Schienbein, I., Staub, F., Steudtner, T., and Svirina, K. (2019). “Revisiting RGEs for general gauge theories”. *Nuclear Physics B* 939, pp. 1–48. DOI: [10.1016/j.nuclphysb.2018.12.001](https://doi.org/10.1016/j.nuclphysb.2018.12.001). arXiv: [1809.06797](https://arxiv.org/abs/1809.06797) [[hep-ph](#)] (cit. on p. 1246).
- Schwarzschild, K. (1916a). “Über das Gravitationsfeld eines Kugel aus inkompressibler Flüssigkeit nach der Einsteinschen Theorie”. *Sitzungsberichte der Preussische Akademie der Wissenschaften zu Berlin, Klasse für Mathematik, Physik, und Technik* 1916, pp. 424–434. arXiv: [physics/9912033](https://arxiv.org/abs/physics/9912033) (cit. on p. 599).
- (1916b). “Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie”. *Sitzungsberichte der Preussische Akademie der Wissenschaften zu Berlin, Klasse für Mathematik, Physik, und Technik* 1916, pp. 189–196. arXiv: [physics/9905030](https://arxiv.org/abs/physics/9905030) (cit. on p. 134).
- Scolnic, D. M. and Jones, D. O. and Rest, A. and Pan, Y. C. and Chornock, R. and Foley, R. J. and Huber, M. E. and Kessler, R. and Narayan, G. and Riess, A. G. and Rodney, S. et al. (40 authors) (2018). “The Complete Light-curve Sample of Spectroscopically Confirmed SNe Ia from Pan-STARRS1 and Cosmological Constraints from the Combined Pantheon Sample”. *Astrophys. J.* 859, p. 101. DOI: [10.3847/1538-4357/aab9bb](https://doi.org/10.3847/1538-4357/aab9bb). arXiv: [1710.00845](https://arxiv.org/abs/1710.00845) [[astro-ph.CO](#)] (cit. on pp. 229, 230).
- Scott, D. and Moss, A. (2009). “Matter temperature during cosmological recombination”. *Mon. Not. Royal Astr. Soc.* 397, pp. 445–446. DOI: [10.1111/j.1365-2966.2009.14939.x](https://doi.org/10.1111/j.1365-2966.2009.14939.x). arXiv: [0902.3438](https://arxiv.org/abs/0902.3438) [[astro-ph.CO](#)] (cit. on p. 877).

- Seager, S., Sasselov, D. D., and Scott, D. (1999). “A new calculation of the recombination epoch”. *Astrophys. J.* 523, pp. L1–L5. DOI: [10.1086/312250](#). arXiv: [astro-ph/9909275](#) (cit. on pp. 884, 885).
- (2000). “How exactly did the universe become neutral?” *Astrophys. J. Suppl.* 128, pp. 407–430. DOI: [10.1086/313388](#). arXiv: [astro-ph/9912182](#) (cit. on pp. 884, 885).
- Seljak, U. and Zaldarriaga, M. (1996). “A line of sight integration approach to cosmic microwave background anisotropies”. *Astrophys. J.* 469, pp. 437–444. DOI: [10.1086/177793](#). arXiv: [astro-ph/9603033](#) (cit. on pp. 832, 933).
- (1997). “Signature of gravity waves in polarization of the microwave background”. *Phys. Rev. Lett.* 78, pp. 2054–2057. DOI: [10.1103/PhysRevLett.78.2054](#). arXiv: [astro-ph/9609169](#) (cit. on p. 964).
- Senovilla, J. M. M. (1998). “Singularity theorems and their consequences”. *Gen. Rel. Grav.* 30, pp. 701–848. DOI: [10.1023/A%3A1018801101244](#). arXiv: [1801.04912 \[gr-qc\]](#) (cit. on pp. 545, 562).
- Shapiro, I. I. (1964). “Fourth test of general relativity”. *Phys. Rev. Lett.* 13, pp. 789–791. DOI: [10.1103/PhysRevLett.13.789](#) (cit. on p. 87).
- Shibata, M. and Nakamura, T. (1995). “Evolution of three-dimensional gravitational waves: Harmonic slicing case”. *Phys. Rev. D* 52, pp. 5428–5444. DOI: [10.1103/PhysRevD.52.5428](#) (cit. on pp. 496, 536).
- Shinkai, H.-a. (2009). “Formulations of the Einstein equations for numerical simulations”. *J. Korean Phys. Soc.* 54, pp. 2513–2528. DOI: [10.3938/jkps.54.2513](#). arXiv: [0805.0068 \[gr-qc\]](#) (cit. on p. 536).
- Shirokov, D. S. (2018). “Classification of Lie algebras of specific type in complexified Clifford algebras”. *Linear and Multilinear Algebra* 66, pp. 1870–1887. arXiv: [1704.03713 \[math-ph\]](#) (cit. on pp. 343, 345, 1065).
- Slansky, R. (1981). “Group theory for unified model building”. *Phys. Rept.* 79, pp. 1–128. DOI: [10.1016/0370-1573\(81\)90092-2](#) (cit. on p. 1015).
- Smarr, L. and York James W., J. (1978). “Kinematical conditions in the construction of space-time”. *Phys. Rev. D* 17, pp. 2529–2551. DOI: [10.1103/PhysRevD.17.2529](#) (cit. on p. 500).
- Sopuerta, C. F., Spherhake, U., and Laguna, P. (2006). “Hydro-without-hydro framework for simulations of black hole-neutron star binaries”. *Class. Quant. Grav.* 23, S579–S598. DOI: [10.1088/0264-9381/23/16/S15](#). arXiv: [gr-qc/0605018](#) (cit. on p. 496).
- Sorkin, E. and Piran, T. (2001). “The effects of pair creation on charged gravitational collapse”. *Phys. Rev. D* 63, 084006. DOI: [10.1103/PhysRevD.63.084006](#). arXiv: [gr-qc/0009095](#) (cit. on p. 649).
- Stephani, H., Kramer, D., MacCallum, M., Hoenselaers, C., and Herlt, E. (2003). *Exact solutions of Einstein’s field equations, 2nd edition*. Cambridge, England: Cambridge University Press. DOI: [10.1017/CBO9780511535185](#) (cit. on pp. 657, 667).
- Study, E. and Cartan, É. (1908). *Nombres complexes*. Paris, France: Editions Jacques Gabay (2005), pp. 329–468 (cit. on p. 1219).
- Stueckelberg, E. C. G. (1941). “Remarque à propos de la création de paires de particules en théorie de de relativité”. *Helv. Phys. Acta* 14, pp. 588–594 (cit. on p. 1129).
- Susskind, L. (2003). “Black Holes and the Information Paradox”. *Sci. Am.* 13, pp. 18–23 (cit. on p. 131).
- Susskind, L., Thorlacius, L., and Uglum, J. (1993). “The stretched horizon and black hole complementarity”. *Phys. Rev. D* 48, pp. 3743–3761. DOI: [10.1103/PhysRevD.48.3743](#). arXiv: [hep-th/9306069](#) (cit. on pp. 171, 173).

- Szekeres, G. (1960). “On the singularities of a Riemann manifold”. *Publ. Mat. Debrecen* 7, pp. 285–301 (cit. on p. 156).
- Takenaka, A. et al. (The Super-Kamiokande Collaboration, 214 authors) (2020). “Search for Proton Decay via $p \rightarrow e^+ \pi^0$ and $p \rightarrow \mu^+ \pi^0$ with an enlarged fiducial volume in Super-Kamiokande I-IV”. *Phys. Rev. D* 102, 112011. DOI: [10.1103/physrevd.102.112011](https://doi.org/10.1103/physrevd.102.112011). arXiv: [2010.16098 \[hep-ex\]](https://arxiv.org/abs/2010.16098) (cit. on pp. 998, 1250).
- Tanabashi, M. et al. (Particle Data Group, 227 authors) (2018). “2018 Review of particle physics”. *Phys. Rev. D* 98, 030001. DOI: [10.1103/PhysRevD.98.030001](https://doi.org/10.1103/PhysRevD.98.030001) (cit. on pp. 1156, 1247).
- Tangherlini, F. R. (1963). “Schwarzschild field in n dimensions and the dimensionality of space problem”. *Il Nuovo Cimento* 27, pp. 636–651. DOI: [10.1007/BF02784569](https://doi.org/10.1007/BF02784569) (cit. on p. 147).
- Taub, A. H. (1951). “Empty space-times admitting a three parameter group of motions”. *Ann. Math.* 53, pp. 472–490 (cit. on pp. 667, 672).
- Telescope Array Collaboration, 126 authors (2014). “Indications of Intermediate-scale Anisotropy of Cosmic Rays with Energy greater than 57 EeV in the Northern Sky measured with the Surface Detector of the Telescope Array Experiment”. *Astrophys. J.* 790, L21. DOI: [10.1088/2041-8205/790/2/L21](https://doi.org/10.1088/2041-8205/790/2/L21). arXiv: [1404.5890 \[astro-ph.HE\]](https://arxiv.org/abs/1404.5890) (cit. on p. 1249).
- Terrell, J. (1959). “Invisibility of the Lorentz Contraction”. *Phys. Rev.* 116, pp. 1041–1045. DOI: [10.1103/PhysRev.116.1041](https://doi.org/10.1103/PhysRev.116.1041) (cit. on p. 43).
- Thirring, H. (1918). “Republication of: On the formal analogy between the basic electromagnetic equations and Einstein’s gravity equations in first approximation”. *Phys. Z.* 19, pp. 204–205. DOI: [10.1007/s10714-012-1451-3](https://doi.org/10.1007/s10714-012-1451-3) (cit. on p. 780).
- ’t Hooft, G. (1974). “Magnetic monopoles in unified gauge theories”. *Nuclear Physics B* 79, pp. 276–284 (cit. on p. 612).
- Thorne, K. S. (1974). “Disk-accretion onto a black hole. II. Evolution of the hole”. *Astrophys. J.* 191, pp. 507–519. DOI: [10.1086/152991](https://doi.org/10.1086/152991) (cit. on p. 712).
- (1994). *Black holes and time warps: Einstein’s outrageous legacy*. W. W. Norton. 619 pp. DOI: [10.1063/1.2808700](https://doi.org/10.1063/1.2808700) (cit. on pp. 10, 134).
- Tremaine, S. and Gunn, J. E. (1979). “Dynamical Role of Light Neutral Leptons in Cosmology”. *Phys. Rev. Lett.* 42, pp. 407–410 (cit. on p. 1205).
- Tyutin, I. V. (1975). *Gauge Invariance in Field Theory and Statistical Physics in Operator Formalism*. arXiv: [0812.0580 \[hep-th\]](https://arxiv.org/abs/0812.0580) (cit. on p. 1149).
- Unruh, W. G. (1976). “Notes on black hole evaporation”. *Phys. Rev. D* 14, pp. 870–892. DOI: [10.1103/PhysRevD.14.870](https://doi.org/10.1103/PhysRevD.14.870) (cit. on pp. 174, 1116).
- Walker, A. G. (1937). “On Milne’s theory of world-structure”. *Proc. London Math. Soc.* 2 42, pp. 90–127 (cit. on p. 236).
- Wallace, C. S., Hamilton, A. J. S., and Polhemus, G. (2008). “Huge entropy production inside black holes”. arXiv: [0801.4415 \[gr-qc\]](https://arxiv.org/abs/0801.4415) (cit. on pp. 632, 652).
- Weinberg, S. (2004). “The making of the Standard Model”. *The European Physical Journal C* 34, pp. 5–13. DOI: [10.1140/epjc/s2004-01761-1](https://doi.org/10.1140/epjc/s2004-01761-1). arXiv: [hep-ph/0401010](https://arxiv.org/abs/hep-ph/0401010) (cit. on p. 1199).
- Weinberg, S. (1967). “A Model of Leptons”. *Phys. Rev. Lett.* 19, pp. 1264–1266. DOI: [10.1103/PhysRevLett.19.1264](https://doi.org/10.1103/PhysRevLett.19.1264) (cit. on pp. 1199, 1231).

- (1995). *The Quantum Theory of Fields*. Cambridge, England: Cambridge University Press (cit. on p. 1097).
- Weisberg, J. M. and Taylor, J. H. (2005). “Relativistic binary pulsar B1913+16: Thirty years of observations and analysis”. *ASP Conf. Ser.* 328, pp. 25–31. arXiv: [astro-ph/0407149](#) (cit. on p. 788).
- Wentzel, G. (1926). “Eine Verallgemeinerung der Quantenbedingungen für die Zwecke der Wellenmechanik”. *Zeitschrift für Physik* 38, pp. 518–529. DOI: [10.1007/BF01397171](#) (cit. on p. 900).
- Weyl, H. (1917). “Zur Gravitationstheorie”. *Ann. Phys. (Berlin)* 54, pp. 117–145. DOI: [10.1002/andp.19173591804](#) (cit. on p. 192).
- White, S. D. M., Frenk, C. S., and Davis, M. (1983). “Clustering in a Neutrino Dominated Universe”. *Astrophys. J.* 274, pp. L1–L5. DOI: [10.1086/161425](#) (cit. on p. 930).
- Wilczek, F. (1998). “SO(10) marshals the particles”. *Nature* 394, p. 15. DOI: [10.1038/27761](#) (cit. on pp. 1203–1205).
- Will, C. M. (2005). “The confrontation between general relativity and experiment”. *Living Rev. Rel.* 9.3. arXiv: [gr-qc/0510072](#) (cit. on p. 51).
- Wong, W. Y., Moss, A., and Scott, D. (2008). “How well do we understand cosmological recombination?”. *Mon. Not. Roy. Astron. Soc.* 386, pp. 1023–1028. DOI: [10.1111/j.1365-2966.2008.13092.x](#). arXiv: [0711.1357 \[astro-ph\]](#) (cit. on p. 884).
- Xing, Z.-z. (2020). “Flavor structures of charged fermions and massive neutrinos”. *Phys. Rept.* 854, pp. 1–147. arXiv: [1909.09610 \[hep-ph\]](#) (cit. on p. 1211).
- Yamatsu, N. (2017). “Special grand unification”. *Prog. Theor. Expt. Phys.*, 061B01. DOI: [10.1093/ptep/ptx088](#). arXiv: [1704.08827 \[hep-ph\]](#) (cit. on p. 1016).
- Yanagida, T. (1980). “Horizontal Symmetry and Masses of Neutrinos”. *Prog. Theor. Phys.* 64, pp. 1103–1105. DOI: [10.1143/PTP.64.1103](#) (cit. on p. 1209).
- Yang, Y.-B., Liang, J., Bi, Y.-J., Chen, Y., Draper, T., Liu, K.-F., and Liu, Z. (2018). “Proton Mass Decomposition from the QCD Energy Momentum Tensor”. *Phys. Rev. Lett.* 121.21, p. 212001. DOI: [10.1103/PhysRevLett.121.212001](#). arXiv: [1808.08677 \[hep-lat\]](#) (cit. on p. 1209).
- York Jr., J. W. (1979). “Kinematics and dynamics of general relativity”. *Sources of Gravitational Radiation*. Ed. by Smarr, L. L. Cambridge, England: Cambridge University Press, pp. 83–126 (cit. on p. 500).
- York Jr., J. W. (1972). “Role of conformal three geometry in the dynamics of gravitation”. *Phys. Rev. Lett.* 28, pp. 1082–1085. DOI: [10.1103/PhysRevLett.28.1082](#) (cit. on p. 466).
- Zaldarriaga, M. and Seljak, U. (1997). “An all sky analysis of polarization in the microwave background”. *Phys. Rev. D* 55, pp. 1830–1840. DOI: [10.1103/PhysRevD.55.1830](#). arXiv: [astro-ph/9609170](#) (cit. on p. 964).
- (1998). “Gravitational lensing effect on cosmic microwave background polarization”. *Phys. Rev. D* 58, 023003. DOI: [10.1103/PhysRevD.58.023003](#). arXiv: [astro-ph/9803150](#) (cit. on p. 995).
- Zeldovich, Y. B. (1972). “A hypothesis, unifying the structure and the entropy of the Universe”. *Mon. Not. Roy. Astron. Soc.* 160, 1P–3P (cit. on p. 855).
- Zurek, W. H. (1982). “Entropy Evaporated by a Black Hole”. *Phys. Rev. Lett.* 49, pp. 1683–1686 (cit. on p. 747).