

# General Relativity, Black Holes, and Cosmology

Andrew J. S. Hamilton



---

## Contents

<i>Table of contents</i>	<i>page</i> iii
<i>List of illustrations</i>	xviii
<i>List of tables</i>	xxiii
<i>List of exercises and concept questions</i>	xxiv
<i>Legal notice</i>	1
<i>Notation</i>	2
<b>PART ONE FUNDAMENTALS</b>	5
<i>Concept Questions</i>	7
<i>What's important?</i>	9
<b>1 Special Relativity</b>	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	30
1.10 4-vectors	32
1.11 Energy-momentum 4-vector	34
1.12 Photon energy-momentum	36
1.13 What things look like at relativistic speeds	38
1.14 Occupation number, phase-space volume, intensity, and flux	43
1.15 How to program Lorentz transformations on a computer	45

<i>Concept Questions</i>	47
<i>What's important?</i>	49
<b>2 Fundamentals of General Relativity</b>	50
2.1 Motivation	51
2.2 The postulates of General Relativity	52
2.3 Implications of Einstein's principle of equivalence	54
2.4 Metric	56
2.5 Timelike, spacelike, proper time, proper distance	57
2.6 Orthonormal tetrad basis $\gamma_m$	57
2.7 Basis of coordinate tangent vectors $e_\mu$	58
2.8 4-vectors and tensors	59
2.9 Covariant derivatives	62
2.10 Torsion	65
2.11 Connection coefficients in terms of the metric	66
2.12 Torsion-free covariant derivative	67
2.13 Mathematical aside: What if there is no metric?	70
2.14 Coordinate 4-velocity	70
2.15 Geodesic equation	71
2.16 Coordinate 4-momentum	72
2.17 Affine parameter	72
2.18 Affine distance	73
2.19 Riemann tensor	75
2.20 Ricci tensor, Ricci scalar	78
2.21 Einstein tensor	78
2.22 Bianchi identities	78
2.23 Covariant conservation of the Einstein tensor	79
2.24 Einstein equations	79
2.25 Summary of the path from metric to the energy-momentum tensor	80
2.26 Energy-momentum tensor of a perfect fluid	80
2.27 Newtonian limit	81
<b>3 More on the coordinate approach</b>	88
3.1 Weyl tensor	88
3.2 Evolution equations for the Weyl tensor, and gravitational waves	89
3.3 Geodesic deviation	90
<b>4 Action principle for point particles</b>	93
4.1 Principle of least action for point particles	93
4.2 Generalized momentum	95
4.3 Lagrangian for a test particle	95
4.4 Massless test particle	97

4.5	Effective Lagrangian for a test particle	98
4.6	Nice Lagrangian for a test particle	99
4.7	Action for a charged test particle in an electromagnetic field	100
4.8	Symmetries and constants of motion	102
4.9	Conformal symmetries	103
4.10	(Super-)Hamiltonian	106
4.11	Conventional Hamiltonian	107
4.12	Conventional Hamiltonian for a test particle	107
4.13	Effective (super-)Hamiltonian for a test particle with electromagnetism	109
4.14	Nice (super-)Hamiltonian for a test particle with electromagnetism	109
4.15	Derivatives of the action	111
4.16	Hamilton-Jacobi equation	112
4.17	Canonical transformations	112
4.18	Symplectic structure	114
4.19	Symplectic scalar product and Poisson brackets	116
4.20	(Super-)Hamiltonian as a generator of evolution	116
4.21	Infinitesimal canonical transformations	117
4.22	Constancy of phase-space volume under canonical transformations	118
4.23	Poisson algebra of integrals of motion	118
	<i>Concept Questions</i>	120
	<i>What's important?</i>	122
<b>5</b>	<b>Observational Evidence for Black Holes</b>	123
<b>6</b>	<b>Ideal Black Holes</b>	125
6.1	Definition of a black hole	125
6.2	Ideal black hole	126
6.3	No-hair theorem	126
<b>7</b>	<b>Schwarzschild Black Hole</b>	128
7.1	Schwarzschild metric	128
7.2	Stationary, static	129
7.3	Spherically symmetric	130
7.4	Energy-momentum tensor	131
7.5	Birkhoff's theorem	131
7.6	Horizon	132
7.7	Proper time	133
7.8	Redshift	134
7.9	“Schwarzschild singularity”	134
7.10	Weyl tensor	135
7.11	Singularity	135
7.12	Gullstrand-Painlevé metric	137

7.13	Embedding diagram	143
7.14	Schwarzschild spacetime diagram	144
7.15	Gullstrand-Painlevé spacetime diagram	146
7.16	Eddington-Finkelstein spacetime diagram	146
7.17	Kruskal-Szekeres spacetime diagram	147
7.18	Antihorizon	149
7.19	Analytically extended Schwarzschild geometry	149
7.20	Penrose diagrams	153
7.21	Penrose diagrams as guides to spacetime	154
7.22	Future and past horizons	156
7.23	Oppenheimer-Snyder collapse to a black hole	156
7.24	Apparent horizon	157
7.25	True horizon	158
7.26	Penrose diagrams of Oppenheimer-Snyder collapse	159
7.27	Illusory horizon	160
7.28	Collapse of a shell of matter on to a black hole	161
7.29	The illusory horizon and black hole thermodynamics	164
7.30	Rindler space and Rindler horizons	164
7.31	Rindler observers who start at rest, then accelerate	167
7.32	Killing vectors	171
7.33	Killing tensors	175
7.34	Lie derivative	175
<b>8</b>	<b>Reissner-Nordström Black Hole</b>	181
8.1	Reissner-Nordström metric	181
8.2	Energy-momentum tensor	182
8.3	Weyl tensor	183
8.4	Horizons	183
8.5	Gullstrand-Painlevé metric	183
8.6	Radial null geodesics	185
8.7	Finkelstein coordinates	186
8.8	Kruskal-Szekeres coordinates	187
8.9	Analytically extended Reissner-Nordström geometry	190
8.10	Penrose diagram	190
8.11	Antiverse: Reissner-Nordström geometry with negative mass	192
8.12	Ingoing, outgoing	192
8.13	The inflationary instability	193
8.14	The X point	195
8.15	Extremal Reissner-Nordström geometry	196
8.16	Super-extremal Reissner-Nordström geometry	197

8.17	Reissner-Nordström geometry with imaginary charge	198
<b>9</b>	<b>Kerr-Newman Black Hole</b>	200
9.1	Boyer-Lindquist metric	200
9.2	Oblate spheroidal coordinates	201
9.3	Time and rotation symmetries	203
9.4	Ring singularity	203
9.5	Horizons	203
9.6	Angular velocity of the horizon	205
9.7	Ergospheres	205
9.8	Turnaround radius	205
9.9	Antiverse	206
9.10	Sisytube	206
9.11	Extremal Kerr-Newman geometry	206
9.12	Super-extremal Kerr-Newman geometry	208
9.13	Energy-momentum tensor	208
9.14	Weyl tensor	209
9.15	Electromagnetic field	209
9.16	Principal null congruences	209
9.17	Finkelstein coordinates	210
9.18	Doran coordinates	210
9.19	Penrose diagram	213
	<i>Concept Questions</i>	214
	<i>What's important?</i>	216
<b>10</b>	<b>Homogeneous, Isotropic Cosmology</b>	217
10.1	Observational basis	217
10.2	Cosmological Principle	222
10.3	Friedmann-Lemaître-Robertson-Walker metric	223
10.4	Spatial part of the FLRW metric: informal approach	223
10.5	Comoving coordinates	225
10.6	Spatial part of the FLRW metric: more formal approach	226
10.7	FLRW metric	228
10.8	Einstein equations for FLRW metric	228
10.9	Newtonian “derivation” of Friedmann equations	228
10.10	Hubble parameter	230
10.11	Critical density	232
10.12	Omega	232
10.13	Types of mass-energy	233
10.14	Redshifting	235
10.15	Evolution of the cosmic scale factor	237

10.16	Age of the Universe	238
10.17	Conformal time	239
10.18	Looking back along the lightcone	240
10.19	Hubble diagram	240
10.20	Recombination	243
10.21	Horizon	243
10.22	Inflation	246
10.23	Evolution of the size and density of the Universe	248
10.24	Evolution of the temperature of the Universe	248
10.25	Neutrino mass	253
10.26	Occupation number, number density, and energy-momentum	255
10.27	Occupation numbers in thermodynamic equilibrium	258
10.28	Maximally symmetric spaces	263
<b>PART TWO TETRAD APPROACH TO GENERAL RELATIVITY</b>		273
<i>Concept Questions</i>		275
<i>What's important?</i>		277
<b>11</b>	<b>The tetrad formalism</b>	278
11.1	Tetrad	278
11.2	Vierbein	279
11.3	The metric encodes the vierbein	279
11.4	Tetrad transformations	280
11.5	Tetrad vectors and tensors	281
11.6	Index and naming conventions for vectors and tensors	283
11.7	Gauge transformations	284
11.8	Directed derivatives	284
11.9	Tetrad covariant derivative	285
11.10	Relation between tetrad and coordinate connections	287
11.11	Antisymmetry of the tetrad connections	287
11.12	Torsion tensor	287
11.13	No-torsion condition	288
11.14	Tetrad connections in terms of the vierbein	288
11.15	Torsion-free covariant derivative	289
11.16	Riemann curvature tensor	289
11.17	Ricci, Einstein, Bianchi	292
11.18	Expressions with torsion	293
<b>12</b>	<b>Spin and Newman-Penrose tetrads</b>	297
12.1	Spin tetrad formalism	297

12.2	Newman-Penrose tetrad formalism	300
12.3	Weyl tensor	302
12.4	Petrov classification of the Weyl tensor	305
<b>13</b>	<b>The geometric algebra</b>	<b>307</b>
13.1	Products of vectors	308
13.2	Geometric product	309
13.3	Reverse	311
13.4	The pseudoscalar and the Hodge dual	311
13.5	General wedge and dot products of multivectors	312
13.6	Reflection	314
13.7	Rotation	315
13.8	A multivector rotation is an active rotation	318
13.9	A rotor is a spin- $\frac{1}{2}$ object	319
13.10	Generator of a rotation	319
13.11	2D rotations and complex numbers	320
13.12	Quaternions	321
13.13	3D rotations and quaternions	323
13.14	Pauli matrices	325
13.15	Pauli spinors as quaternions, or scaled rotors	326
13.16	Spin axis	328
<b>14</b>	<b>The spacetime algebra</b>	<b>329</b>
14.1	Spacetime algebra	329
14.2	Complex quaternions	331
14.3	Lorentz transformations and complex quaternions	332
14.4	Spatial Inversion ( $P$ ) and Time Reversal ( $T$ )	336
14.5	How to implement Lorentz transformations on a computer	337
14.6	Killing vector fields of Minkowski space	341
14.7	Dirac matrices	344
14.8	Dirac spinors	345
14.9	Dirac spinors as complex quaternions	346
14.10	Non-null Dirac spinor	349
14.11	Null Dirac Spinor	350
<b>15</b>	<b>Geometric Differentiation and Integration</b>	<b>353</b>
15.1	Covariant derivative of a multivector	354
15.2	Riemann tensor of bivectors	356
15.3	Torsion tensor of vectors	357
15.4	Covariant spacetime derivative	357
15.5	Torsion-full and torsion-free covariant spacetime derivative	359
15.6	Differential forms	360

15.7	Wedge product of differential forms	361
15.8	Exterior derivative	362
15.9	An alternative notation for differential forms	363
15.10	Hodge dual form	365
15.11	Relation between coordinate- and tetrad-frame volume elements	366
15.12	Generalized Stokes' theorem	366
15.13	Exact and closed forms	369
15.14	Generalized Gauss' theorem	369
15.15	Dirac delta-function	370
<b>16</b>	<b>Action principle for electromagnetism and gravity</b>	372
16.1	Euler-Lagrange equations for a generic field	373
16.2	Super-Hamiltonian formalism	375
16.3	Conventional Hamiltonian formalism	375
16.4	Symmetries and conservation laws	376
16.5	Electromagnetic action	377
16.6	Electromagnetic action in forms notation	384
16.7	Gravitational action	390
16.8	Variation of the gravitational action	393
16.9	Matter energy-momentum and the Einstein equations with matter	395
16.10	Spin angular-momentum	396
16.11	Trading coordinates and momenta	403
16.12	Lagrangian as opposed to Hamiltonian formulation	405
16.13	Gravitational action in multivector notation	406
16.14	Gravitational action in multivector forms notation	410
16.15	Space+time (3+1) split in multivector forms notation	424
16.16	Loop Quantum Gravity	434
16.17	Bianchi identities in multivector forms notation	440
<b>17</b>	<b>Conventional Hamiltonian (3+1) approach</b>	446
17.1	ADM formalism	447
17.2	ADM gravitational equations of motion	457
17.3	Conformally scaled ADM	463
17.4	Bianchi spacetimes	465
17.5	Friedmann-Lemaître-Robertson-Walker spacetimes	470
17.6	BKL oscillatory collapse	472
17.7	BSSN formalism	480
17.8	Pretorius formalism	484
17.9	$M+N$ split	486
17.10	2+2 split	488

<b>18</b>	<b>Singularity theorems</b>	489
18.1	Congruences	489
18.2	Raychaudhuri equations	491
18.3	Raychaudhuri equations for a timelike geodesic congruence	492
18.4	Raychaudhuri equations for a null geodesic congruence	494
18.5	Sachs optical coefficients	496
18.6	Hypersurface-orthogonality for a timelike congruence	497
18.7	Hypersurface-orthogonality for a null congruence	500
18.8	Focusing theorems	503
18.9	Singularity theorems	504
	<i>Concept Questions</i>	508
	<i>What's important?</i>	509
<b>19</b>	<b>Black hole waterfalls</b>	510
19.1	Tetrads move through coordinates	510
19.2	Gullstrand-Painlevé waterfall	511
19.3	Boyer-Lindquist tetrad	517
19.4	Doran waterfall	519
<b>20</b>	<b>General spherically symmetric spacetimes</b>	525
20.1	Spherical spacetime	525
20.2	Spherical line-element	525
20.3	Rest diagonal line-element	527
20.4	Comoving diagonal line-element	528
20.5	Tetrad connections	529
20.6	Riemann, Einstein, and Weyl tensors	530
20.7	Einstein equations	531
20.8	Choose your frame	532
20.9	Interior mass	532
20.10	Energy-momentum conservation	534
20.11	Structure of the Einstein equations	535
20.12	Comparison to ADM (3+1) formulation	538
20.13	Spherical electromagnetic field	538
20.14	General relativistic stellar structure	539
20.15	Freely-falling dust	541
20.16	Naked singularities in dust collapse	542
20.17	Self-similar spherically symmetric spacetime	545
<b>21</b>	<b>The interiors of spherical black holes</b>	559
21.1	The mechanism of mass inflation	559
21.2	The far future?	562
21.3	Self-similar models of the interior structure of black holes	564

21.4	Instability at outer horizon?	578
<b>22</b>	<b>Ideal rotating black holes</b>	579
22.1	Separable geometries	579
22.2	Horizons	581
22.3	Conditions from Hamilton-Jacobi separability	582
22.4	Electrovac solutions from separation of Einstein's equations	585
22.5	Electrovac solutions of Maxwell's equations	589
22.6	$\Lambda$ -Kerr-Newman boundary conditions	591
22.7	Taub-NUT geometry	594
<b>23</b>	<b>Trajectories in ideal rotating black holes</b>	597
23.1	Hamilton-Jacobi equation	597
23.2	Particle with magnetic charge	599
23.3	Killing vectors and Killing tensor	599
23.4	Turnaround	600
23.5	Constraints on the Hamilton-Jacobi parameters $P_t$ and $P_x$	600
23.6	Principal null congruences	601
23.7	Carter integral $\mathcal{Q}$	602
23.8	Penrose process	605
23.9	Constant latitude trajectories in the Kerr-Newman geometry	605
23.10	Circular orbits in the Kerr-Newman geometry	606
23.11	General solution for circular orbits	607
23.12	Circular geodesics (orbits for particles with zero electric charge)	612
23.13	Null circular orbits	616
23.14	Marginally stable circular orbits	617
23.15	Circular orbits at constant latitude in the Antiverse	617
23.16	Circular orbits at the horizon of an extremal black hole	618
23.17	Equatorial circular orbits in the Kerr geometry	620
23.18	Thin disk accretion	622
23.19	Circular orbits in the Reissner-Nordström geometry	626
23.20	Hypersurface-orthogonal congruences	627
23.21	The principal null and Doran congruences	632
23.22	Pretorius-Israel double-null congruence	634
<b>24</b>	<b>The interiors of rotating black holes</b>	638
24.1	Nonlinear evolution	638
24.2	Focussing along principal null directions	639
24.3	Conformally separable geometries	639
24.4	Conditions from conformal Hamilton-Jacobi separability	639
24.5	Tetrad-frame connections	639
24.6	Inevitability of mass inflation	640

24.7	The black hole particle accelerator	641
	<i>Concept Questions</i>	644
	<i>What's important?</i>	646
<b>25</b>	<b>Perturbations and gauge transformations</b>	647
25.1	Notation for perturbations	647
25.2	Vierbein perturbation	647
25.3	Gauge transformations	648
25.4	Tetrad metric assumed constant	648
25.5	Perturbed coordinate metric	649
25.6	Tetrad gauge transformations	649
25.7	Coordinate gauge transformations	650
25.8	Scalar, vector, tensor decomposition of perturbations	652
<b>26</b>	<b>Perturbations in a flat space background</b>	655
26.1	Classification of vierbein perturbations	655
26.2	Metric, tetrad connections, and Einstein and Weyl tensors	657
26.3	Spin components of the Einstein tensor	660
26.4	Too many Einstein equations?	660
26.5	Action at a distance?	661
26.6	Comparison to electromagnetism	662
26.7	Harmonic gauge	666
26.8	Newtonian (Copernican) gauge	667
26.9	Synchronous gauge	669
26.10	Newtonian potential	670
26.11	Dragging of inertial frames	671
26.12	Quadrupole pressure	674
26.13	Gravitational waves	676
26.14	Energy-momentum carried by gravitational waves	678
	<i>Concept Questions</i>	684
<b>27</b>	<b>An overview of cosmological perturbations</b>	685
<b>28</b>	<b>Cosmological perturbations in a flat Friedmann-Lemaître-Robertson-Walker background</b>	690
28.1	Unperturbed line-element	690
28.2	Comoving Fourier modes	691
28.3	Classification of vierbein perturbations	691
28.4	Residual global gauge freedoms	693
28.5	Metric, tetrad connections, and Einstein tensor	695
28.6	Gauge choices	699
28.7	ADM gauge choices	699
28.8	Conformal Newtonian (Copernican) gauge	699

28.9	Conformal synchronous gauge	705
<b>29</b>	<b>Cosmological perturbations: a simplest set of assumptions</b>	707
29.1	Perturbed FLRW line-element	708
29.2	Energy-momenta of perfect fluids	708
29.3	Entropy conservation at superhorizon scales	711
29.4	Unperturbed background	714
29.5	Generic behaviour of non-baryonic cold dark matter	715
29.6	Generic behaviour of radiation	716
29.7	Equations for the simplest set of assumptions	718
29.8	On the numerical computation of cosmological power spectra	721
29.9	Analytic solutions in various regimes	722
29.10	Superhorizon scales	723
29.11	Radiation-dominated, adiabatic initial conditions	726
29.12	Radiation-dominated, isocurvature initial conditions	730
29.13	Subhorizon scales	731
29.14	Fluctuations that enter the horizon during the matter-dominated epoch	734
29.15	Matter-dominated regime	737
29.16	Baryons post-recombination	738
29.17	Matter with dark energy	739
29.18	Matter with dark energy and curvature	740
29.19	Primordial power spectrum	743
29.20	Matter power spectrum	744
29.21	Nonlinear evolution of the matter power spectrum	747
29.22	Statistics of random fields	748
<b>30</b>	<b>Non-equilibrium processes in the FLRW background</b>	754
30.1	Conditions around the epoch of recombination	755
30.2	Overview of recombination	756
30.3	Energy levels and ionization state in thermodynamic equilibrium	757
30.4	Occupation numbers	761
30.5	Boltzmann equation	761
30.6	Collisions	763
30.7	Non-equilibrium recombination	765
30.8	Recombination: Peebles approximation	768
30.9	Recombination: Seager et al. approximation	773
30.10	Sobolev escape probability	775
<b>31</b>	<b>Cosmological perturbations: the hydrodynamic approximation</b>	777
31.1	Electron-photon (Thomson) scattering	779
31.2	Summary of equations in the hydrodynamic approximation	780
31.3	Standard cosmological parameters	785

*Contents*

xv

31.4	The photon-baryon fluid in the tight-coupling approximation	787
31.5	WKB approximation	789
31.6	Including quadrupole pressure in the momentum conservation equation	790
31.7	Photon diffusion (Silk damping)	791
31.8	Viscous baryon drag damping	792
31.9	Photon-baryon wave equation with dissipation	793
31.10	Baryon loading	794
31.11	Neutrinos	796
<b>32</b>	<b>Cosmological perturbations: Boltzmann treatment</b>	797
32.1	Summary of equations in the Boltzmann treatment	799
32.2	Boltzmann equation in a perturbed FLRW geometry	802
32.3	Non-baryonic cold dark matter	805
32.4	Boltzmann equation for the temperature fluctuation	807
32.5	Spherical harmonics of the temperature fluctuation	809
32.6	The Boltzmann equation for massless particles	809
32.7	Energy-momentum tensor for massless particles	810
32.8	Nonrelativistic electron-photon (Thomson) scattering	810
32.9	The photon collision term for electron-photon scattering	811
32.10	Boltzmann equation for photons	815
32.11	Baryons	815
32.12	Boltzmann equation for relativistic neutrinos	816
32.13	Truncating the photon Boltzmann hierarchy	818
32.14	Massive neutrinos	819
32.15	Appendix: Legendre polynomials	820
<b>33</b>	<b>Fluctuations in the Cosmic Microwave Background</b>	822
33.1	Radiative transfer of CMB photons	822
33.2	Harmonics of the CMB photon distribution	824
33.3	CMB in real space	833
33.4	Observing CMB power	838
33.5	Large-scale CMB fluctuations (Sachs-Wolfe effect)	838
33.6	Radiative transfer of neutrinos	840
33.7	Appendix: Integrals over spherical Bessel functions	843
<b>34</b>	<b>Polarization of the Cosmic Microwave Background</b>	844
34.1	Photon polarization	844
34.2	Spin sign convention	845
34.3	Photon density matrix	846
34.4	Temperature fluctuation for polarized photons	848
34.5	Boltzmann equations for polarized photons	849
34.6	Spherical harmonics of the polarized photon distribution	849

34.7	Vector and tensor Einstein equations	852
34.8	Reality conditions on the polarized photon distribution	852
34.9	Polarized Thomson scattering	853
34.10	Summary of Boltzmann equations for polarized photons	858
34.11	Radiative transfer of the polarized CMB	858
34.12	Harmonics of the polarized CMB photon distribution	860
34.13	Harmonics of the polarized CMB in real space	862
34.14	Polarized CMB power spectra	864
34.15	Appendix: Spin-weighted spherical harmonics	866
<b>PART THREE SPINORS</b>		871
<b>35</b>	<b>The super geometric algebra</b>	873
35.1	Spin basis vectors in 3D	873
35.2	Spin weight	874
35.3	Pauli representation of spin basis vectors	874
35.4	Spinor basis elements	875
35.5	Pauli spinor	876
35.6	Spinor metric	876
35.7	Row basis spinors	877
35.8	Inner products of spinor basis elements	877
35.9	Lowering and raising spinor indices	878
35.10	Scalar products of Pauli spinors	878
35.11	Outer products of spinor basis elements	879
35.12	The 3D super geometric algebra	880
35.13	<i>C</i> -conjugate Pauli spinor	881
35.14	Scalar products of spinors and conjugate spinors	882
35.15	<i>C</i> -conjugate multivectors	883
35.16	No self-conjugate Majorana Pauli spinor	884
<b>36</b>	<b>Super spacetime algebra</b>	885
36.1	Newman-Penrose formalism	885
36.2	Chiral representation of $\gamma$ -matrices	887
36.3	Spinor basis elements	888
36.4	Dirac, Weyl, and Majorana spinors	890
36.5	Spinor scalar product	891
36.6	Super spacetime algebra	894
36.7	<i>C</i> -conjugation	898
36.8	(Anti)symmetry of the scalar product of spinors	903
36.9	Discrete transformations $P, T$	904

36.10	Self-conjugate Majorana spinor	906
36.11	Real, or Majorana, representation of $\gamma$ -matrices	908
<b>37</b>	<b>Geometric Differentiation and Integration of Spinors</b>	<b>912</b>
37.1	Covariant derivative of a Dirac spinor	912
37.2	Covariant derivative in a spinor basis	912
37.3	Ashtekhar variables	913
37.4	Covariant spacetime derivative of a spinor	913
37.5	Covariant spacetime derivative of a true scalar	914
37.6	Stokes' and Gauss' theorems for spinors	914
37.7	Gauss' theorem for true scalars	915
<b>38</b>	<b>Action principle for fields</b>	<b>916</b>
38.1	Dirac spinor field	916
38.2	Dirac field with electromagnetism	919
38.3	<i>CPT</i> -conjugates: Antiparticles	922
38.4	Negative mass particles?	924
38.5	Super-Hamiltonian as generator of evolution	926
38.6	Majorana spinor field	927
38.7	Electromagnetic field	929
38.8	A scalar product of derivatives of multivectors of different grade is a total divergence	936
38.9	Yang-Mills field	936
38.10	Real scalar field	942
38.11	Complex scalar field	944
38.12	Conformally coupled scalar field	946
38.13	Spin- $\frac{3}{2}$ field	947
38.14	(Super)-Hamiltonian	953
38.15	Conventional Hamiltonian	953
	<i>Bibliography</i>	955
	<i>Index</i>	967

---

## 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	31
1.17	The longest proper time between two events is a straight line	34
1.18	Superluminal motion of the M87 jet	39
1.19	The rules of 4-dimensional perspective	41
1.20	Tachyon spacetime diagram	45
2.1	The principle of equivalence implies that gravity curves spacetime	51
2.2	A 2-sphere must be covered with at least two charts	53
2.3	The principle of equivalence implies the gravitational redshift and the gravitational bending of light	55
2.4	Tetrad vectors $\gamma_m$ and tangent vectors $e_\mu$	58
2.5	Derivatives of tangent vectors $e_\mu$ defined by parallel transport	63
2.6	Shapiro time delay	85
2.7	Lensing diagram	86
2.8	The appearance of a source lensed by a point lens	86
4.1	Action principle	94

4.2	Rindler wedge	104
6.1	Fishes in a black hole waterfall	125
7.1	The singularity is not a point	136
7.2	Waterfall model of a Schwarzschild black hole	137
7.3	A body cannot remain rigid as it approaches the Schwarzschild singularity	142
7.4	Embedding diagram of the Schwarzschild geometry	143
7.5	Schwarzschild spacetime diagram	145
7.6	Gullstrand-Painlevé spacetime diagram	146
7.7	Finkelstein spacetime diagram	147
7.8	Kruskal-Szekeres spacetime diagram	148
7.9	Morph Finkelstein to Kruskal-Szekeres spacetime diagram	149
7.10	Embedding diagram of the analytically extended Schwarzschild geometry	150
7.11	Analytically extended Kruskal-Szekeres spacetime diagram	150
7.12	Sequence of embedding diagrams of the analytically extended Schwarzschild geometry	151
7.13	Penrose spacetime diagram	152
7.14	Morph Kruskal-Szekeres to Penrose spacetime diagram	153
7.15	Penrose spacetime diagram of the analytically extended Schwarzschild geometry	153
7.16	Penrose diagram of the Schwarzschild geometry	155
7.17	Penrose diagram of the analytically extended Schwarzschild geometry	155
7.18	Oppenheimer-Snyder collapse of a pressureless star	157
7.19	Spacetime diagrams of Oppenheimer-Snyder collapse	158
7.20	Penrose diagrams of Oppenheimer-Snyder collapse	159
7.21	Penrose diagram of a collapsed spherical star	160
7.22	Visualization of falling into a Schwarzschild black hole	162
7.23	Collapse of a shell on to a pre-existing black hole	163
7.24	Finkelstein spacetime diagram of a shell collapsing on to a pre-existing black hole	164
7.25	Rindler diagram	165
7.26	Penrose Rindler diagram	166
7.27	Spacetime diagram of Minkowski space showing observers who start at rest and then accelerate	168
7.28	Formation of the Rindler illusory horizon	168
7.29	Penrose diagram of Rindler space	169
7.30	Killing vector field on a 2-sphere	173
8.1	Waterfall model of a Reissner-Nordström black hole	184
8.2	Spacetime diagram of the Reissner-Nordström geometry	186
8.3	Finkelstein spacetime diagram of the Reissner-Nordström geometry	187
8.4	Kruskal spacetime diagram of the Reissner-Nordström geometry	188
8.5	Kruskal spacetime diagram of the analytically extended Reissner-Nordström geometry	189
8.6	Penrose diagram of the Reissner-Nordström geometry	191
8.7	Penrose diagram illustrating why the Reissner-Nordström geometry is subject to the inflationary instability	193
8.8	Waterfall model of an extremal Reissner-Nordström black hole	195
8.9	Penrose diagram of the extremal Reissner-Nordström geometry	196
8.10	Waterfall model of an extremal Reissner-Nordström black hole	197

8.11	Penrose diagram of the Reissner-Nordström geometry with imaginary charge $Q$	199
9.1	Geometry of Kerr and Kerr-Newman black holes	202
9.2	Contours of constant $\rho$ , and their normals, in Boyer-Lindquist coordinates	204
9.3	Geometry of extremal Kerr and Kerr-Newman black holes	207
9.4	Geometry of a super-extremal Kerr black hole	208
9.5	Waterfall model of a Kerr black hole	211
9.6	Penrose diagram of the Kerr-Newman geometry	212
10.1	Hubble diagram of Type Ia supernovae	218
10.2	Power spectrum of fluctuations in the CMB	220
10.3	Power spectrum of galaxies	221
10.4	Embedding diagram of the FLRW geometry	224
10.5	Poincaré disk	227
10.6	Newtonian picture of the Universe as a uniform density gravitating ball	229
10.7	Behaviour of the mass-energy density of various species as a function of cosmic time	234
10.8	Distance versus redshift in the FLRW geometry	241
10.9	Spacetime diagram of FRLW Universe	244
10.10	Evolution of redshifts of objects at fixed comoving distance	245
10.11	Cosmic scale factor and Hubble distance as a function of cosmic time	249
10.12	Mass-energy density of the Universe as a function of cosmic time	250
10.13	Temperature of the Universe as a function of cosmic time	253
10.14	A massive fermion flips between left- and right-handed as it propagates through spacetime	255
10.15	Embedding spacetime diagram of de Sitter space	264
10.16	Penrose diagram of de Sitter space	267
10.17	Embedding spacetime diagram of anti de Sitter space	268
10.18	Penrose diagram of anti de Sitter space	270
11.1	Tetrad vectors $\gamma_m$	278
11.2	Derivatives of tetrad vectors $\gamma_m$ defined by parallel transport	285
13.1	Vectors, bivectors, and trivectors	308
13.2	Reflection of a vector through an axis	315
13.3	Rotation of a vector by a bivector	316
13.4	Right-handed rotation of a vector by angle $\theta$	317
14.1	Lorentz boost of a vector by rapidity $\theta$	334
14.2	Killing trajectories in Minkowski space	342
15.1	Partition of unity	367
17.1	Cosmic scale factors in BKL collapse	479
18.1	Expansion, vorticity, and shear	497
18.2	Formation of caustics in a hypersurface-orthogonal timelike congruence	498
18.3	Caustics in the galaxy NGC 474	499
18.4	Formation of caustics in a hypersurface-orthogonal null congruence	502
18.5	Spacetime diagram of the dog-leg proposition	505
18.6	Null boundary of the future of a 2-dimensional spacelike surface	506
19.1	Waterfall model of Schwarzschild and Reissner-Nordström black holes	512
19.2	Waterfall model of a Kerr black hole	523

20.1	Spacetime diagram of a naked singularity in dust collapse	543
21.1	Spacetime diagram illustrating qualitatively the three successive phases of mass inflation	560
21.2	An uncharged baryonic plasma falls into an uncharged spherical black hole	569
21.3	A charged, non-conducting plasma falls into a charged spherical black hole	570
21.4	A charged plasma with near critical conductivity falls into a charged spherical black hole, creating huge entropy inside the horizon	571
21.5	An accreting spherical charged black hole that creates even more entropy inside the horizon	572
21.6	1	573
21.7	1	574
21.8	1	575
21.9	1	576
21.10	1	577
21.11	1	578
22.1	Geometry of a Kerr-NUT black hole	595
23.1	Values of $1/P$ for circular orbits of a charged particle about a Kerr-Newman black hole	610
23.2	Location of stable and unstable circular orbits in the Kerr geometry	613
23.3	Location of stable and unstable circular orbits in the super-extremal Kerr geometry	614
23.4	Radii of null circular orbits for a Kerr black hole	616
23.5	Radii of marginally stable circular orbits for a Kerr black hole	618
23.6	Values of the Hamilton-Jacobi parameter $P_t$ for circular orbits in the equatorial plane of a near-extremal Kerr black hole	621
23.7	Energy and angular momentum on the ISCO	622
23.8	Accretion efficiency of a Kerr black hole	623
23.9	Outgoing and ingoing null coordinates	632
23.10	Pretorius-Israel double-null hypersurface-orthogonal congruence	634
24.1	Particles falling from infinite radius can be either outgoing or ingoing at the inner horizon only up to a maximum latitude	642
26.1	The two polarizations of gravitational waves	676
28.1	Evolution of the tensor potential $h_{ab}$	703
29.1	Evolution of dark matter and radiation in the simple model	719
29.2	Overdensities and velocities in the simple approximation	720
29.3	Regimes in the evolution of fluctuations	722
29.4	Superhorizon scales	723
29.5	Evolution of the scalar potential $\Phi$ at superhorizon scales	725
29.6	Radiation-dominated regime	726
29.7	Evolution of the potential $\Phi$ and the radiation monopole $\Theta_0$	727
29.8	Evolution of the dark matter overdensity $\delta_c$	728
29.9	Subhorizon scales	731
29.10	Growth of the dark matter overdensity $\delta_c$ through matter-radiation equality	733
29.11	Fluctuations that enter the horizon in the matter-dominated regime	735
29.12	Evolution of the potential $\Phi$ and the radiation monopole $\Theta_0$ for long wavelength modes	736
29.13	Matter-dominated regime	737
29.14	Growth factor $g(a)$	742

29.15	Matter power spectrum for the simple model	747
30.1	Ion fractions in thermodynamic equilibrium	759
30.2	Recombination of Hydrogen	770
30.3	Departure coefficients in the recombination of Hydrogen	771
30.4	Recombination of Hydrogen and Helium	774
31.1	Overdensities and velocities in the hydrodynamic approximation	778
31.2	Photon and neutrino multipoles in the hydrodynamic approximation	779
31.3	Evolution of matter and photon monopoles in the hydrodynamic approximation	782
31.4	Matter power spectrum in the hydrodynamic approximation	783
32.1	Overdensities and velocities from a Boltzmann computation	797
32.2	Photon and neutrino multipoles	798
32.3	Evolution of matter and photon monopoles in a Boltzmann computation	800
32.4	Difference $\Psi - \Phi$ in scalar potentials	801
32.5	Matter power spectrum	802
33.1	Visibility function	824
33.2	Factors in the solution of the radiative transfer equation	826
33.3	ISW integrand	827
33.4	CMB transfer functions $T_\ell(\eta_0, k)$ for a selection of harmonics $\ell$	829
33.4	(continued)	830
33.5	Thomson scattering source functions at recombination	831
33.6	CMB transfer functions in the rapid recombination approximation	832
33.7	CMB power spectrum in the rapid recombination approximation	836
33.8	Multipole contributions to the CMB power spectrum	837
34.1	Thomson scattering of polarized light	854
34.2	Angles between photon momentum, scattered photon momentum, and wavevector	856

---

## Tables

1.1	Trip across the Universe	29
10.1	Cosmic inventory	232
10.2	Properties of universes dominated by various species	234
10.3	Evolution of cosmic scale factor in universes dominated by various species	238
10.4	Effective entropy-weighted number of relativistic particle species	252
12.1	Petrov classification of the Weyl tensor	306
17.1	Classification of Bianchi spaces	467
17.2	Bianchi vierbein	474
23.1	Signs of $P_t$ and $P_x$ in various regions of the Kerr-Newman geometry	601

---

## 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	*Proper time, proper distance	31
1.13	Scalar product	33
1.14	The principle of longest proper time	33
1.15	Superluminal jets	39
1.16	The rules of 4-dimensional perspective	41
1.17	Circles on the sky	42
1.18	Lorentz transformation preserves angles on the sky	43
1.19	The aberration of starlight	43
1.20	*Apparent (affine) distance	43
1.21	Brightness of a star	44
1.22	Tachyons	45
2.1	The equivalence principle implies the gravitational redshift of light, Part 1	54
2.2	The equivalence principle implies the gravitational redshift of light, Part 2	55
2.3	*Parallel transport when torsion is present	66
2.4	*Can the metric be Minkowski in the presence of torsion?	68
2.5	Covariant curl and coordinate curl	69
2.6	Covariant divergence and coordinate divergence	69
2.7	Gravitational redshift in a stationary metric	73
2.8	Gravitational redshift in Rindler space	74

2.9	Gravitational redshift in a uniformly rotating space	74
2.10	Derivation of the Riemann tensor	75
2.11	Jacobi identity	79
2.12	Special and general relativistic corrections for clocks on satellites	81
2.13	Equations of motion in weak gravity	82
2.14	Deflection of light by the Sun	84
2.15	Shapiro time delay	84
2.16	Gravitational lensing	85
3.1	Number of Bianchi identities	90
3.2	Wave equation for the Riemann tensor	90
4.1	*Redundant time coordinates?	95
4.2	*Throw a clock up in the air	97
4.3	*Conventional Lagrangian	98
4.4	Geodesics in Rindler space	104
4.5	*Action vanishes along a null geodesic, but its gradient does not	111
4.6	*How many integrals of motion can there be?	119
7.1	Schwarzschild metric in isotropic form	129
7.2	Derivation of the Schwarzschild metric	131
7.3	*Going forwards or backwards in time inside the horizon	133
7.4	*Is the singularity of a Schwarzschild black hole a point?	135
7.5	*Separation between infallers who fall in at different times	136
7.6	Geodesics in the Schwarzschild geometry	138
7.7	General relativistic precession of Mercury	140
7.8	A body cannot remain rigid as it approaches the Schwarzschild singularity	141
7.9	*Penrose diagram of Minkowski space	155
7.10	*Penrose diagram of a thin spherical shell collapsing on to a Schwarzschild black hole	163
7.11	*Spherical Rindler space	167
7.12	Rindler illusory horizon	170
7.13	Area of the Rindler horizon	171
7.14	*What use is a Lie derivative?	176
7.15	Equivalence of expressions for the Lie derivative	178
7.16	Lie derivative of the metric	179
7.17	Lie derivative of the inverse metric	180
8.1	*Units of charge of a charged black hole	182
8.2	Blueshift of a photon crossing the inner horizon of a Reissner-Nordström black hole	194
10.1	Isotropic (Poincaré) form of the FLRW metric	227
10.2	Omega in photons	233
10.3	Mass-energy in a FLRW Universe	235
10.4	*Mass of a ball of photons or of vacuum	235
10.5	Geodesics in the FLRW geometry	236
10.6	Age of a FLRW universe containing matter and vacuum	238
10.7	Age of a FLRW universe containing radiation and matter	238
10.8	Relation between conformal time and cosmic scale factor	239

10.9	Hubble diagram	243
10.10	Horizon size at recombination	247
10.11	The horizon problem	247
10.12	Relation between horizon and flatness problems	247
10.13	Distribution of non-interacting particles initially in thermodynamic equilibrium	258
10.14	The first law of thermodynamics with non-conserved particle number	259
10.15	Number, energy, pressure, and entropy of a relativistic ideal gas at zero chemical potential	260
10.16	A relation between thermodynamic integrals	260
10.17	Relativistic particles in the early Universe had approximately zero chemical potential	261
10.18	Entropy per particle	261
10.19	Photon temperature at high redshift versus today	262
10.20	Cosmic Neutrino Background	262
10.21	Maximally symmetric spaces	271
10.22	*Milne Universe	272
10.23	*Stationary FLRW metrics with different curvature constants describe the same spacetime	272
11.1	*Schwarzschild vierbein	280
11.2	Generators of Lorentz transformations are antisymmetric	281
11.3	Riemann tensor	290
11.4	Antisymmetry of the Riemann tensor	290
11.5	Cyclic symmetry of the Riemann tensor	290
11.6	Symmetry of the Riemann tensor	291
11.7	Number of components of the Riemann tensor	291
11.8	*Must connections vanish if Riemann vanishes?	291
11.9	Tidal forces falling into a Schwarzschild black hole	294
11.10	Totally antisymmetric tensor	295
13.1	Schur's lemma	312
13.2	*How fast do bivectors rotate?	318
13.3	Rotation of a vector	321
13.4	3D rotation matrices	324
13.5	*Properties of Pauli matrices	325
13.6	Translate a rotor into an element of $SU(2)$	326
13.7	Translate a Pauli spinor into a quaternion	327
13.8	Translate a quaternion into a Pauli spinor	327
13.9	Orthonormal eigenvectors of the spin operator	328
14.1	Null complex quaternions	331
14.2	Nilpotent complex quaternions	332
14.3	Lorentz boost	335
14.4	Factor a Lorentz rotor into a boost and a rotation	335
14.5	Topology of the group of Lorentz rotors	336
14.6	Interpolate a Lorentz transformation	340
14.7	Spline a Lorentz transformation	340
14.8	The wrong way to implement a Lorentz transformation	340
14.10	Translate a Dirac spinor into a complex quaternion	347

14.11	Translate a complex quaternion into a Dirac spinor	348
14.13	*The boost axis of a null spinor is Lorentz-invariant	351
14.14	*What makes Weyl spinors special?	352
15.1	*Commutator versus wedge product of multivectors	354
16.1	*Can the coordinate metric be Minkowski in the presence of torsion?	401
16.2	*Why the names energy-momentum and spin angular-momentum?	401
16.3	Energy-momentum and spin angular-momentum of the electromagnetic field	401
16.4	Energy-momentum and spin angular-momentum of a Dirac field	402
16.5	Energy-momentum and spin angular-momentum of a Majorana field	402
16.6	Electromagnetic field in the presence of torsion	402
16.7	Dirac (or Majorana) spinor field in the presence of torsion	403
16.8	Commutation of multivector forms	411
16.9	*Scalar product of the interval form $\mathbf{dx}$	414
16.10	Triple products involving products of the interval form $\mathbf{dx}$	414
16.11	Lie derivative of a form	424
16.12	Gravitational equations in arbitrary spacetime dimensions	437
16.13	Volume of a ball and area of a sphere	439
17.1	*Does Nature pick out a preferred foliation of time?	449
17.2	Energy and momentum constraints	462
17.3	Kasner spacetime	475
17.4	Schwarzschild interior as a Bianchi spacetime	475
17.5	Kasner spacetime for a perfect fluid	476
17.6	Oscillatory Belinskii-Khalatnikov-Lifshitz (BKL) instability	478
17.7	Geodesics in Bianchi spacetimes	480
18.1	Raychaudhuri equations for a non-geodesic timelike congruence	494
18.2	*How do singularity theorems apply to the Kerr geometry?	506
18.3	*How do singularity theorems apply to the Reissner-Nordström geometry?	507
19.1	Tetrad frame of a rotating wheel	510
19.2	Coordinate transformation from Schwarzschild to Gullstrand-Painlevé	513
19.3	Velocity of a person who free-falls radially from rest	513
19.4	Dragging of inertial frames around a Kerr-Newman black hole	519
19.5	River model of the Friedmann-Lemaître-Robertson-Walker metric	523
19.6	Program geodesics in a rotating black hole	524
20.1	Apparent horizon	527
20.2	Birkhoff's theorem	536
20.3	Naked singularities in spherical spacetimes	536
20.4	Constant density star	540
20.5	Oppenheimer-Snyder collapse	542
22.1	Explore other separable solutions	585
22.2	Explore separable solutions in an arbitrary number $N$ of spacetime dimensions	585
22.3	Area of the horizon	596
23.1	Near the Kerr-Newman singularity	603
23.2	When must $t$ and $\phi$ progress forwards on a geodesic?	604

23.3	Inside the sisytube	604
23.4	Gödel's Universe	604
23.5	Negative energy trajectories outside the horizon	605
23.6	*Are principal null geodesics circular orbits?	620
23.7	Icarus	624
23.8	Interstellar	625
23.9	Expansion, vorticity, and shear along the principal null congruences of the $\Lambda$ -Kerr-Newman geometry	636
24.1	*Which Einstein equations are redundant?	641
24.2	Can accretion fuel ingoing and outgoing streams at the inner horizon?	641
25.1	*Non-infinitesimal tetrad transformations in perturbation theory?	650
25.2	*Should not the Lie derivative of a tetrad tensor be a tetrad tensor?	651
25.3	*Variation of unperturbed quantities under coordinate gauge transformations?	652
26.1	*Are gauge-invariant potentials Lorentz-invariant?	663
26.2	*What parts of Maxwell's equations can be discarded?	665
26.3	Einstein tensor in harmonic gauge	667
26.4	*Independent evolution of scalar, vector, and tensor modes	669
26.5	Gravity Probe B and the geodetic and frame-dragging precession of gyroscopes	672
26.6	Equality of the two scalar potentials outside a spherical body	675
26.7	*Units of the gravitational quadrupole radiation formula	678
26.8	Hulse-Taylor binary	680
28.1	*Global curvature as a perturbation?	695
28.2	*Can the Universe at large rotate?	695
28.3	*Evolution of vector perturbations in FLRW spacetimes	701
28.4	Evolution of tensor perturbations (gravitational waves) in FLRW spacetimes	702
28.5	*Scalar, vector, tensor components of energy-momentum conservation	704
28.6	*What frame does the CMB define?	705
28.7	*Are congruences of comoving observers in cosmology hypersurface-orthogonal?	706
29.1	Entropy perturbation	710
29.2	*Entropy perturbation when number is conserved	711
29.3	Relation between entropy and $\zeta$	712
29.4	*If the Friedmann equations enforce conservation of entropy, where does the entropy of the Universe come from?	712
29.5	*What is meant by the horizon in cosmology?	715
29.6	Redshift of matter-radiation equality	715
29.7	Generic behaviour of dark matter	716
29.8	Generic behaviour of radiation	717
29.9	*Can neutrinos be treated as a fluid?	717
29.10	Program the equations for the simplest set of cosmological assumptions	718
29.11	Radiation-dominated fluctuations	729
29.12	*Does the radiation monopole oscillate after recombination?	736
29.13	Growth of baryon fluctuations after recombination	738
29.14	*Curvature scale	741

29.15	Power spectrum of matter fluctuations: simple approximation	745
30.1	Proton and neutron fractions	755
30.2	*Level populations of hydrogen near recombination	760
30.3	*Ionization state of hydrogen near recombination	760
30.4	*Atomic structure notation	761
30.5	*Dimensional analysis of the collision rate	764
30.6	Detailed balance	765
30.7	Recombination	772
31.1	Thomson scattering rate	780
31.2	Program the equations in the hydrodynamic approximation	782
31.3	Power spectrum of matter fluctuations: hydrodynamic approximation	782
31.4	Effect of massive neutrinos on the matter power spectrum	783
31.5	Behaviour of radiation in the presence of damping	795
31.6	Diffusion scale	795
31.7	Generic behaviour of neutrinos	796
32.1	Program the Boltzmann equations	801
32.2	Power spectrum of matter fluctuations: Boltzmann treatment	801
32.3	Boltzmann equation factors in a general gauge	805
32.4	Moments of the non-baryonic cold dark matter Boltzmann equation	806
32.5	Initial conditions in the presence of neutrinos	817
33.1	CMB power spectrum in the instantaneous and rapid recombination approximation	835
33.2	CMB power spectra from CAMB	837
33.3	Cosmic Neutrino Background	842
34.1	*Elliptically polarized light	847
34.2	*Fluctuations with $ m  \geq 3$ ?	853
34.3	Photon diffusion including polarization	857
34.4	Boltzmann code including polarization	858
34.5	*Scalar, vector, tensor power spectra?	865
34.6	CMB polarized power spectrum	866
36.1	*Boost and spin weight	887
36.2	*Lorentz transformation of the phase of a spinor	890
36.3	*Chiral scalar	896
38.1	Pseudoscalar electromagnetic Lagrangian?	933
38.2	Chiral electromagnetic Lagrangian?	934



---

## Legal notice

If you have obtained a copy of this proto-book from anywhere other than my website

<http://jila.colorado.edu/~ajsh/>

then that copy is illegal. This version of the proto-book is linked at

[http://jila.colorado.edu/~ajsh/astr5770\\_14/notes.html](http://jila.colorado.edu/~ajsh/astr5770_14/notes.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 2014-15

---

## 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.  $\partial_m$  denotes tetrad directed derivatives, which do not commute.  $D_{\mu}$  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  $\varepsilon^a_{bc}$  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  $c = \text{universal constant}$  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  $\alpha$  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  $\alpha$  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.

# 1

---

## 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 <http://casa.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  $2\pi$ , 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!

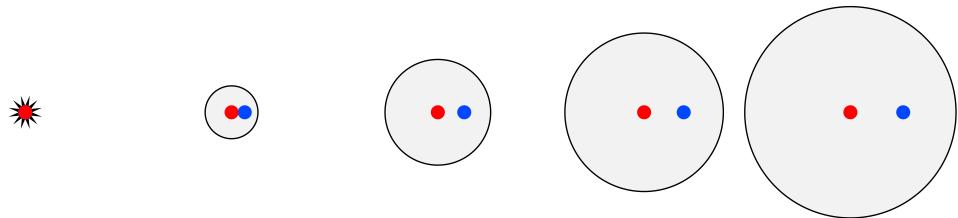


Figure 1.1 Vermilion emits a flash of light, which (from left to right) expands away from her in all directions. Since the speed of light is constant in all directions, she finds herself at the centre of the expanding sphere of light. Cerulean is moving to the right at half of the speed of light relative 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

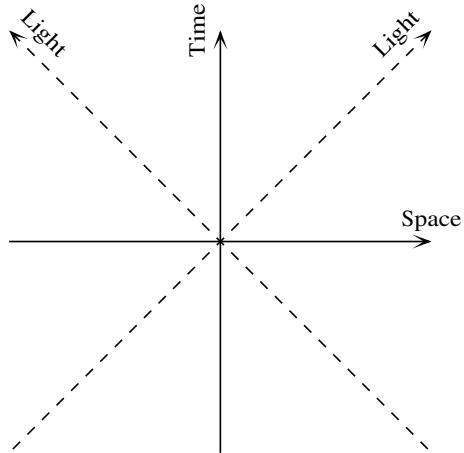


Figure 1.2 A spacetime diagram shows events in space and time. In a spacetime diagram, time goes upward, while space dimensions are horizontal. Really there should be 3 space dimensions, but usually it suffices to show 1 spatial dimension, as here. In a spacetime diagram, the units of space and time 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

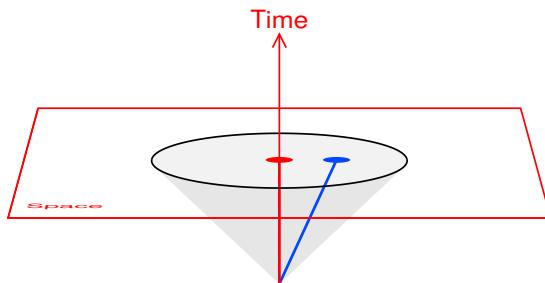


Figure 1.3 Spacetime diagram of Vermilion emitting a flash of light. This is a spacetime diagram version of the situation illustrated in Figure 1.1. The lines along which Vermilion and Cerulean move through spacetime are called their worldlines. Each point in 4-dimensional spacetime is called an event. 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 in 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

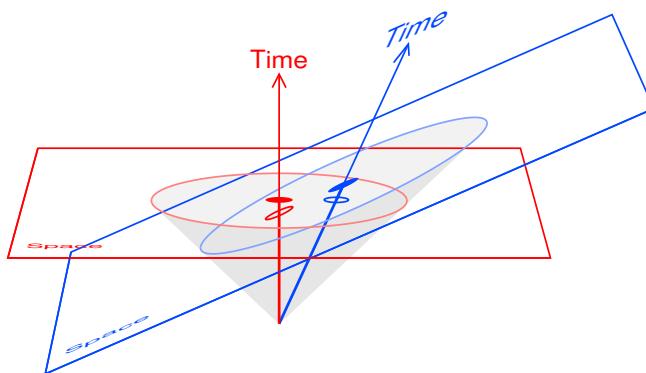


Figure 1.4 The solution to how both Vermilion and Cerulean can consider themselves to be at the centre of the lightcone. Cerulean's spacetime is skewed compared to Vermilion's. Cerulean is in the centre of the lightcone, according to the way Cerulean perceives space and time, while Vermilion remains 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.

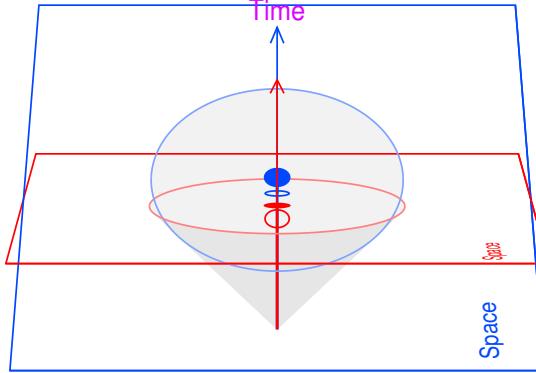


Figure 1.5 Same as Figure 1.4, but with Cerulean moving into the page instead of to the right. This is just Figure 1.4 spatially rotated by  $90^\circ$  in the horizontal plane. Distances perpendicular to the direction of motion are unchanged.

## 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

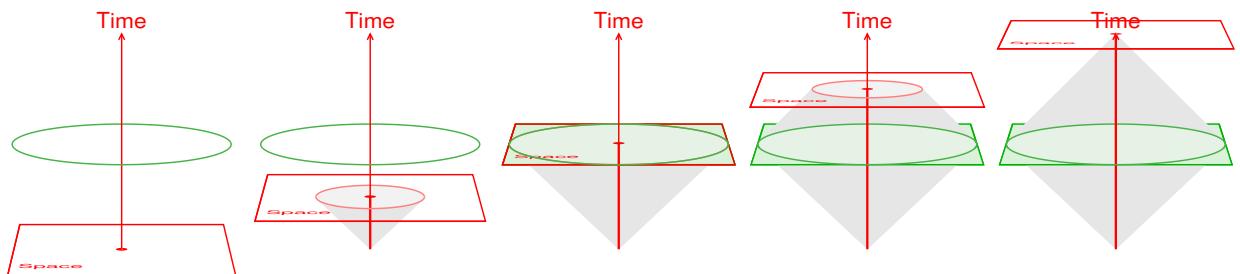


Figure 1.6 How Vermilion defines hypersurfaces of simultaneity. She surrounds herself with (green) mirrors all at the same distance. She sends out a light beam, which reflects off the mirrors, and returns to her all at the same moment. She knows that the mirrors are all at the same distance precisely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

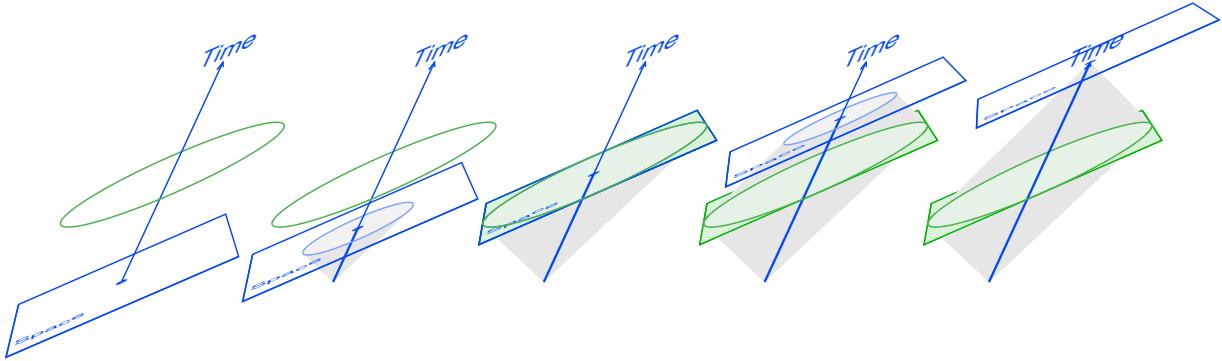


Figure 1.7 Cerulean defines hypersurfaces of simultaneity using the same operational setup as Vermilion: he bounces light off (green) mirrors all at the same distance from him, arranging them so that the light returns to him all at the same time. But from Vermilion's frame, Cerulean's experiment looks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

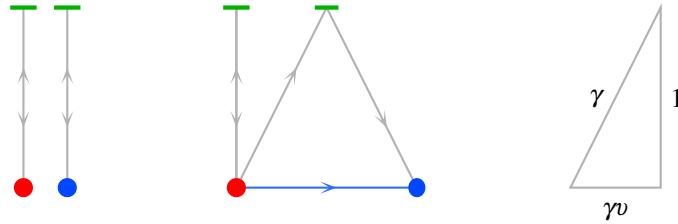


Figure 1.8 Vermilion and Cerulean construct identical clocks, consisting of a light beam that bounces off a (green) mirror and returns to them. In the left panel, Cerulean is at rest relative to Vermilion. They both agree that their clocks are identical. In the middle panel, Cerulean is moving to the 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  $\gamma$ , 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  $\gamma$ , than her own. The sides of the triangle formed by the distance 1 to the mirror, the length  $\gamma$  of the lightpath to Cerulean's clock, and the distance  $\gamma v$  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**  $\gamma$ , 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  $\gamma$  ticks. Cerulean is moving at speed  $v$ , so Vermilion thinks he moves a distance of  $\gamma v$  ticks during the  $\gamma$  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,  $\gamma$ , and  $\gamma v$ , 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  $\gamma$  is related to Cerulean's velocity  $v$  by

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

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.

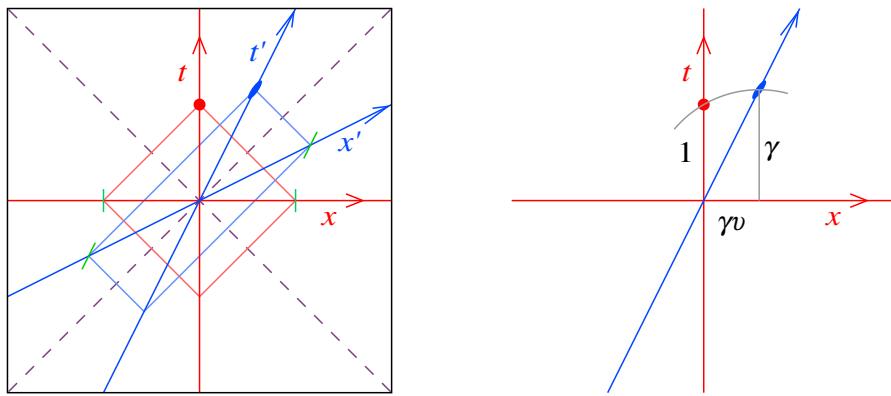


Figure 1.9 Spacetime diagram representing the experiments shown in Figures 1.6 and 1.7. The right panel shows a detail of how the spacetime diagram can be drawn using only a straight edge and a compass. If Cerulean's position is drawn first, then Vermilion's position follows from drawing the arc as shown.

**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 lightrays 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  $\gamma$  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

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

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  $\tau$  experienced by a moving person is a factor  $\gamma$  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

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

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

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

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

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

which says that the observer measures the front end of the rocket to be a distance  $l/\gamma$  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  $\gamma$ .

**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.

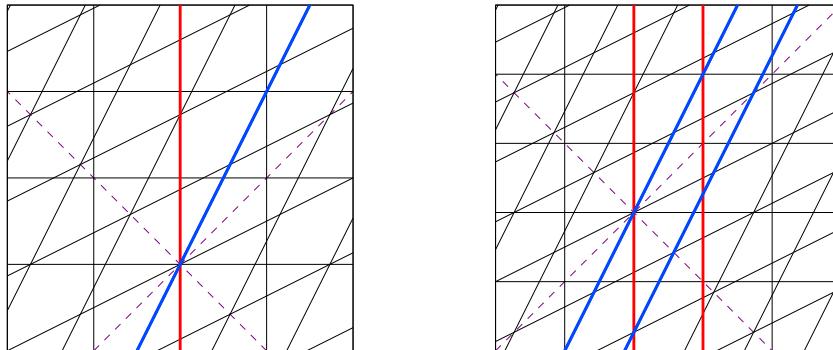


Figure 1.10 (Left) Time dilation, and (right) Lorentz contraction spacetime diagrams.

**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?

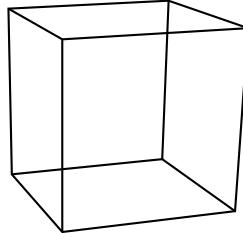


Figure 1.11 A cube. Are the lengths of its sides all equal?

**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" when just arriving at Alpha, and the twin's "now" 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  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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

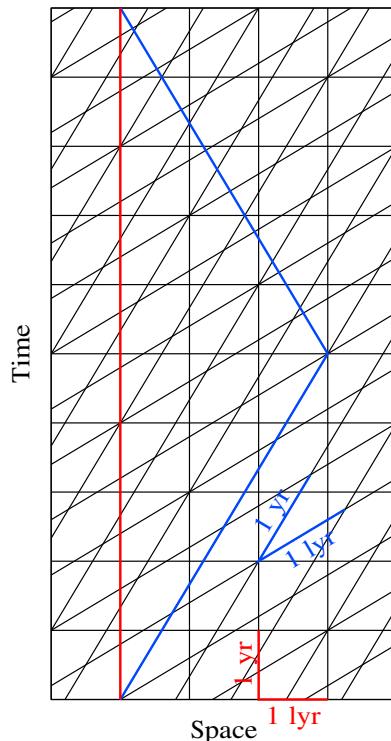


Figure 1.12 Twin paradox spacetime diagram.

wristwatch during this time? What is the ratio of these two times? Express your answers first in terms of  $\ell$ ,  $v$ , and  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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  $\ell$ ,  $v$ , and  $\gamma$ , 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

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

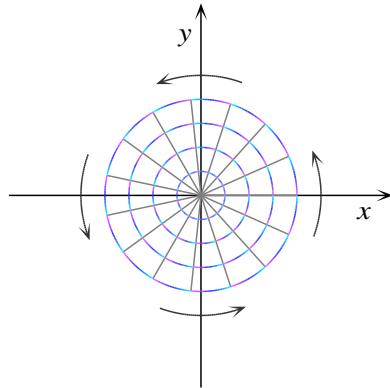


Figure 1.13 A wheel.

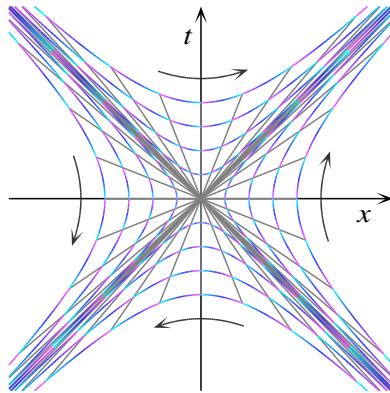


Figure 1.14 A spacetime wheel.

### 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  $\alpha$  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  $\alpha$  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  $\alpha$  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  $\alpha$ , 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

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

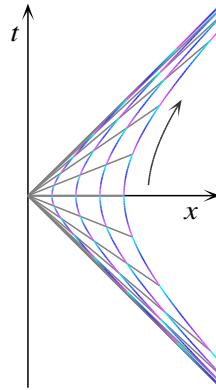


Figure 1.15 The right quadrant of the spacetime wheel represents the worldlines and lines of simultaneity of persons who accelerate in the  $x$  direction with uniform acceleration in their own frames.

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

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

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  $0.97g = 9.5 \text{ m s}^{-2}$ .

Table 1.1 *Trip across the Universe.*

Time elapsed on spaceship in years	Time elapsed on Earth in years	Distance travelled in lightyears	To
$\alpha$	$\sinh \alpha$	$\cosh \alpha - 1$	
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	$2.7 \times 10^4$	$2.7 \times 10^4$	Centre of Milky Way
15.4	$2.44 \times 10^6$	$2.44 \times 10^6$	Andromeda galaxy
18.4	$4.9 \times 10^7$	$4.9 \times 10^7$	Virgo cluster
19.2	$1.1 \times 10^8$	$1.1 \times 10^8$	Coma cluster
25.3	$5 \times 10^{10}$	$5 \times 10^{10}$	Edge of observable Universe

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.

## 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  $\Delta s$  between two events separated by  $\{\Delta t, \Delta x, \Delta y, \Delta z\}$  is given by

$$\begin{aligned}\Delta s^2 &= -\Delta t^2 + \Delta r^2 \\ &= -\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2.\end{aligned}\tag{1.26}$$

A quantity such as  $\Delta s^2$  that remains unchanged under any Lorentz transformation is called a **scalar**. You should check yourself that  $\Delta s^2$  is unchanged under Lorentz transformations (see Exercise 1.13). Lorentz transformations can be defined as linear spacetime transformations that leave  $\Delta s^2$  invariant.

The single scalar spacetime squared interval  $\Delta s^2$  replaces the two scalar quantities

$$\begin{array}{ll}\text{time interval} & \Delta t \\ \text{spatial interval} & \Delta r = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}\end{array}\tag{1.27}$$

of classical Galilean spacetime.

### 1.9.1 Timelike, lightlike, spacelike

The scalar spacetime distance squared  $\Delta s^2$ , equation (1.26), between two events can be negative, zero, or positive. A spacetime interval  $\{\Delta t, \Delta x, \Delta y, \Delta z\} \equiv \{\Delta t, \Delta r\}$  is called

$$\begin{array}{ll} \text{timelike} & \text{if } \Delta t > \Delta r \text{ or equivalently if } \Delta s^2 < 0 , \\ \text{null or lightlike} & \text{if } \Delta t = \Delta r \text{ or equivalently if } \Delta s^2 = 0 , \\ \text{spacelike} & \text{if } \Delta t < \Delta r \text{ or equivalently if } \Delta s^2 > 0 , \end{array} \quad (1.28)$$

as illustrated in Figure 1.16.

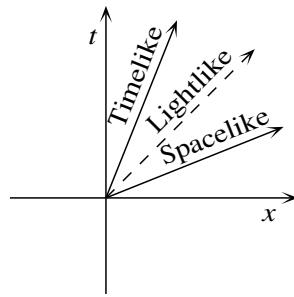


Figure 1.16 Spacetime diagram illustrating timelike, lightlike, and spacelike intervals.

### 1.9.2 Proper time, proper distance

The scalar spacetime distance squared  $\Delta s^2$  has a physical meaning.

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

$$\Delta\tau = \sqrt{-\Delta s^2} = \sqrt{\Delta t^2 - \Delta r^2} . \quad (1.29)$$

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

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

$$\Delta l = \sqrt{\Delta s^2} = \sqrt{\Delta r^2 - \Delta t^2} . \quad (1.30)$$

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

**Concept question 1.12 Proper time, proper distance.** Justify the assertions (1.29) and (1.30).

### 1.9.3 Minkowski metric

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

$$\Delta x^m \equiv \{\Delta t, \Delta \mathbf{r}\} \equiv \{\Delta t, \Delta x, \Delta y, \Delta z\} . \quad (1.31)$$

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

$$\Delta s^2 = \eta_{mn} \Delta x^m \Delta x^n , \quad (1.32)$$

where  $\eta_{mn}$  is the **Minkowski metric**

$$\eta_{mn} \equiv \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad (1.33)$$

Equation (1.32) 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  $\Delta x^m$  transforms as

$$\Delta x^m \rightarrow \Delta x'^m = L^m{}_n \Delta x^n , \quad (1.34)$$

where  $L^m{}_n$  denotes a Lorentz transformation. The paired indices  $n$  on the right hand side of equation (1.36), one lowered and one raised, are implicitly summed over. In matrix notation,  $L^m{}_n$  is a  $4 \times 4$  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

$$a^m \equiv \{a^t, a^x, a^y, a^z\} , \quad (1.35)$$

that transforms under Lorentz transformations like an interval  $\Delta x^m$  of spacetime,

$$a^m \rightarrow a'^m = L^m{}_n a^n . \quad (1.36)$$

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,

$$a_m \equiv \eta_{mn} a^n . \quad (1.37)$$

With the Minkowski metric (1.33), the covariant components  $a_m$  are

$$a_m = \{-a^t, a^x, a^y, a^z\} , \quad (1.38)$$

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

$$\begin{aligned} a^m a_m &\equiv \eta_{mn} a^m a^n \\ &= a_t a^t + a_x a^x + a_y a^y + a_z a^z \\ &= -(a^t)^2 + (a^x)^2 + (a^y)^2 + (a^z)^2 \end{aligned} \quad (1.39)$$

is a Lorentz scalar, a fact you will prove in Exercise 1.13.

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 1.13 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 1.14 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  $\alpha$  rather than 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

$$\tau^2 - (\tau_1 + \tau_2)^2 = 4\tau_1 \tau_2 \sinh^2 \left( \frac{\alpha_2 - \alpha_1}{2} \right) , \quad (1.40)$$

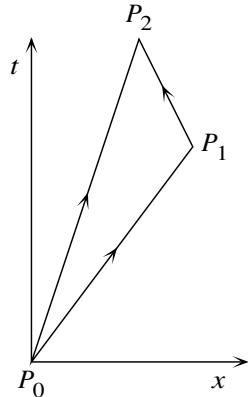


Figure 1.17 The longest proper time between  $P_0$  and  $P_2$  is a straight line.

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  $\Delta x^m$  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:

$$\left. \begin{array}{l} \text{energy} = \text{time component} \\ \text{momentum} = \text{space component} \end{array} \right\} \text{of a 4-vector.} \quad (1.41)$$

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

$$\frac{\text{energy}}{\text{momentum}} = \text{constant} \quad (1.42)$$

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  $\mathbf{v}$ ,

$$\mathbf{p} = m\mathbf{v} = m \frac{d\mathbf{r}}{dt} . \quad (1.43)$$

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  $p^n$ :

$$\begin{aligned} p^n &= m \frac{dx^n}{d\tau} \\ &= m \left\{ \frac{dt}{d\tau}, \frac{d\mathbf{r}}{d\tau} \right\} \\ &= m \{\gamma, \gamma\mathbf{v}\} . \end{aligned} \quad (1.44)$$

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

$$p^n = \{E, \mathbf{p}\} = \{m\gamma, m\gamma\mathbf{v}\} . \quad (1.45)$$

### 1.11.2 Special relativistic energy

From equation (1.45), the special relativistic energy  $E$  is the product of the rest mass and the Lorentz  $\gamma$ -factor,

$$E = m\gamma \quad (\text{units } c = 1) , \quad (1.46)$$

or, restoring standard units,

$$E = mc^2\gamma . \quad (1.47)$$

For small velocities  $v$ , the Taylor expansion of the Lorentz factor  $\gamma$  is

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = 1 + \frac{1}{2} \frac{v^2}{c^2} + \dots . \quad (1.48)$$

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

$$\begin{aligned} E &= mc^2 \left( 1 + \frac{1}{2} \frac{v^2}{c^2} + \dots \right) \\ &= mc^2 + \frac{1}{2} mv^2 + \dots . \end{aligned} \quad (1.49)$$

The first term,  $mc^2$ , is the rest-mass energy. The second term,  $\frac{1}{2}mv^2$ , is the non-relativistic 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

$$E = mc^2 , \quad (1.50)$$

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

$$\begin{aligned} p_n p^n &= -E^2 + p^2 \\ &= -m^2(\gamma^2 - \gamma^2 v^2) \\ &= -m^2 , \end{aligned} \quad (1.51)$$

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 (1.50) 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

$$m = 0 . \quad (1.52)$$

Its scalar energy-momentum squared is thus zero,

$$p_n p^n = -E^2 + p^2 = -m^2 = 0 . \quad (1.53)$$

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

$$p \equiv |\mathbf{p}| = E . \quad (1.54)$$

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

$$\begin{aligned} p^n &= \{E, \mathbf{p}\} \\ &= E\{1, \mathbf{n}\} \\ &= h\nu\{1, \mathbf{n}\} \end{aligned} \quad (1.55)$$

where  $\nu$  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

$$\begin{pmatrix} p'^t \\ p'^x \\ p'^y \\ p'^z \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p^t \\ p^x \\ p^y \\ p^z \end{pmatrix} = \begin{pmatrix} \gamma(p^t - vp^x) \\ \gamma(p^x - vp^t) \\ p^y \\ p^z \end{pmatrix} . \quad (1.56)$$

Equivalently

$$h\nu' \begin{pmatrix} 1 \\ n'^x \\ n'^y \\ n'^z \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} h\nu \begin{pmatrix} 1 \\ n^x \\ n^y \\ n^z \end{pmatrix} = h\nu \begin{pmatrix} \gamma(1 - n^x v) \\ \gamma(n^x - v) \\ n^y \\ n^z \end{pmatrix} . \quad (1.57)$$

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

### 1.12.2 Redshift

The wavelength  $\lambda$  of a photon is related to its frequency  $\nu$  by

$$\lambda = c/\nu . \quad (1.58)$$

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

$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} . \quad (1.59)$$

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

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} . \quad (1.60)$$

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

$$\frac{1}{1 + z} \equiv \frac{\lambda_{\text{em}}}{\lambda_{\text{obs}}} = \frac{\nu_{\text{obs}}}{\nu_{\text{em}}} . \quad (1.61)$$

### 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 (1.57),

$$h\nu_{\text{em}} = h\nu_{\text{obs}}\gamma(1 - n^x v) . \quad (1.62)$$

The redshift factor is therefore

$$\begin{aligned} 1 + z &= \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} \\ &= \gamma(1 - n^x v) \\ &= \gamma(1 - \mathbf{n} \cdot \mathbf{v}) . \end{aligned} \quad (1.63)$$

Equation (1.63) is the general formula for the special relativistic Doppler shift. In special cases,

$$1 + z = \begin{cases} \sqrt{\frac{1-v}{1+v}} & \text{velocity directly towards observer } (\mathbf{v} \text{ aligned with } \mathbf{n}), \\ \gamma & \text{velocity in the transverse direction } (\mathbf{v} \cdot \mathbf{n} = 0), \\ \sqrt{\frac{1+v}{1-v}} & \text{velocity directly away from observer } (\mathbf{v} \text{ anti-aligned with } \mathbf{n}) . \end{cases} \quad (1.64)$$

## 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 1.15.

**Exercise 1.15 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  $x-y$ -plane. Let the jet be moving at velocity  $v$  at angle  $\theta$  away from the direction towards us on Earth, so that its spatial velocity relative to Earth is  $\mathbf{v} \equiv \{v_x, v_y\} = \{-v \cos \theta, v \sin \theta\}$ .
2. In Earth coordinates  $\{t, x, y\}$ , the jet moves in time  $t$  a distance  $\mathbf{l} = \{l_x, l_y\} = \mathbf{vt}$ . 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

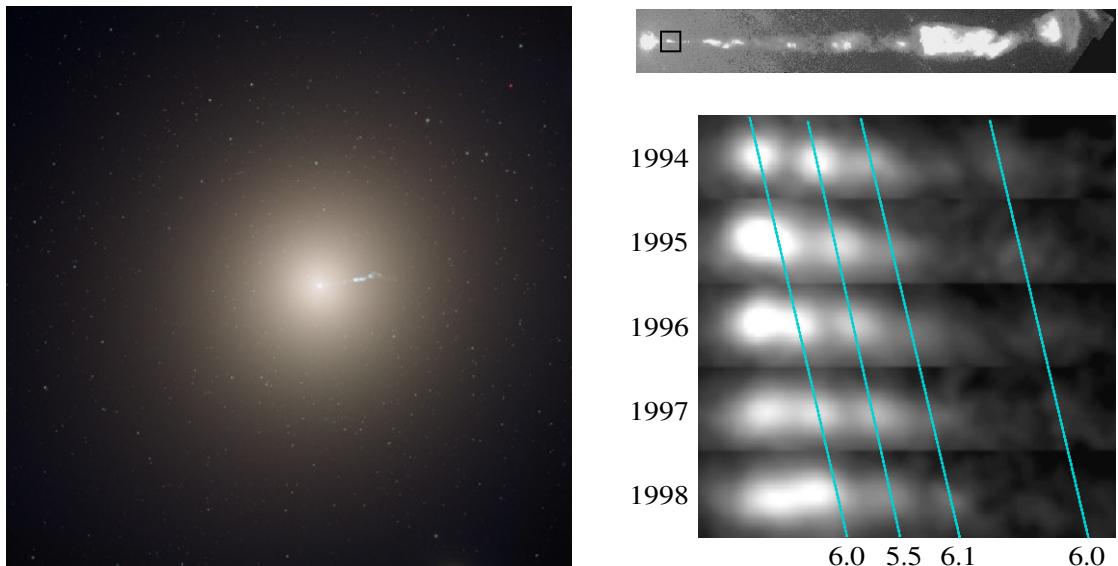


Figure 1.18 The left panel shows an image of the galaxy M87 taken with the Advanced Camera for Surveys on the Hubble Space Telescope. A jet, bluish compared to the starry background of the galaxy, emerges from the galaxy's central nucleus. Radio observations, not shown here, reveal that there is a second 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  $6c$ . 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.

velocity on the sky is

$$v_{\text{app}} = \frac{v \sin \theta}{1 - v \cos \theta}. \quad (1.65)$$

3. Sketch the apparent velocity  $v_{\text{app}}$  as a function of  $\theta$  for some given velocity  $v$ . In terms of  $v$  and the Lorentz factor  $\gamma$ , what are the values of  $\theta$  and of  $v_{\text{app}}$  at the point where  $v_{\text{app}}$  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  $\theta$ ? By expressing  $v$  in terms of  $v_{\text{app}}$  and  $\theta$  using equation (1.65), show that the blueshift factor is

$$\frac{1}{1+z} = \sqrt{1 + 2v_{\text{app}} \cot \theta - v_{\text{app}}^2}. \quad (1.66)$$

[Hint: Remember to use the correct redshift formula, equation (1.63).]

5. In terms of  $v_{\text{app}}$ , at what value of  $\theta$  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  $\theta$ ? 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. 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,

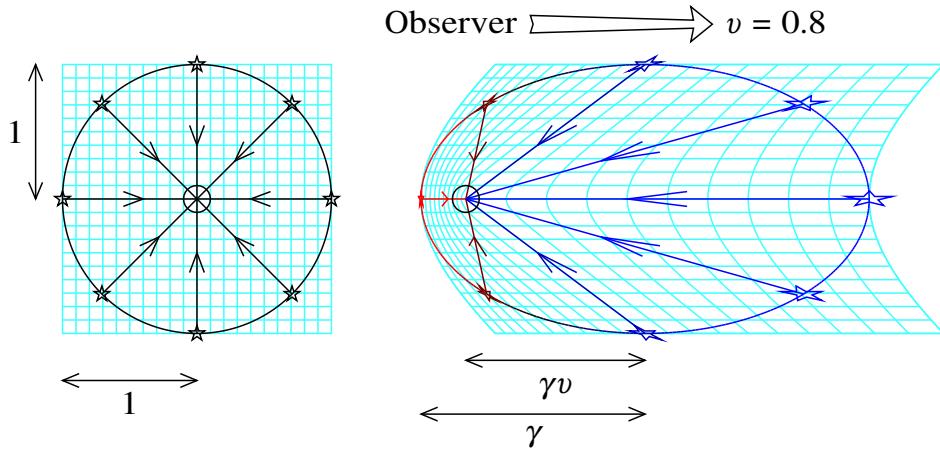


Figure 1.19 The rules of 4-dimensional perspective. In special relativity, the scene seen by an observer moving through the scene (right) is relativistically beamed compared to the scene seen by an observer at rest relative to the scene (left). On the left, the observer at the center of the circle is 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.

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 1.16 The rules of 4-dimensional perspective.

- 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  $\theta$  to the  $x$ -direction and in the primed frame at angle  $\theta'$  to the  $x$ -direction. Show that  $\mu' \equiv \cos \theta'$  and  $\mu \equiv \cos \theta$  are related by

$$\mu' = \frac{\mu + v}{1 + v\mu} . \quad (1.67)$$

- 3. Redshift.** By what factor  $a = E'/E$  is the observed photon frequency from the object changed? Express your answer as a function of  $\gamma$ ,  $v$ , and  $\mu$ .
- 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),

$$F(E) \equiv \frac{dN(E)}{dt d\Omega d\ln E} , \quad (1.68)$$

goes as

$$\frac{F'(E')}{F(E)} = \frac{E'}{E} \frac{d\mu}{d\mu'} = a^3 . \quad (1.69)$$

[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 1.17 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  $\theta$  be the angle between the  $x$ -direction and the direction  $\mathbf{m}$  to the center of the circle, and let  $\alpha$  be the angle between the circle axis  $\mathbf{m}$  and the photon direction  $\mathbf{n}$ . Show that the angle  $\theta'$  in the primed frame is given by

$$\tan \theta' = \frac{\sin \theta}{\gamma(v \cos \alpha + \cos \theta)} , \quad (1.70)$$

and that the angular radius  $\alpha'$  in the primed frame is given by

$$\tan \alpha' = \frac{\sin \alpha}{\gamma(\cos \alpha + v \cos \theta)} . \quad (1.71)$$

[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 1.18 Lorentz transformation preserves angles on the sky.** From equation (1.67), show that the angular metric  $do^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$  on the sky Lorentz transforms as

$$do'^2 = \frac{1 - v^2}{(1 + v \cos \theta)^2} do^2 . \quad (1.72)$$

This kind of transformation, which multiplies the metric by an overall factor, called a conformal factor, is called a **conformal transformation**. The conformal transformation (1.72) of the angular metric shrinks and expands patches on the sky while preserving their shapes, that is, while preserving angles between lines.

**Exercise 1.19 The aberration of starlight.** The aberration of starlight was discovered by James Bradley (1728) through precision measurements of the position of  $\gamma$  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 1.20 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 1.16 asked you to discover how the appearance of an emitter changes when the observer boosts into a different frame. The change (1.69) 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  $\hbar^3$ ; 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,

$$dN(t, \mathbf{r}, \mathbf{p}) = f(t, \mathbf{r}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3}. \quad (1.73)$$

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**  $I_\nu$  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  $\nu$ . Often intensity is quoted per unit wavelength  $\lambda$  or per unit energy  $E$  instead of per unit frequency  $\nu$ , and the intensity is subscripted accordingly,  $I_\lambda$  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

$$I_E \equiv \frac{E dN}{dt dA do dE} = cf \frac{gp^3}{(2\pi\hbar)^3}, \quad (1.74)$$

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 (1.74) reproduces the brightness factor  $a^3 \equiv (E'/E)^3$  in equation (1.69).

Stars typically appear to astronomers as point sources. Astronomers define the **flux**  $F_\nu$  from a source to be the intensity  $I_\nu$  integrated over the solid angle of the source. Again, flux is often quoted per unit wavelength  $\lambda$  or per unit energy  $E$ , and subscripted accordingly,  $F_\lambda$  or  $F_E$ .

**Concept question 1.21 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

$$F_E \equiv \frac{E dN}{dt dA dE} = c \frac{gp^3}{(2\pi\hbar)^3} \int f do. \quad (1.75)$$

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

$$F_E \propto E, \quad (1.76)$$

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,

$$\int F_E d \ln E \propto E . \quad (1.77)$$

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

$$\int F_E dE \propto E^2 . \quad (1.78)$$


---

## 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. Thus 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, as will be demonstrated in Chapter 13.

---

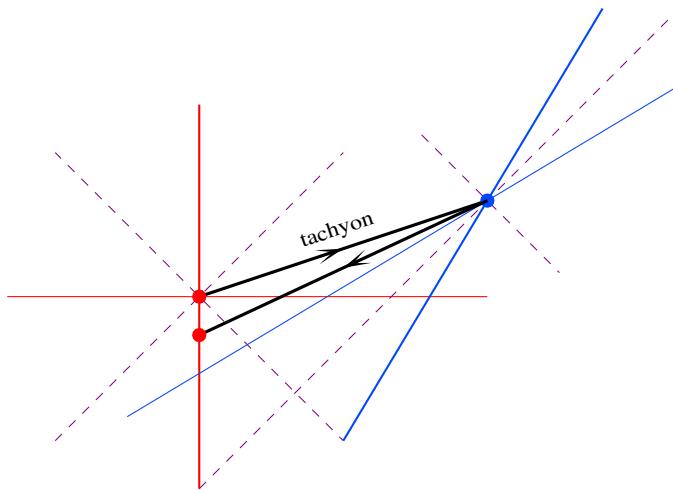


Figure 1.20 Tachyon spacetime diagram.

**Exercise 1.22 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  $x^\mu$ ; (b) a coordinate interval  $dx^\mu$ ; (c) proper time  $\tau$ ?
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.

## 2

---

### 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  $e_\mu$  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  $e_\mu$  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 1.22. 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

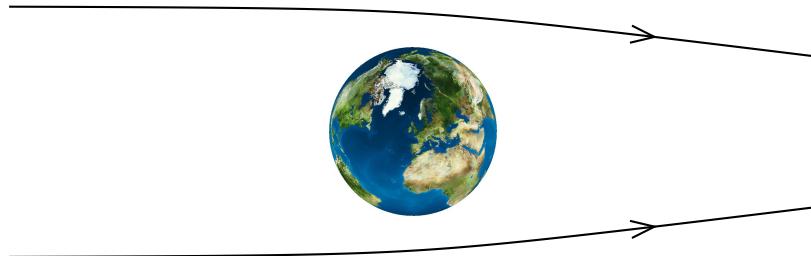


Figure 2.1 Particles initially on parallel trajectories passing either side of the Earth are caused to converge by the Earth's gravity. According to Einstein's principle of equivalence, the situation is equivalent to one where the particles are moving in straight lines in local free-fall frames. This 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  $x^\mu = \{\theta, \phi\}$

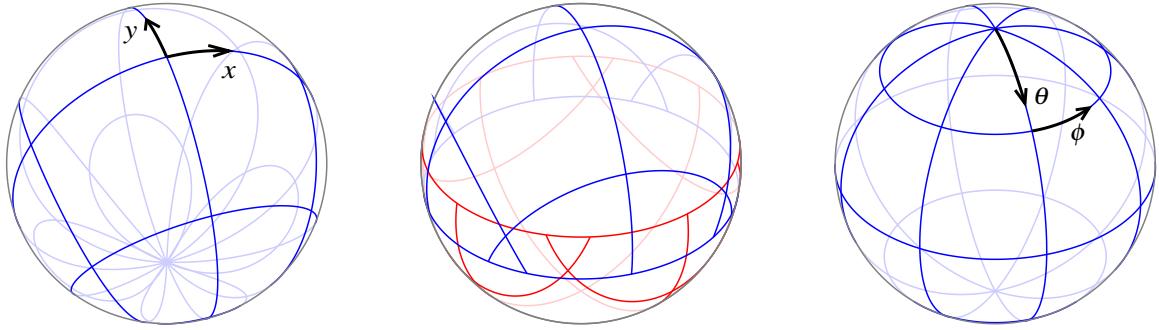


Figure 2.2 The 2-sphere is a 2-manifold, a topological space that is locally homeomorphic to Euclidean 2-space  $\mathbb{R}^2$ . Any attempt to cover the surface of a 2-sphere with a single chart, that is, with coordinates  $x$  and  $y$  such that each point on the sphere is specified by a unique coordinate  $\{x, y\}$ , 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 differentiably. The right panel shows standard polar coordinates  $\theta, \phi$  on the 2-sphere. The polar coordinatization fails at the north and south poles, where lines of longitude cross, the azimuthal angle  $\phi$  is not unique, and a person passing smoothly through the pole would see the azimuthal angle jump by  $\pi$ . 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  $\pi$ . 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 zero. In other words, Einstein’s principle of equivalence asserts the **existence of locally inertial frames**.

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  $4 \times 4$  symmetric matrix of equations

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

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  $\Phi$  is the Newtonian gravitational potential, and  $\rho$  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, §2.27.

## 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 2.7.

---

**Exercise 2.1 The equivalence principle implies the gravitational redshift of light, Part 1.** A rigorous general relativistic version of this exercise is Exercise 2.8. 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 potential  $\Phi$  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)

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

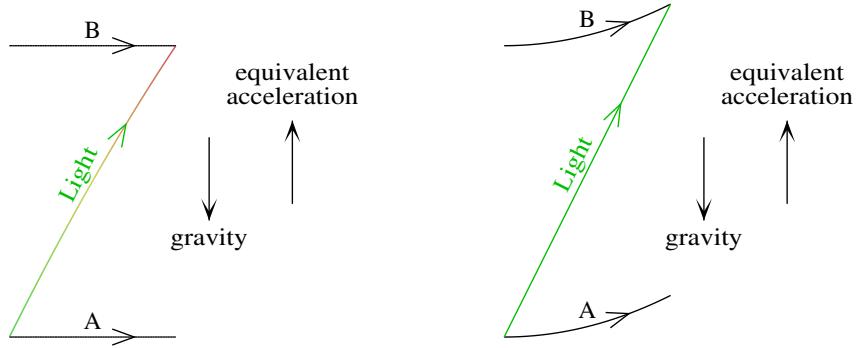


Figure 2.3 Einstein's principle of equivalence implies the gravitational redshift of light, and the gravitational bending of light. In the left panel, persons A and B are at rest relative to each other in a uniform gravitational field. They are shown moving to the right to bring out the evolution of 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.

**Exercise 2.2 The equivalence principle implies the gravitational redshift of light, Part 2.** A rigorous general relativistic version of this exercise is Exercise 2.9. 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  $\gamma$ -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  $\omega$ , 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 non-relativistic 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 2.13.

## 2.4 Metric

Postulate (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) 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  $\eta_{mn}$  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 (1.26) for the spacetime distance  $\Delta s^2$  held for finite intervals  $\Delta x^m$ , 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 differentiably related, so locally inertial intervals  $d\xi^m$  and coordinate intervals  $dx^\mu$  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  $dx^\mu$  as

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

where  $g_{\mu\nu}$  is the **metric**, a  $4 \times 4$  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  $dx^\mu$  to a proper measurement of an interval of time or space.

## 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  $dx^\mu$ . The scalar spacetime distance squared can be negative, zero, or positive, and accordingly timelike, lightlike, or spacelike:

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

## 2.6 Orthonormal tetrad basis $\gamma_m$

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  $\xi^m$  be a set of locally inertial coordinates at a point of spacetime. Define the vectors  $\gamma_m$ , 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  $\gamma_m$  is a 4-dimensional object, with both magnitude and direction. The basis vectors  $\gamma_m$  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

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

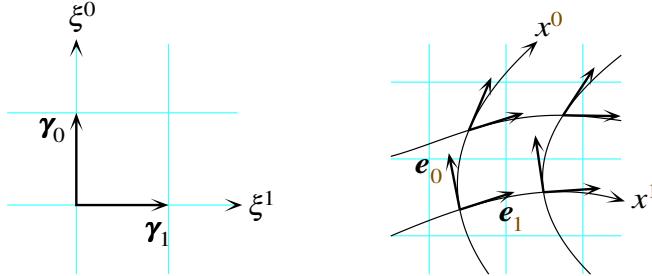


Figure 2.4 (Left) The tetrad vectors  $\gamma_m$  form an orthonormal basis of vectors tangent to a set of locally inertial coordinates  $\xi^m$  at a point. (Right) The coordinate tangent vectors  $e_\mu$  are the basis of vectors tangent to the coordinates at each point. The background square grid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

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  $dx$  invariant under Lorentz transformation, the basis vectors  $\gamma_m$  must transform as a covariant Lorentz vector.

The scalar length squared of the abstract vector interval  $dx$  is

$$ds^2 = dx \cdot dx = \gamma_m \cdot \gamma_n d\xi^m d\xi^n . \quad (2.20)$$

Since this must reproduce the locally inertial metric (2.10), the scalar products of the tetrad vectors  $\gamma_m$  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 $e_\mu$

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

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

are defined to satisfy

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

The relation (2.11) between coordinate intervals  $dx^\mu$  and locally inertial coordinate intervals  $d\xi^m$  implies that the coordinate tangent vectors  $e_\mu$  must be related to the orthonormal tetrad vectors  $\gamma_m$  by

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

Like the tetrad axes  $\gamma_m$ , each coordinate tangent axis  $e_\mu$  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  $dx$  is

$$ds^2 = dx \cdot dx = e_\mu \cdot e_\nu dx^\mu dx^\nu , \quad (2.25)$$

from which it follows that the scalar products of the coordinate tangent axes  $e_\mu$  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  $\gamma_m$ , the coordinate tangent vectors  $e_\mu$  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). Third, vectors in the tangent space can be differentiated with respect to coordinates, as will be elucidated in §2.9.3.

Some texts represent the tangent vectors  $e_\mu$  with the notation  $\partial_\mu$ , on the grounds that  $e_\mu$  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  $\partial_\mu$  as a derivative and  $\partial_\mu$  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  $dx^\mu$  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

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

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  $A^{\mu}$  with raised index are called its **contravariant** components, while those  $A_{\mu}$  with lowered indices are called its **covariant** components,

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

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

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

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

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

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

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

The dot products of the dual and tangent basis elements are

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

### 2.8.5 Covariant coordinate 4-vector

Under a general coordinate transformation (2.27), the covariant components  $A_{\mu}$  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  $A_{\mu}$  is consistent with the transformation law (2.29) for the contravariant components  $A^{\mu}$ .

You can check that the tangent vectors  $e_{\mu}$  transform as a covariant coordinate 4-vector.

### 2.8.6 Scalar product

If  $A^\mu$  and  $B^\mu$  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  $\mathbf{B} = e_\nu B^\nu$  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  $A^\mu$  (with a dummy index  $\mu$ ) and  $\mathbf{A} \equiv A^\mu e_\mu$  vectors. Although  $A^\mu$  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  $A^\mu$ .

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

1. A **coordinate vector**  $A^\mu$ , identified by greek (brown) indices  $\mu$ , 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  $A^\mu$  and  $\mathbf{A}$  as vectors, this book follows the standard physics practice of mentally regarding  $A^\mu$  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  $\binom{n}{m}$ .

## 2.9 Covariant derivatives

### 2.9.1 Derivative of a coordinate scalar

Suppose that  $\Phi$  is a coordinate scalar. Then the coordinate derivative of  $\Phi$  is a coordinate 4-vector

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

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 covariant coordinate 4-vector  $A^{\mu}$  is not a tensor

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

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

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

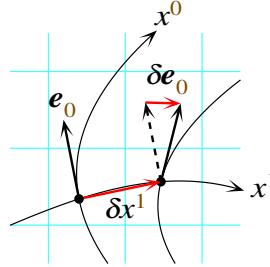


Figure 2.5 The change  $\delta e_0$  in the tangent vector  $e_0$  over a small interval  $\delta x^1$  of spacetime is defined to be the difference between the tangent vector  $e_0(x^1 + \delta x^1)$  at the shifted position  $x^1 + \delta x^1$  and the tangent vector  $e_0(x^1)$  at the original position  $x^1$ , 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.

The last line of equation (2.44) assumes that it is legitimate to differentiate the tangent vectors  $e_\mu$ , but what does this mean? The partial derivatives of basis vectors  $e_\mu$  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-translating** 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  $\delta x$  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  $x^\mu$  on the parent manifold and tangent space coordinates  $\xi^m$  extrapolated from a locally inertial frame about each point. The tangent space coordinates  $\xi^m$  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  $e_\mu$  that appear on the right hand side of equation (2.44) define the **coordinate connection coefficients**  $\Gamma_{\mu\nu}^\kappa$ ,

$$\frac{\partial e_\mu}{\partial x^\nu} \equiv \Gamma_{\mu\nu}^\kappa e_\kappa$$

not a coordinate tensor .

(2.46)

The definition (2.46) shows that the connection coefficients express how each tangent vector  $e_\mu$  changes, relative to parallel-transport, when shifted along an interval  $\delta x^\nu$ .

#### 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

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

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

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

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

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

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,

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

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

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

#### 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  $\Gamma$  term for each contravariant index, and a negative  $\Gamma$  term for each covariant index.

## 2.10 Torsion

### 2.10.1 No-torsion condition

The existence of locally inertial frames requires that it must be possible to arrange not only that the tangent axes  $e_\mu$  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  $e_\mu$  are orthonormal, and unchanged to linear order:

$$e_\mu \cdot e_\nu = \eta_{\mu\nu} , \quad (2.53a)$$

$$\frac{\partial e_\mu}{\partial x^\nu} = 0 . \quad (2.53b)$$

In view of the definition (2.46) of the connection coefficients, the second condition (2.53b) is equivalent to the vanishing of all the connection coefficients:

$$\Gamma_{\mu\nu}^\kappa = 0 . \quad (2.54)$$

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  $4 \times 4$  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 (2.53a), which is a symmetric  $4 \times 4$  matrix equation with 10 degrees of freedom. The additional  $16 - 10 = 6$  degrees of freedom are Lorentz transformations, which rotate the tangent axes  $e_\mu$ , but leave the metric  $\eta_{\mu\nu}$  unchanged.

Just as it is possible to reorient the tangent axes  $e_\mu$  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  $\partial e_\mu / \partial x^\nu$  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  $4 \times 4$  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 (2.54) 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 (2.57), 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 (2.57) for the torsion tensor, the no-torsion condition that the connection coefficients are symmetric in their last two indices

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

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 (2.53) 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 2.3 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  $\nu$ . Parallel-transport along that direction can be defined relative to such a frame. For any given direction  $\nu$ , 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  $\Phi$

$$[D_\kappa, D_\lambda] \Phi = S_{\kappa\lambda}^\mu \frac{\partial \Phi}{\partial x^\mu} \quad \text{a coordinate tensor .} \quad (2.56)$$

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

$$S_{\kappa\lambda}^\mu = \Gamma_{\kappa\lambda}^\mu - \Gamma_{\lambda\kappa}^\mu \quad \text{a coordinate tensor} \quad (2.57)$$

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.

## 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  $e_\mu$ . 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

$$\begin{aligned} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} &= \frac{\partial e_\lambda \cdot e_\mu}{\partial x^\nu} \\ &= e_\lambda \cdot \frac{\partial e_\mu}{\partial x^\nu} + e_\mu \cdot \frac{\partial e_\lambda}{\partial x^\nu} \\ &= e_\lambda \cdot e_\kappa \Gamma_{\mu\nu}^\kappa + e_\mu \cdot e_\kappa \Gamma_{\lambda\nu}^\kappa \\ &= g_{\lambda\kappa} \Gamma_{\mu\nu}^\kappa + g_{\mu\kappa} \Gamma_{\lambda\nu}^\kappa \\ &= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} , \end{aligned} \quad (2.58)$$

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,

$$\Gamma_{\lambda\mu\nu} \equiv g_{\lambda\kappa} \Gamma_{\mu\nu}^\kappa . \quad (2.59)$$

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

$$\begin{aligned} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} &= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} + \Gamma_{\lambda\nu\mu} + \Gamma_{\nu\lambda\mu} - \Gamma_{\mu\nu\lambda} - \Gamma_{\nu\mu\lambda} \\ &= 2\Gamma_{\lambda\mu\nu} - S_{\lambda\mu\nu} - S_{\mu\nu\lambda} - S_{\nu\mu\lambda} , \end{aligned} \quad (2.60)$$

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

$$\Gamma_{\lambda\mu\nu} = \frac{1}{2} \left( \frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} + S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda} \right) \quad \text{not a coordinate tensor .} \quad (2.61)$$

If torsion vanishes, as general relativity assumes, then

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

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. Torsion arises naturally in extensions of general relativity such as supergravity (Nieuwenhuizen, 1981), which unify the Poincaré group by adjoining supersymmetry transformations.

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 (2.85) is the torsion-free covariant derivative, equation (2.87). 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 (2.62). 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  $\mathring{\Gamma}_{\lambda\mu\nu}^\lambda$  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!)  $K_{\lambda\mu\nu}$ ,

$$\Gamma_{\lambda\mu\nu} = \mathring{\Gamma}_{\lambda\mu\nu} + K_{\lambda\mu\nu} \quad \text{not a coordinate tensor .} \quad (2.63)$$

From equation (2.61), the contortion tensor  $K_{\lambda\mu\nu}$  is related to the torsion tensor  $S_{\lambda\mu\nu}$  by

$$K_{\lambda\mu\nu} = \frac{1}{2}(S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda}) = -S_{\nu\lambda\mu} + \frac{3}{2}S_{[\lambda\mu\nu]} \quad \text{a coordinate tensor .} \quad (2.64)$$

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

$$K_{\lambda\mu\nu} = -K_{\mu\lambda\nu} , \quad (2.65)$$

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  $K_{\lambda\mu\nu}$ ,

$$S_{\lambda\mu\nu} = K_{\lambda\mu\nu} - K_{\lambda\nu\mu} = -K_{\mu\nu\lambda} + 3K_{[\lambda\mu\nu]} \quad \text{a coordinate tensor .} \quad (2.66)$$

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

$$D_\nu A^\kappa \equiv \mathring{D}_\nu A^\kappa + K_{\mu\nu}^\kappa A^\mu \quad \text{a coordinate tensor .} \quad (2.67)$$

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

**Concept question 2.4 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 (2.53b). Impose instead the weaker condition that the derivatives of the metric (i.e. scalar products of tangent axes) vanish,

$$\frac{\partial g_{\lambda\mu}}{\partial x^\nu} = 0 . \quad (2.68)$$

Must torsion vanish under this weaker condition? **Answer.** No. In fact torsion may exist even in flat (Minkowski) space, where the metric is everywhere Minkowski,  $g_{\lambda\mu} = \eta_{\lambda\mu}$ . The condition (2.68) of vanishing metric derivatives is equivalent to the vanishing of the torsion-free connections,

$$\frac{1}{2} \frac{\partial g_{\lambda\mu}}{\partial x^\nu} = \Gamma_{(\lambda\mu)\nu} = \mathring{\Gamma}_{(\lambda\mu)\nu} + K_{(\lambda\mu)\nu} = \mathring{\Gamma}_{(\lambda\mu)\nu} = 0 . \quad (2.69)$$

Thus the condition (2.68) of vanishing metric derivatives imposes no condition on torsion.

**Exercise 2.5 Covariant curl and coordinate curl.** Show that the covariant curl of a covariant vector  $A_\lambda$  is

$$D_\kappa A_\lambda - D_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} + S_{\kappa\lambda}^\mu A_\mu . \quad (2.70)$$

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

$$\mathring{D}_\kappa A_\lambda - \mathring{D}_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} . \quad (2.71)$$

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 (2.70) 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 2.6 Covariant divergence and coordinate divergence.** Show that the covariant divergence of a contravariant vector  $A^\mu$  is

$$D_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu} + S_{\mu\nu}^\nu A^\mu , \quad (2.72)$$

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

$$\mathring{D}_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu} . \quad (2.73)$$

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 (2.72) and the torsion term on the right hand side of equation (2.72) are both tensors, the torsion-free covariant divergence (2.73) is a tensor even in the presence of torsion.

**Solution.** The covariant divergence is

$$D_\mu A^\mu = \frac{\partial A^\mu}{\partial x^\mu} + \Gamma_{\mu\nu}^\nu A^\mu . \quad (2.74)$$

From equation (2.61),

$$\begin{aligned} \Gamma_{\mu\nu}^\nu &= \frac{1}{2} g^{\lambda\nu} \frac{\partial g_{\lambda\nu}}{\partial x^\mu} + S_{\mu\nu}^\nu \\ &= \frac{\partial \ln |\sqrt{-g}|}{\partial x^\mu} + S_{\mu\nu}^\nu . \end{aligned} \quad (2.75)$$

The second line of equations (2.75) follows because for any matrix  $M$ ,

$$\begin{aligned}\delta \ln |M| &= \ln |M + \delta M| - \ln |M| \\ &= \ln |M^{-1}(M + \delta M)| \\ &= \ln |1 + M^{-1}\delta M| \\ &= \ln(1 + \text{Tr } M^{-1}\delta M) \\ &= \text{Tr } M^{-1}\delta M.\end{aligned}\tag{2.76}$$

The torsion-free covariant divergence is

$$\mathring{D}_{\mu} A^{\mu} = \frac{\partial A^{\mu}}{\partial x^{\mu}} + \mathring{\Gamma}_{\mu\nu}^{\nu} A^{\mu},\tag{2.77}$$

where the torsion-free coordinate connection is

$$\mathring{\Gamma}_{\mu\nu}^{\nu} = \frac{1}{2} g^{\lambda\nu} \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} = \frac{\partial \ln |\sqrt{-g}|}{\partial x^{\mu}}.\tag{2.78}$$


---

## 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  $e_{\mu}$ , and to define a dual vector space with basis  $e^{\mu}$  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

$$x^{\mu}(\tau),\tag{2.79}$$

where  $\tau$  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**  $u^{\mu}$  by

$u^{\mu} \equiv \frac{dx^{\mu}}{d\tau}$

a coordinate 4-vector .

$$\tag{2.80}$$

The magnitude squared of the 4-velocity is constant

$$u_{\mu} u^{\mu} = g_{\mu\nu} \frac{dx^{\mu}}{d\tau} \frac{dx^{\nu}}{d\tau} = \frac{ds^2}{d\tau^2} = -1.\tag{2.81}$$

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 (2.81) can be regarded as an integral of motion associated with conservation of particle rest mass.

## 2.15 Geodesic equation

Let  $\mathbf{u} \equiv e_\mu u^\mu$  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

$$\boxed{\frac{d\mathbf{u}}{d\tau} = 0} . \quad (2.82)$$

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  $e_\mu = \{e_0, e_1, e_2, e_3\}$  are constant.

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

$$\begin{aligned} \frac{d\mathbf{u}}{d\tau} &= \frac{dx^\nu}{d\tau} \frac{\partial \mathbf{u}}{\partial x^\nu} \\ &= u^\nu e_\kappa D_\nu u^\kappa \\ &= u^\nu e_\kappa \left( \frac{\partial u^\kappa}{\partial x^\nu} + \Gamma_{\mu\nu}^\kappa u^\mu \right) \\ &= e_\kappa \left( \frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu \right) . \end{aligned} \quad (2.83)$$

The geodesic equation is then

$$\boxed{\frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu = 0} . \quad (2.84)$$

Another way of writing the geodesic equation is

$$\frac{Du^\kappa}{D\tau} = 0 , \quad (2.85)$$

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

$$\frac{D}{D\tau} \equiv u^\nu D_\nu . \quad (2.86)$$

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  $\mathring{D}/D\tau = u^\mu \mathring{D}_\mu$ ,

$$\boxed{\frac{\mathring{D}u^\kappa}{D\tau} = 0} . \quad (2.87)$$

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

$$\boxed{p^\mu \equiv mu^\mu = m \frac{dx^\mu}{d\tau}} \quad \text{a coordinate 4-vector .} \quad (2.88)$$

The momentum squared is, from equation (2.81),

$$p_\mu p^\mu = m^2 u_\mu u^\mu = -m^2 \quad (2.89)$$

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  $p^\mu$  remains finite. Define the **affine parameter**  $\lambda$  by

$$\boxed{\lambda \equiv \frac{\tau}{m}} \quad \text{a coordinate scalar} \quad (2.90)$$

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

$$\boxed{p^\mu = \frac{dx^\mu}{d\lambda}} . \quad (2.91)$$

The geodesic equation is then in coordinate-independent notation

$$\frac{dp}{d\lambda} = 0 , \quad (2.92)$$

or in component form

$$\frac{dp^\kappa}{d\lambda} + \Gamma_{\mu\nu}^\kappa p^\mu p^\nu = 0 , \quad (2.93)$$

which works for massless as well as massive particles.

Another way of writing this is

$$\frac{Dp^\kappa}{D\lambda} = 0 , \quad (2.94)$$

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

$$\frac{D}{D\lambda} \equiv p^\nu D_\nu . \quad (2.95)$$

In the presence of torsion, the connection in the geodesic equation (2.93) should be interpreted as the torsion-free connection  $\mathring{\Gamma}_{\mu\nu}^\kappa$ , and the covariant derivative in equations (2.94) and (2.95) 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.

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

**Exercise 2.7 Gravitational redshift in a stationary metric.** Let  $x^\mu \equiv \{t, x^\alpha\}$  constitute time  $t$  and spatial coordinates  $x^\alpha$  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

$$u^\nu = \{\gamma, 0, 0, 0\} , \quad \gamma \equiv \frac{1}{\sqrt{-g_{tt}}} . \quad (2.96)$$

2. Argue that the proper energy  $E$  of a particle, massless or massive, with energy-momentum 4-vector  $p^\nu$  seen by a comoving observer with 4-velocity  $u^\nu$ , equation (2.96), is

$$E = -u^\nu p_\nu . \quad (2.97)$$

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

$$\frac{E_{\text{obs}}}{E_{\text{em}}} = \frac{\gamma_{\text{obs}}}{\gamma_{\text{em}}} . \quad (2.98)$$

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

**Exercise 2.8 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  $l, \alpha$  by

$$t = l \sinh \alpha , \quad x = l \cosh \alpha . \quad (2.99)$$

Show that the line-element in Rindler coordinates is

$$ds^2 = -l^2 d\alpha^2 + dl^2 + dy^2 + dz^2 . \quad (2.100)$$

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 (2.98),

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

3. Can Rindler space be considered equivalent to a spacetime containing a uniform gravitational field? Do Rindler observers all accelerate at the same rate?

**Exercise 2.9 Gravitational redshift in a uniformly rotating space.** Start with Minkowski space in cylindrical coordinates  $\{t, r, \phi, z\}$ ,

$$ds^2 = -dt^2 + dr^2 + r^2 d\phi^2 + dz^2 . \quad (2.102)$$

Define a uniformly rotating azimuthal angle  $\chi$  by

$$\chi \equiv \phi - \omega t , \quad (2.103)$$

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

$$ds^2 = -dt^2 + dr^2 + r^2(d\chi + \omega dt)^2 + dz^2 . \quad (2.104)$$

1. A comoving observer in the uniformly rotating system follows a worldline at constant  $r$ ,  $\chi$ , 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 (2.98),

$$\frac{E_{\text{obs}}}{E_{\text{em}}} = \frac{\gamma_{\text{em}}}{\gamma_{\text{obs}}} , \quad (2.105)$$

where

$$\gamma = \frac{1}{\sqrt{1 - v^2}} , \quad v = \omega r . \quad (2.106)$$

2. What happens where  $v > 1$ ?
-

## 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,

$$[D_\kappa, D_\lambda] A_\mu = S_{\kappa\lambda}^\nu D_\nu A_\mu + R_{\kappa\lambda\mu\nu} A^\nu \quad \text{a coordinate tensor .} \quad (2.107)$$

If torsion vanishes, as general relativity assumes, then the definition (2.107) reduces to

$$[D_\kappa, D_\lambda] A_\mu = R_{\kappa\lambda\mu\nu} A^\nu \quad \text{a coordinate tensor .} \quad (2.108)$$

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

$$R_{\kappa\lambda\mu\nu} = \frac{\partial \Gamma_{\mu\nu\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{\mu\nu\kappa}}{\partial x^\lambda} + \Gamma_{\mu\lambda}^\pi \Gamma_{\pi\nu\kappa} - \Gamma_{\mu\kappa}^\pi \Gamma_{\pi\nu\lambda} \quad \text{a coordinate tensor .} \quad (2.109)$$

This is the formula that allows the Riemann tensor to be calculated from the connection coefficients. The same formula (2.109) remains valid if torsion does not vanish, but the connection coefficients  $\Gamma_{\lambda\mu\nu}$  themselves are given by (2.61) in place of (2.62).

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

$$[D_\kappa, D_\lambda] \rightarrow \left[ \frac{\partial}{\partial x^\kappa}, \frac{\partial}{\partial x^\lambda} \right] = 0 \quad \text{in flat space} \quad (2.110)$$

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

**Exercise 2.10 Derivation of the Riemann tensor.** Confirm expression (2.109) 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 11.3–11.6 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 (2.56), which general relativity assumes vanishes, while the commutator of the covariant derivative acting on a vector defined the Riemann tensor, equation (2.108). 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

$$[D_\kappa, D_\lambda] A^{\pi\rho\dots}_{\mu\nu\dots} = S^\sigma_{\kappa\lambda} D_\sigma A^{\pi\rho\dots}_{\mu\nu\dots} + R_{\kappa\lambda\mu}{}^\sigma A^{\pi\rho\dots}_{\sigma\nu\dots} + R_{\kappa\lambda\nu}{}^\sigma A^{\pi\rho\dots}_{\mu\sigma\dots} - R_{\kappa\lambda\sigma}{}^\pi A^{\sigma\rho\dots}_{\mu\nu\dots} - R_{\kappa\lambda\sigma}{}^\rho A^{\pi\sigma\dots}_{\mu\nu\dots}. \quad (2.111)$$

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

$$[D_\kappa, D_\lambda] = S^\mu_{\kappa\lambda} D_\mu + \hat{R}_{\kappa\lambda} \quad (2.112)$$

where the Riemann curvature operator  $\hat{R}_{\kappa\lambda}$  is an operator whose action on any tensor is specified by equation (2.111). 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 (2.111) 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, do not vanish.

Thus in a locally inertial frame the Riemann tensor is

$$\begin{aligned} R_{\kappa\lambda\mu\nu} &= \frac{\partial\Gamma_{\mu\nu\lambda}}{\partial x^\kappa} - \frac{\partial\Gamma_{\mu\nu\kappa}}{\partial x^\lambda} \\ &= \frac{1}{2} \left( \frac{\partial^2 g_{\mu\nu}}{\partial x^\kappa \partial x^\lambda} + \frac{\partial^2 g_{\mu\lambda}}{\partial x^\kappa \partial x^\nu} - \frac{\partial^2 g_{\nu\lambda}}{\partial x^\kappa \partial x^\mu} - \frac{\partial^2 g_{\mu\nu}}{\partial x^\lambda \partial x^\kappa} - \frac{\partial^2 g_{\mu\kappa}}{\partial x^\lambda \partial x^\nu} + \frac{\partial^2 g_{\nu\kappa}}{\partial x^\lambda \partial x^\mu} \right) \\ &= \frac{1}{2} \left( \frac{\partial^2 g_{\mu\lambda}}{\partial x^\kappa \partial x^\nu} - \frac{\partial^2 g_{\nu\lambda}}{\partial x^\kappa \partial x^\mu} - \frac{\partial^2 g_{\mu\kappa}}{\partial x^\lambda \partial x^\nu} + \frac{\partial^2 g_{\nu\kappa}}{\partial x^\lambda \partial x^\mu} \right). \end{aligned} \quad (2.113)$$

You can check that the bottom line of equation (2.113):

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

$$R_{\kappa\lambda\mu\nu} + R_{\kappa\nu\lambda\mu} + R_{\kappa\mu\nu\lambda} = 0. \quad (2.114)$$

Actually, as shown in Exercise 11.6, the last two of the four symmetries, the symmetric symmetry and the cyclic symmetry, imply each other. The first three of the four symmetries can be expressed compactly

$$R_{\kappa\lambda\mu\nu} = R_{([\kappa\lambda][\mu\nu])}, \quad (2.115)$$

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

$$A_{[\kappa\lambda]} \equiv \frac{1}{2} (A_{\kappa\lambda} - A_{\lambda\kappa}), \quad A_{(\kappa\lambda)} \equiv \frac{1}{2} (A_{\kappa\lambda} + A_{\lambda\kappa}). \quad (2.116)$$

The symmetries (2.115) 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  $6 \times 6$  symmetric matrix, which has 21 degrees of freedom. The final, cyclic symmetry of the Riemann tensor, equation (2.114), which can be abbreviated

$$R_{\kappa[\lambda\mu\nu]} = 0, \quad (2.117)$$

removes 1 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 (2.109), 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 11.4–11.6.

## 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

$$R_{\kappa\mu} \equiv g^{\lambda\nu} R_{\kappa\lambda\mu\nu} \quad \text{a coordinate tensor .} \quad (2.118)$$

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

$$R_{\kappa\mu} = R_{\mu\kappa} \quad (2.119)$$

and therefore has 10 independent components.

The **Ricci scalar** is

$$R \equiv g^{\kappa\mu} R_{\kappa\mu} \quad \text{a coordinate tensor (a scalar) .} \quad (2.120)$$

## 2.21 Einstein tensor

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

$$G_{\kappa\mu} \equiv R_{\kappa\mu} - \frac{1}{2} g_{\kappa\mu} R \quad \text{a coordinate tensor .} \quad (2.121)$$

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

$$G_{\kappa\mu} = G_{\mu\kappa} , \quad (2.122)$$

and thus has 10 independent components.

## 2.22 Bianchi identities

The Jacobi identity

$$[D_\kappa, [D_\lambda, D_\mu]] + [D_\lambda, [D_\mu, D_\kappa]] + [D_\mu, [D_\kappa, D_\lambda]] = 0 \quad (2.123)$$

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

$$D_\kappa R_{\lambda\mu\nu\pi} + D_\lambda R_{\mu\kappa\nu\pi} + D_\mu R_{\kappa\lambda\nu\pi} = 0 . \quad (2.124)$$

The torsion-free Bianchi identities can be written in shorthand

$$D_{[\kappa} R_{\lambda\mu]\nu\pi} = 0 . \quad (2.125)$$

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 (2.117), 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 2.11 Jacobi identity.** Prove the Jacobi identity (2.123).

---

## 2.23 Covariant conservation of the Einstein tensor

The most important consequence of the torsion-free Bianchi identities (2.125) is obtained from the double contraction

$$g^{\kappa\nu} g^{\lambda\pi} (D_\kappa R_{\lambda\mu\nu\pi} + D_\lambda R_{\mu\kappa\nu\pi} + D_\mu R_{\kappa\lambda\nu\pi}) = -D^\kappa R_{\kappa\mu} - D^\lambda R_{\lambda\mu} + D_\mu R = 0 , \quad (2.126)$$

or equivalently

$$D^\kappa G_{\kappa\mu} = 0 , \quad (2.127)$$

where  $G_{\kappa\mu}$  is the Einstein tensor, equation (2.121). Equation (2.127) is a primary motivation for the form of the Einstein equations, since it implies energy-momentum conservation, equation (2.129).

## 2.24 Einstein equations

Einstein's equations are

$$\boxed{G_{\kappa\mu} = 8\pi G T_{\kappa\mu}} \quad \text{a coordinate tensor equation .} \quad (2.128)$$

What motivates the form of Einstein's equations?

1. The equation is generally covariant.
2. For vanishing torsion, the Bianchi identities (2.125) guarantee covariant conservation of the Einstein tensor, equation (2.127), which in turn guarantees covariant conservation of energy-momentum,

$$\boxed{D^\kappa T_{\kappa\mu} = 0} . \quad (2.129)$$

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 (2.128) 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.

## 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 (2.62).
3. Compute the Riemann tensor  $R_{\kappa\lambda\mu\nu}$  from equation (2.109).
4. Compute the Ricci tensor  $R_{\kappa\mu}$  from equation (2.118), the Ricci scalar  $R$  from equation (2.120), and the Einstein tensor  $G_{\kappa\mu}$  from equation (2.121).
5. The Einstein equations (2.128) 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 an **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

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} \quad (2.130)$$

where

$$\begin{aligned} \rho &\text{ is the proper mass-energy density ,} \\ p &\text{ is the proper pressure .} \end{aligned} \quad (2.131)$$

The expression (2.130) is valid only in the locally inertial rest frame of the fluid. An expression that is valid in any frame is

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu} , \quad (2.132)$$

where  $u^\mu$  is the 4-velocity of the fluid. Equation (2.132) 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 Newtonian limit

The Newtonian limit is obtained in the limit of a weak gravitational field and non-relativistic (pressureless) matter. In Cartesian coordinates, the metric in the Newtonian limit is (see Chapter 26)

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)(dx^2 + dy^2 + dz^2), \quad (2.133)$$

in which

$$\Phi(x, y, z) = \text{Newtonian potential} \quad (2.134)$$

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

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

$$G_{tt} = 2\nabla^2\Phi, \quad (2.135)$$

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 (2.135) of the Einstein tensor plugged into Einstein's equations (2.128) implies Poisson's equation (2.4).

**Exercise 2.12 Special and general relativistic corrections for clocks on satellites.** The metric just above the surface of the Earth is well-approximated by

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad (2.136)$$

where

$$\Phi(r) = -\frac{GM}{r} \quad (2.137)$$

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 non-relativistic 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  $\Phi$ . Regard  $v^2$  as first order in  $\Phi$ . 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)

$$u_t = -E, \quad u_\phi = L, \quad u_\mu u^\mu = -1. \quad (2.138)$$

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

$$u^r = \sqrt{2(\Delta E - U)}, \quad (2.139)$$

where  $\Delta E$  is the energy per unit mass of the particle excluding its rest mass energy,

$$\Delta E = E - 1, \quad (2.140)$$

and the effective potential  $U$  is

$$U = \Phi + \frac{L^2}{2r^2}. \quad (2.141)$$

[Hint: Neglect air resistance. Remember to work to first order in  $\Phi$ . Treat  $\Delta E$  and  $L^2$  as first order in  $\Phi$ . 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

$$\frac{d\tau_{\text{satellite}}}{dt} - 1 = -\Phi_{\oplus} (f_{\text{GR}} + f_{\text{SR}}) , \quad (2.142)$$

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

$$\Phi_{\oplus} = -\frac{GM_{\oplus}}{c^2 R_{\oplus}} . \quad (2.143)$$

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

$$\frac{d\tau_{\text{satellite}}}{dt} - \frac{d\tau_{\text{person}}}{dt} = -\Phi_{\oplus} (f_{\text{GR}} + f_{\text{SR}}) . \quad (2.144)$$

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  $\text{ms yr}^{-1}$ , 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 2.13 Equations of motion in weak gravity.** Take the metric to be the Newtonian metric (2.133) with the Newtonian potential  $\Phi(x, y, z)$  a function only of the spatial coordinates  $x, y, z$ , not of time  $t$ , equation (2.134).

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  $\Gamma_{\mu\nu}^{\kappa} = \Gamma_{\nu\mu}^{\kappa}$ )

$$\Gamma_{t\alpha}^t = \Gamma_{tt}^{\alpha} = \Gamma_{\beta\alpha}^{\alpha} = -\Gamma_{\beta\alpha}^{\beta} = -\Gamma_{\alpha\alpha}^{\alpha} = \frac{\partial \Phi}{\partial x^{\alpha}} \quad (\alpha \neq \beta = x, y, z) . \quad (2.145)$$

[Hint: Work to linear order in  $\Phi$ .]

2. Consider a massive, non-relativistic 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

$$u^t = 1 + \frac{1}{2}u^2 - \Phi , \quad (2.146)$$

whereas

$$u_t = - \left( 1 + \frac{1}{2} u^2 + \Phi \right) \quad (2.147)$$

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  $\Phi$ . Note that  $u^2$  is of linear order in  $\Phi$ .]

3. **Equation of motion of a massive particle.** From the geodesic equation

$$\frac{du^\kappa}{d\tau} + \Gamma_{\mu\nu}^\kappa u^\mu u^\nu = 0 \quad (2.148)$$

show that

$$\frac{du^\alpha}{dt} = - \frac{\partial \Phi}{\partial x^\alpha} \quad \alpha = x, y, z . \quad (2.149)$$

Why is it legitimate to replace  $d\tau$  by  $dt$ ? Show further that

$$\frac{du^t}{dt} = - 2 u^\alpha \frac{\partial \Phi}{\partial x^\alpha} \quad (2.150)$$

with implicit summation over  $\alpha = x, y, z$ . Does the result agree with what you would expect from equation (2.146)?

4. For a massless particle, the proper time along a geodesic is zero, and the affine parameter  $\lambda$  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  $p^\mu$  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

$$v^t = (1 - 2\Phi)v , \quad (2.151)$$

whereas

$$v_t = -v , \quad (2.152)$$

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

$$\frac{dv^\kappa}{d\lambda} + \Gamma_{\mu\nu}^\kappa v^\mu v^\nu = 0 \quad (2.153)$$

show that the spatial components  $\mathbf{v} \equiv \{v^x, v^y, v^z\}$  satisfy

$$\frac{d\mathbf{v}}{d\lambda} = 2 \mathbf{v} \times (\mathbf{v} \times \nabla \Phi) , \quad (2.154)$$

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 (2.154). 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

$$-u^\mu v_\mu = 1 - \Phi . \quad (2.155)$$

Does the observed photon have higher or lower energy in a deeper potential well?

**Exercise 2.14 Deflection of light by the Sun.**

1. Consider light that passes by a spherical mass  $M$  sufficiently far away that the potential  $\Phi$  is always weak. The potential at distance  $r$  from the spherical mass can be approximated by the Newtonian potential

$$\Phi = -\frac{GM}{r} . \quad (2.156)$$

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 (2.154) becomes

$$\frac{dv^y}{d\lambda} = v^x \frac{dv^y}{dx} = -2(v^x)^2 \frac{\partial \Phi}{\partial y} . \quad (2.157)$$

Integrate this equation to show that

$$\frac{\Delta v^y}{v^x} = -\frac{4GM}{y} . \quad (2.158)$$

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 2.15 Shapiro time delay.** The three classic tests of general relativity are the gravitational redshift (Exercise 2.7), the gravitational bending of light around the Sun (Exercise 2.14), and the precession of Mercury (Exercise 7.7). 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 2.13 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  $\Phi$  is (units  $c = 1$ )

$$v^\mu \equiv \left\{ \frac{dt}{d\lambda}, \frac{d\mathbf{x}}{d\lambda} \right\} = \{v^t, \mathbf{v}\} = \{1 - 2\Phi, \mathbf{v}\} , \quad (2.159)$$

where  $\mathbf{v}$  is a 3-vector of unit magnitude. Equation (2.159) implies that

$$\frac{dt}{dl} = 1 - 2\Phi , \quad (2.160)$$

where  $dl \equiv |d\mathbf{x}|$  is the magnitude of the 3-vector interval  $d\mathbf{x}$ . The Shapiro time delay comes from the  $2\Phi$  correction.

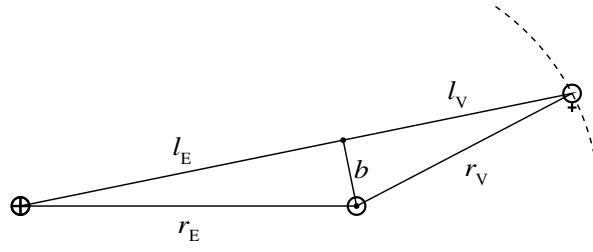


Figure 2.6 A person on Earth sends out a radio signal that passes by the Sun, bounces off the planet Venus, and returns to Earth.

1. **Time delay.** The potential  $\Phi$  at distance  $r$  from the Sun is

$$\Phi = -\frac{GM_{\odot}}{r} . \quad (2.161)$$

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  $\Delta t$  is, with units of  $c$  restored,

$$\Delta t = \frac{2}{c}(l_E + l_V) + \frac{4GM_{\odot}}{c^3} \ln \left[ \frac{(r_E + l_E)(r_V + l_V)}{b^2} \right] , \quad (2.162)$$

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 (2.162) is the Newtonian expectation, while the last term in equation (2.162) 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 2.16 Gravitational lensing.** In Exercise 2.14 you found that, in the weak field limit, light

passing a spherical mass  $M$  at impact parameter  $y$  is deflected by angle

$$\Delta\phi = \frac{4GM}{yc^2} . \quad (2.163)$$

1. **Lensing equation.** Argue that the deflection angle  $\Delta\phi$  is related to the angles  $\alpha$  and  $\beta$  illustrated in

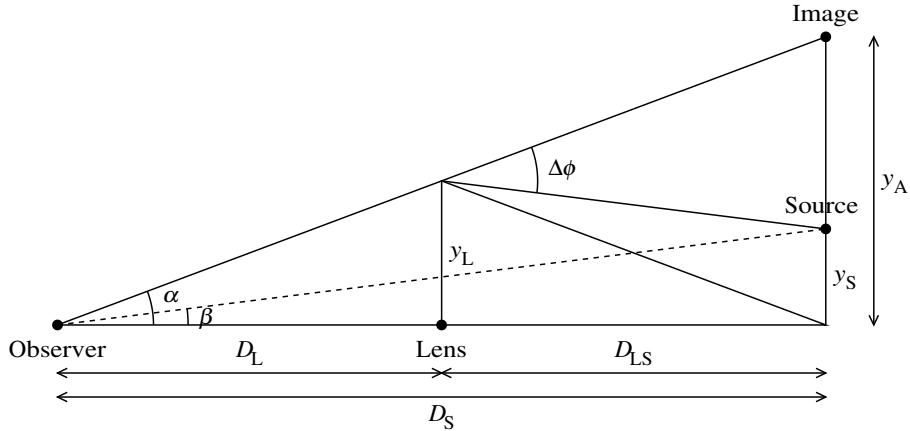


Figure 2.7 Lensing diagram.

the lensing diagram in Figure 2.7 by

$$\alpha D_S = \beta D_S + \Delta\phi D_{LS} . \quad (2.164)$$

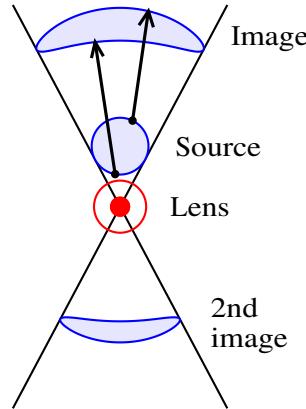


Figure 2.8 The appearance of a source lensed by a point lens. The lens in this case is a black hole, whose physical size is the filled circle, and whose apparent (lensed) size is the surrounding unfilled circle.

Hence or otherwise obtain the “lensing equation” in the form commonly used by astronomers

$$\beta = \alpha - \frac{\alpha_E^2}{\alpha}, \quad (2.165)$$

where

$$\alpha_E = \left( \frac{4GM}{c^2} \frac{D_{LS}}{D_L D_S} \right)^{1/2}. \quad (2.166)$$

2. **Solutions.** Equation (2.165) has two solutions for the apparent angles  $\alpha$  in terms of  $\beta$ . 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  $\mu$  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

$$\mu = \frac{y_A dy_A}{y_S dy_S}. \quad (2.167)$$

What is the magnification of a small source in terms of  $\alpha$  and  $\alpha_E$ ? 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 (2.115) 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 anstisymmetric components  $G_{[\kappa\mu]}$  of the Einstein tensor vanish if torsion vanishes, and likewise the 10 anstisymmetric 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.

### 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 (2.124) are

$$D_{[\kappa} R_{\lambda\mu]\nu}{}^\kappa = D_\kappa R_{\lambda\mu\nu}{}^\kappa + D_\lambda R_{\mu\nu} - D_\nu R_{\lambda\mu} = 0. \quad (3.3)$$

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 of independent torsion-free Bianchi identities, it follows that the contracted Bianchi identities (3.3) are equivalent to the full set of Bianchi identities (2.125). 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),

$$D_{[\kappa} R_{\lambda\mu]}{}^{\nu\pi} = (18 \delta_{[\kappa}^\rho \delta_\lambda^\sigma \delta_{\mu]}^{[\nu} \delta_\tau^{\pi]} + 9 \delta_\tau^\rho \delta_{[\kappa}^\sigma \delta_{\lambda]}^\nu \delta_{\mu]}^{\pi}) D_{[\nu} R_{\rho\sigma]}{}^{\tau\mu} \quad (\text{4D spacetime}). \quad (3.4)$$

If the Riemann tensor is separated into its trace (Ricci) and traceless (Weyl) parts, equation (3.1), then the contracted Bianchi identities (3.3) become the Weyl evolution equations

$$D^\kappa C_{\kappa\lambda\mu\nu} = J_{\lambda\mu\nu}, \quad (3.5)$$

where  $J_{\lambda\mu\nu}$  is the Weyl current

$$J_{\lambda\mu\nu} \equiv \frac{1}{2} (D_\mu G_{\lambda\nu} - D_\nu G_{\lambda\mu}) - \frac{1}{6} (g_{\lambda\nu} D_\mu G - g_{\lambda\mu} D_\nu G). \quad (3.6)$$

The Weyl evolution equations (3.5) 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 (2.114), valid for vanishing torsion, which implies the cyclic symmetry

$$J_{[\lambda\mu\nu]} = 0. \quad (3.7)$$

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 (3.5) on  $\lambda\mu$  yields zero on the left hand side, so that the contracted Weyl current satisfies

$$J^\lambda{}_{\lambda\nu} = 0. \quad (3.8)$$

This doubly-contracted Bianchi identity, which is the same as equation (2.127), enforces conservation of energy-momentum. Unlike the cyclic symmetry (3.7), which follows from the cyclic symmetry of the Riemann tensor and is not a differential condition on the Riemann tensor, equations (3.8) constitute a non-trivial set of 4 differential conditions on the Einstein tensor. Besides the algebraic relations (3.7) and (3.8), the Weyl current satisfies 6 differential identities comprising the conservation law

$$D^\lambda J_{\lambda\mu\nu} = 0 \quad (3.9)$$

in view of equation (3.5) and the antisymmetry of  $C_{\kappa\lambda\mu\nu}$  with respect to the indices  $\kappa\lambda$ . The Weyl current conservation law (3.9) follows from the form (3.6) of the Weyl current, coupled with covariant conservation of the Einstein tensor, equation (2.127), so does not impose any additional non-trivial conditions on the Riemann tensor. The Weyl current conservation law (3.9) is the gravitational analogue of the conservation law for electric current that follows from Maxwell's equations.

The 4 relations (3.8) and the 6 identities (3.9) account for 10 of the 20 contracted torsion-free Bianchi identities (3.3). The remaining 10 equations comprise Maxwell-like equations (3.5) for the evolution of the 10 components of the Weyl tensor.

Whereas the Einstein equations relating the Einstein tensor to the energy-momentum tensor are postulated equations of general relativity, the 10 evolution equations for the Weyl tensor, and the 4 equations enforcing covariant conservation of the Einstein tensor, follow mathematically from the Bianchi identities, and do not represent additional assumptions of the theory.

**Exercise 3.1 Number of Bianchi identities.** Confirm the counting of degrees of freedom.

**Exercise 3.2 Wave equation for the Riemann tensor.** From the Bianchi identities, show that the Riemann tensor satisfies the covariant wave equation

$$\square R_{\kappa\lambda\mu\nu} = D_\kappa D_\mu R_{\lambda\nu} - D_\kappa D_\nu R_{\lambda\mu} + D_\lambda D_\nu R_{\kappa\mu} - D_\lambda D_\mu R_{\kappa\nu} , \quad (3.10)$$

where  $\square$  is the D'Alembertian operator, the 4-dimensional wave operator

$$\square \equiv D^\pi D_\pi . \quad (3.11)$$

Show that contracting equation (3.10) with  $g^{\lambda\nu}$  yields the identity  $\square R_{\kappa\mu} = \square R_{\kappa\mu}$ . Conclude that the wave equation (3.10) is non-trivial only for the trace-free part of the Riemann tensor, the Weyl tensor  $C_{\kappa\lambda\mu\nu}$ . Show that the wave equation for the Weyl tensor is

$$\begin{aligned} \square C_{\kappa\lambda\mu\nu} &= (D_\kappa D_\mu - \frac{1}{2} g_{\kappa\mu} \square) R_{\lambda\nu} - (D_\kappa D_\nu - \frac{1}{2} g_{\kappa\nu} \square) R_{\lambda\mu} \\ &\quad + (D_\lambda D_\nu - \frac{1}{2} g_{\lambda\nu} \square) R_{\kappa\mu} - (D_\lambda D_\mu - \frac{1}{2} g_{\lambda\mu} \square) R_{\kappa\nu} \\ &\quad + \frac{1}{6} (g_{\kappa\mu} g_{\lambda\nu} - g_{\kappa\nu} g_{\lambda\mu}) \square R . \end{aligned} \quad (3.12)$$

Conclude that in a vacuum, where  $R_{\kappa\mu} = 0$ ,

$$\square C_{\kappa\lambda\mu\nu} = 0 . \quad (3.13)$$

### 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  $\delta x^\mu$  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  $\delta x^\kappa$  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  $\mu$  is put downstairs so that the final equation (3.20) looks cosmetically better, but of course since everything is covariant the  $\mu$  index could just as well be put upstairs everywhere):

$$\begin{aligned} 0 &= \delta \frac{Du_\mu}{D\tau} \\ &= \delta x^\kappa D_\kappa (u^\lambda D_\lambda u_\mu) \\ &= \delta u^\lambda D_\lambda u_\mu + \delta x^\kappa u^\lambda D_\kappa D_\lambda u_\mu. \end{aligned} \quad (3.16)$$

On the second line, the covariant difference  $\delta$  between quantities a small distance  $\delta x^\kappa$  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  $\delta u^\lambda$ . Next, consider the covariant acceleration of the interval  $\delta x_\mu$ , which is the covariant proper time derivative of the covariant velocity difference  $\delta u_\mu$ :

$$\begin{aligned} \frac{D^2 \delta x_\mu}{D\tau^2} &= \frac{D\delta u_\mu}{D\tau} \\ &= u^\lambda D_\lambda (\delta x^\kappa D_\kappa u_\mu) \\ &= \delta u^\kappa D_\kappa u_\mu + \delta x^\kappa u^\lambda D_\lambda D_\kappa u_\mu. \end{aligned} \quad (3.17)$$

As in the previous equation (3.16), on the second line  $D/D\tau$  has been set equal to  $u^\lambda D_\lambda$ , while  $\delta$  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  $\delta u^\mu$ , equation (3.15). Subtracting (3.16) from (3.17) gives

$$\frac{D^2 \delta x_\mu}{D\tau^2} = \delta x^\kappa u^\lambda [D_\lambda, D_\kappa] u_\mu , \quad (3.18)$$

or equivalently

$$\frac{D^2 \delta x_\mu}{D\tau^2} + S_{\kappa\lambda}^\nu \delta x^\kappa u^\lambda D_\nu u_\mu + R_{\kappa\lambda\mu\nu} \delta x^\kappa u^\lambda u^\nu = 0 . \quad (3.19)$$

If torsion vanishes as general relativity assumes, then

$$\boxed{\frac{D^2 \delta x_\mu}{D\tau^2} + R_{\kappa\lambda\mu\nu} \delta x^\kappa u^\lambda u^\nu = 0} , \quad (3.20)$$

which is the desired equation of geodesic deviation.

# 4

---

## 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  $\lambda$ . In non-relativistic mechanics it is usual to take the parameter  $\lambda$  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  $\lambda$ , as illustrated in Figure 4.1. The parameter  $\lambda$  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**

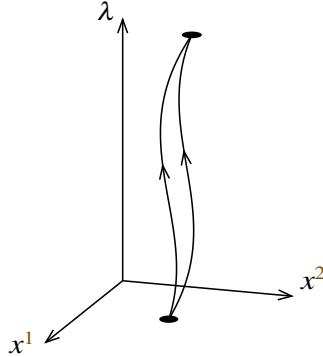


Figure 4.1 The action principle considers various paths through spacetime between fixed initial and final conditions, and chooses that path that minimizes the action.

$L(x^\mu, dx^\mu/d\lambda)$  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  $\lambda$ . The action from an initial state at  $\lambda_i$  to a final state at  $\lambda_f$  is thus

$$\boxed{S = \int_{\lambda_i}^{\lambda_f} L \left( x^\mu, \frac{dx^\mu}{d\lambda} \right) d\lambda.} \quad (4.1)$$

The principle of least action demands that the actual path taken by the system between given initial and final coordinates  $x_i^\mu$  and  $x_f^\mu$  is such as to minimize the action. Thus the variation  $\delta S$  of the action must be zero under any change  $\delta x^\mu$  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  $\delta x^\mu$  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  $\lambda$ . But in fact the Euler-Lagrange equations are the same regardless of the choice of  $\lambda$ . An example of the arbitrariness of  $\lambda$  will be seen in §4.3. Since  $\lambda$  can be chosen arbitrarily, it is common to choose it in some convenient fashion. For a massive particle,  $\lambda$  can be taken equal to the proper time  $\tau$  of the particle. For a massless particle, whose proper time never progresses,  $\lambda$  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  $\lambda$  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**  $\pi_\mu$  conjugate to the coordinate  $x^\mu$ ,

$$\boxed{\pi_\mu \equiv \frac{\partial L}{\partial(dx^\mu/d\lambda)}} . \quad (4.6)$$

## 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  $x^\mu$  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  $x^\mu$  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

$$\frac{\partial L_m}{\partial(dx^\kappa/d\lambda)} = mg_{\kappa\nu} \frac{dx^\nu}{d\lambda} \frac{d\lambda}{d\tau} = mu_\kappa , \quad (4.10a)$$

$$\frac{\partial L_m}{\partial x^\kappa} = \frac{1}{2} m \frac{\partial g_{\mu\nu}}{\partial x^\kappa} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} \frac{d\lambda}{d\tau} = m\Gamma_{\mu\nu\kappa} u^\mu u^\nu \frac{d\tau}{d\lambda} , \quad (4.10b)$$

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 (2.58). The generalized momentum  $\pi_\kappa$ , equation (4.6), of the test particle coincides with its ordinary momentum  $p_\kappa$ :

$$\pi_\kappa = p_\kappa \equiv mu_\kappa . \quad (4.11)$$

The resulting Euler-Lagrange equations of motion (4.5) are

$$\frac{dmu_\kappa}{d\lambda} = m\Gamma_{\mu\nu\kappa} u^\mu u^\nu \frac{d\tau}{d\lambda} . \quad (4.12)$$

As remarked in §4.1, the choice of the arbitrary parameter  $\lambda$  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 (2.63), gives

$$\frac{du_\kappa}{d\tau} = (\mathring{\Gamma}_{\mu\nu\kappa} + K_{\mu\nu\kappa})u^\mu u^\nu = \mathring{\Gamma}_{\mu\nu\kappa}u^\mu u^\nu , \quad (4.14)$$

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  $\tau$  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 (2.90), which remains finite in the limit  $m \rightarrow 0$ . In terms of the affine parameter  $\lambda$ , the momentum  $p^\kappa$  of a particle can be written

$$p^\kappa \equiv mu^\kappa = \frac{dx^\kappa}{d\lambda} , \quad (4.16)$$

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  $\lambda$  defined by equation (2.90) 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  $\lambda$  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  $x^\alpha$ , and the  $3N$  spatial velocities  $dx^\alpha/dt$ . 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  $\lambda$  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  $\lambda$  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

$$\boxed{L'_m = -\frac{L_m^2}{2m^2} = \frac{1}{2}g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}}. \quad (4.25)$$

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  $p_\mu$ , 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  $\lambda$ , equation (2.90). Thus the  $\lambda$  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  $\lambda$  that labels the path of the particle between its fixed endpoints. The equations of motion are said to be **reparametrization independent**. Introduce, therefore, a parameter  $\mu(\lambda)$ , an arbitrary function of  $\lambda$ , that rescales the parameter  $\lambda$ , and let the action for a test particle of mass  $m$  be

$$S_m = \int \frac{1}{2} \left( g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} - m^2 \right) \mu d\lambda , \quad (4.30)$$

with nice Lagrangian

$$\boxed{L_m = \frac{\mu}{2} \left( g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} - m^2 \right)} . \quad (4.31)$$

Variation of the action with respect to  $x^\mu$  and  $dx^\mu/d\lambda$  yields the Euler-Lagrange equations in the form

$$\frac{d}{d\lambda} \left( g_{\kappa\nu} \frac{dx^\nu}{\mu d\lambda} \right) = \Gamma_{\mu\nu\kappa} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} . \quad (4.32)$$

Variation of the action with respect to the parameter  $\mu$  gives

$$\delta S_m = \int \frac{1}{2} \left( -g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} - m^2 \right) \delta\mu d\lambda , \quad (4.33)$$

and requiring that this be an extremum imposes

$$g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} = -m^2 . \quad (4.34)$$

Equation (4.34) is equivalent to

$$\mu d\lambda = \frac{d\tau}{m} , \quad (4.35)$$

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  $A_\mu$  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  $A_\mu$  is taken to be a prescribed function of the coordinates  $x^\mu$  along the path of the particle, then the Lagrangian  $L_q$  (4.38) is a function of coordinates  $x^\mu$  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  $\pi_\kappa$ , equation (4.6), of the test particle of mass  $m$  and charge  $q$  in the electromagnetic field of potential  $A_\mu$  is, from equations (4.10a) and (4.39a),

$$\pi_\kappa \equiv \frac{\partial(L_m + L_q)}{\partial(dx^\kappa/d\lambda)} = mu_\kappa + qA_\kappa . \quad (4.40)$$

Applied to the Lagrangian  $L = L_m + L_q$ , the Euler-Lagrange equations (4.5) are

$$\frac{d}{d\lambda} (mu_\kappa + qA_\kappa) = \left( m\Gamma_{\mu\nu\kappa} u^\mu u^\nu + q \frac{\partial A_\mu}{\partial x^\kappa} u^\mu \right) \frac{d\tau}{d\lambda} , \quad (4.41)$$

which rearranges to

$$\frac{dmu_\kappa}{d\tau} = m\Gamma_{\mu\nu\kappa} u^\mu u^\nu + qF_{\kappa\mu} u^\mu , \quad (4.42)$$

where the antisymmetric electromagnetic field tensor  $F_{\kappa\mu}$  is defined to be the torsion-free covariant curl of the electromagnetic potential  $A_\mu$ ,

$$F_{\kappa\mu} \equiv \frac{\partial A_\mu}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\mu} . \quad (4.43)$$

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,

$$\frac{\dot{D}mu_\kappa}{D\tau} = qF_{\kappa\mu} u^\mu . \quad (4.44)$$

Equation (4.44), which involves the torsion-free covariant derivative  $\dot{D}/D\tau$ , 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  $\phi$  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  $\phi$  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  $\xi$ . 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  $\xi$ . Then the Euler-Lagrangian equations

of motion (4.5) imply that the derivative of the covariant  $\xi$ -component  $\pi_\xi$  of the conjugate momentum of the particle vanishes along the trajectory of the particle,

$$\frac{d\pi_\xi}{d\lambda} = \frac{\partial L}{\partial \dot{\xi}} = 0 . \quad (4.49)$$

Thus the covariant momentum  $\pi_\xi$  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  $\xi$  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  $\xi$ . The factor  $e^\xi$  is called a conformal factor. The Euler-Lagrangian equation of motion (4.5) for the conformal coordinate  $\xi$  is then

$$\frac{d\pi_\xi}{d\lambda} = \frac{\partial L}{\partial \dot{\xi}} = 2L . \quad (4.52)$$

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  $\xi$ . 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  $\tau$  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 (2.100) 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  $\alpha$  and  $\xi$  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  $\alpha$ ,  $y$ , 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

$$p^\nu p_\nu = -m^2 . \quad (4.60)$$

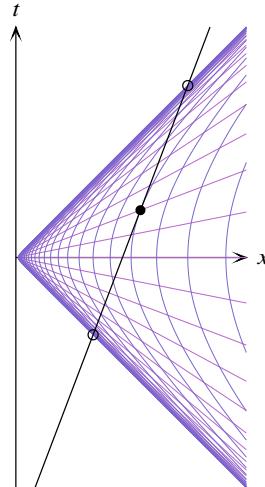


Figure 4.2 Rindler wedge of Minkowski space. Purple and blue lines are lines of constant Rindler time  $\alpha$  and constant Rindler spatial coordinate  $\xi$  respectively. The grid of lines is equally spaced by 0.2 in each of  $\alpha$  and  $\xi$ . The Rindler coordinates  $\alpha$  and  $\xi$ , 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  $\alpha$  and  $\xi$  (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  $\alpha$ - $\xi$  grid are rhombi with null ( $45^\circ$ ) 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  $\mu$  is the positive constant

$$\mu \equiv \sqrt{p_y^2 + p_z^2 + m^2}. \quad (4.62)$$

Equation (4.61) integrates to give  $\xi$  as a function of  $\lambda$ ,

$$e^{2\xi} = \frac{p_{\alpha}^2}{\mu^2} - \mu^2 \lambda^2, \quad (4.63)$$

where a constant of integration has been absorbed without loss of generality into a shift of the zero point of the affine parameter  $\lambda$  along the trajectory of the particle. The coordinate  $\xi$  passes through its maximum value  $\xi_0$  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  $\alpha$ . 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  $\xi$ ,

$$\frac{dp_{\xi}}{d\lambda} = -\mu^2. \quad (4.65)$$

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  $\lambda$  (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  $\alpha$  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  $\alpha_0$  is the value of  $\alpha$  for  $\lambda = 0$ , where  $\xi$  takes its maximum  $\xi_0$ . The Rindler time coordinate  $\alpha$  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  $x^\mu$  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  $x^\mu$  and the  $4N$  generalized momenta  $\pi_\mu$ , 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

$$H \equiv \pi_\mu \frac{dx^\mu}{d\lambda} - L . \quad (4.68)$$

Here, as previously in §4.1, the parameter  $\lambda$  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  $\lambda$  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  $\lambda_i$  and  $\lambda_f$  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**

$$\frac{dx^\mu}{d\lambda} = \frac{\partial H}{\partial \pi_\mu} , \quad \frac{d\pi_\mu}{d\lambda} = - \frac{\partial H}{\partial x^\mu} . \quad (4.72)$$

## 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  $\lambda$  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  $x^\alpha$  and  $3N$  spatial velocities  $dx^\alpha/dt$ . 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  $x^\alpha$  and corresponding  $3N$  generalized momenta  $\pi_\alpha$ . 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  $x^\mu$  and momenta  $p_\mu$ . But the expression (4.68) for the Hamiltonian depends on the arbitrary parameter  $\lambda$ , whereas as seen in §4.3 the coordinates  $x^\mu$  and momenta  $p_\mu$  are (before the least action principle is applied) independent of the choice of  $\lambda$ . There is no way to express the Hamiltonian in the prescribed form without imposing some additional constraint on  $\lambda$ . 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  $\lambda$  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  $\eta_{mn}$ . 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  $\pi_a$  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  $x^\alpha$  and momenta  $p_\alpha$  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  $p_\alpha$  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  $\lambda$  is applied *after* equations of motion are derived. The effective test-particle Lagrangian (4.25), coupled to electromagnetism, is

$$L = L_m + L_q = \frac{1}{2} g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} + qA_\mu \frac{dx^\mu}{d\lambda}, \quad (4.88)$$

where the metric  $g_{\mu\nu}$  and electromagnetic potential  $A_\mu$  are considered to be given functions of the coordinates  $x^\mu$ . The corresponding generalized momentum (4.6) is

$$\pi_\mu = g_{\mu\nu} \frac{dx^\nu}{d\lambda} + qA_\mu. \quad (4.89)$$

The (super-)Hamiltonian (4.68) expressed in terms of coordinates  $x^\mu$  and momenta  $\pi_\mu$  as required is

$$H = \frac{1}{2} g^{\mu\nu} (\pi_\mu - qA_\mu)(\pi_\nu - qA_\nu). \quad (4.90)$$

Hamilton's equations (4.72) are

$$\frac{dx^\mu}{d\lambda} = p^\mu, \quad \frac{dp_\kappa}{d\lambda} = \Gamma_{\mu\nu\kappa} p^\mu p^\nu + qF_{\kappa\mu} p^\mu, \quad (4.91)$$

where  $p_\mu$  is defined by

$$p_\mu \equiv \pi_\mu - qA_\mu. \quad (4.92)$$

The equations of motion (4.91) having been derived from the Hamiltonian (4.90), the parameter  $\lambda$  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

$$L = \frac{\mu}{2} \left( g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} - m^2 \right) + qA_\mu \frac{dx^\mu}{d\lambda}. \quad (4.94)$$

The corresponding generalized momentum (4.6) is

$$\pi_\mu = g_{\mu\nu} \frac{dx^\nu}{\mu d\lambda} + qA_\mu . \quad (4.95)$$

The associated nice (super-)Hamiltonian (4.68) expressed in terms of coordinates  $x^\mu$  and momenta  $\pi_\mu$  as required is

$$H = \frac{\mu}{2} [g^{\mu\nu}(\pi_\mu - qA_\mu)(\pi_\nu - qA_\nu) + m^2] . \quad (4.96)$$

The nice Hamiltonian  $H$ , equation (4.96), depends on the auxiliary parameter  $\mu$  as well as on  $x^\mu$  and  $\pi_\mu$ , 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  $\delta\mu$ :

$$\delta S = [\pi_\mu \delta x^\mu]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left\{ - \left( \frac{d\pi_\mu}{d\lambda} + \frac{\partial H}{\partial x^\mu} \right) \delta x^\mu + \left( \frac{dx^\mu}{d\lambda} - \frac{\partial H}{\partial \pi_\mu} \right) \delta \pi_\mu - \frac{\partial H}{\partial \mu} \delta \mu \right\} d\lambda . \quad (4.97)$$

Requiring that the variation (4.97) of the action vanish under arbitrary variations of the coordinates  $x^\mu$  and momenta  $\pi_\mu$  yields Hamilton's equations (4.72), which here are

$$\frac{dx^\mu}{\mu d\lambda} = p^\mu , \quad \frac{dp_\mu}{\mu d\lambda} = \Gamma_{\mu\nu\kappa} p^\mu p^\nu + qF_{\kappa\mu} p^\mu , \quad (4.98)$$

with  $p_\mu$  defined by

$$p_\mu \equiv \pi_\mu - qA_\mu . \quad (4.99)$$

The condition (4.103) found below, substituted into the first of Hamilton's equations (4.98), implies that  $p^\mu$  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  $\mu$  yields the additional equation of motion

$$\frac{\partial H}{\partial \mu} = 0 . \quad (4.100)$$

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

$$g_{\mu\nu} \frac{dx^\mu}{\mu d\lambda} \frac{dx^\nu}{\mu d\lambda} = -m^2 , \quad (4.102)$$

which in turn is equivalent to

$$\mu d\lambda = \frac{d\tau}{m} , \quad (4.103)$$

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  $\lambda$  along the actual path equals the Lagrangian  $L$ . In addition to being a function of the coordinates  $x^\mu$  along the actual path, the action may also be an explicit function  $S(\lambda, x^\mu)$  of the parameter  $\lambda$ . 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  $\lambda$  is minus the Hamiltonian

$$\frac{\partial S}{\partial \lambda} = -H . \quad (4.107)$$

In the conventional approach where the parameter  $\lambda$  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  $x^\mu$  and generalized momenta  $\pi_\mu$ , and replacing the Hamiltonian  $H$  by  $-\partial S/d\lambda$  in accordance with equation (4.107), and the generalized momenta  $\pi_\mu$  by  $\partial S/\partial x^\mu$  in accordance with equation (4.105).

For the effective Hamiltonian (4.90), the resulting Hamilton-Jacobi equation is

$$-\frac{\partial S}{\partial \lambda} = \frac{1}{2} g^{\mu\nu} \left( \frac{\partial S}{\partial x^\mu} - q A_\mu \right) \left( \frac{\partial S}{\partial x^\nu} - q A_\nu \right) , \quad (4.110)$$

whose left hand side is  $-m^2/2$ , equation (4.93). For the nice Hamiltonian (4.96), the resulting Hamilton-Jacobi equation is

$$-\frac{\partial S}{\mu \partial \lambda} = \frac{1}{2} \left[ g^{\mu\nu} \left( \frac{\partial S}{\partial x^\mu} - q A_\mu \right) \left( \frac{\partial S}{\partial x^\nu} - q A_\nu \right) + m^2 \right] , \quad (4.111)$$

whose left hand side is zero, 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  $x^\mu$  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  $q^\mu$  and  $p_\mu$ ,

$$q^\mu , \quad p_\mu . \quad (4.112)$$

### 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

$$S = \int_{\lambda_i}^{\lambda_f} p_{\mu} dq^{\mu} - H d\lambda , \quad S' = \int_{\lambda_i}^{\lambda_f} p'_{\mu} dq'^{\mu} - H' d\lambda . \quad (4.114)$$

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  $\lambda_i$  and  $\lambda_f$ , 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  $p_{\mu}$ ,  $-p'_{\mu}$ , and  $H - H'$  as partial derivatives of  $F$  with respect to  $q^{\mu}$ ,  $q'^{\mu}$ , and  $\lambda$ . 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  $q^{\mu}$  and  $q'^{\mu}$ . Other generating functions depending on either of  $q^{\mu}$  or  $p_{\mu}$ , and either of  $q'^{\mu}$  or  $p'_{\mu}$ , 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  $q^{\nu}$  but not of the momenta  $p_{\nu}$ , 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  $\lambda$ , 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  $\lambda$  is arbitrary, the Hamiltonian is without loss of generality independent of  $\lambda$ , and there is no physical significance to canonical transformations generated by functions that depend on  $\lambda$ . The super-Hamiltonian  $H(q^\mu, p_\mu)$  is then a scalar, invariant with respect to canonical transformations that do not depend explicitly on  $\lambda$ . This contrasts with the conventional Hamiltonian approach, where the parameter  $\lambda$  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  $t$ .

#### 4.17.2 Evolution is a canonical transformation

The evolution of the system from some initial hypersurface  $\lambda = 0$  to some final hypersurface  $\lambda$  is itself a canonical transformation. This is evident from the fact that Hamilton's equations (4.72) hold for any value of the parameter  $\lambda$ , 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  $p_\mu(\lambda)$ ,

$$q^\mu(0) \rightarrow q'^\mu = q^\mu(\lambda) , \quad p_\mu(0) \rightarrow p'_\mu = p_\mu(\lambda) . \quad (4.121)$$

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),

$$dF = p_\mu(0) dq^\mu(0) - p_\mu(\lambda) dq^\mu(\lambda) - [H(0) - H(\lambda)] d\lambda = dS(0) - dS(\lambda) . \quad (4.122)$$

Thus the canonical transformation from an initial  $\lambda = 0$  to a final  $\lambda$  is generated by the difference in the actions along the actual path of the system,

$$F = S(0) - S(\lambda) . \quad (4.123)$$

### 4.18 Symplectic structure

The generalized coordinates  $q^\mu$  and momenta  $p_\mu$  of a dynamical system of particles have a geometrical structure that transcends the geometrical structure of the underlying spacetime manifold. For  $N$  coordinates  $q^\mu$  and  $N$  momenta  $p_\mu$ , 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**  $\omega_{ij}$ . Let  $z^i$  with  $i = 1, \dots, 2N$  denote the combined set of  $2N$  generalized

coordinates and momenta  $\{q^{\mu}, p_{\mu}\}$ ,

$$\{z^1, \dots, z^N, z^{N+1}, \dots, z^{2N}\} \equiv \{q^1, \dots, q^N, p_1, \dots, p_N\}. \quad (4.124)$$

Hamilton's equations (4.72) can be written

$$\frac{dz^i}{d\lambda} = \omega^{ij} \frac{\partial H}{\partial z^j}, \quad (4.125)$$

where  $\omega^{ij}$  is the antisymmetric symplectic metric (actually the inverse symplectic metric)

$$\omega^{ij} \equiv \delta_{i+N, j} - \delta_{i, j+N} = \begin{cases} 1 & \text{if } z^i = q^{\mu} \text{ and } z^j = p_{\mu}, \\ -1 & \text{if } z^i = p_{\mu} \text{ and } z^j = q^{\mu}, \\ 0 & \text{otherwise.} \end{cases} \quad (4.126)$$

As a matrix, the symplectic metric  $\omega^{ij}$  is the  $2N \times 2N$  matrix

$$\omega^{ij} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad (4.127)$$

where 1 denotes the  $N \times N$  unit matrix. Inverting the inverse symplectic metric  $\omega^{ij}$  yields the symplectic metric  $\omega_{ij}$ , which is the same matrix but flipped in sign,

$$\omega_{ij} \equiv (\omega^{ij})^{-1} = (\omega^{ij})^{\top} = -\omega^{ij} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \quad (4.128)$$

Let  $z'^i$  be another set of generalized coordinates and momenta satisfying Hamilton's equations with the same Hamiltonian  $H$ ,

$$\frac{dz'^i}{d\lambda} = \omega^{ij} \frac{\partial H}{\partial z^j}. \quad (4.129)$$

It is being assumed here that the Hamiltonian  $H$  does not depend explicitly on the parameter  $\lambda$ . In the super-Hamiltonian approach, there is no loss of generality in taking the Hamiltonian  $H$  to be independent of  $\lambda$ , since the parameter  $\lambda$  is arbitrary, without physical significance. The important point about equation (4.129) is that the symplectic metric  $\omega^{ij}$  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

$$\frac{dz'^i}{d\lambda} = \frac{\partial z'^i}{\partial z^k} \frac{dz^k}{d\lambda} = \frac{\partial z'^i}{\partial z^k} \omega_{kl} \frac{\partial H}{\partial z^l} = \frac{\partial z'^i}{\partial z^k} \omega_{kl} \frac{\partial z'^j}{\partial z^l} \frac{\partial H}{\partial z'^j}. \quad (4.130)$$

Comparing equations (4.129) and (4.130) shows that the symplectic matrix  $\omega^{ij}$  is invariant under a canonical transformation,

$$\omega^{ij} = \frac{\partial z'^i}{\partial z^k} \omega_{kl} \frac{\partial z'^j}{\partial z^l}. \quad (4.131)$$

Equation (4.131) can be expressed as the invariance under canonical transformations of

$$\omega^{ij} \frac{\partial}{\partial z^i} \frac{\partial}{\partial z^j} = \omega^{ij} \frac{\partial}{\partial z'^i} \frac{\partial}{\partial z'^j}. \quad (4.132)$$

Equivalently,

$$\omega_{ij} dz^i dz^j = \omega_{ij} dz'^i dz'^j . \quad (4.133)$$

The invariance of the symplectic metric  $\omega_{ij}$  under canonical transformations can be thought of as analogous to the invariance of the Minkowski metric  $\eta_{mn}$  under Lorentz transformations. But whereas the Minkowski metric  $\eta_{mn}$  is symmetric, the symplectic metric  $\omega_{ij}$  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  $g$ ,

$$[f, g] \equiv \omega^{ij} \frac{\partial f}{\partial z^i} \frac{\partial g}{\partial z^j} = \frac{\partial f}{\partial q^\mu} \frac{\partial g}{\partial p_\mu} - \frac{\partial f}{\partial p_\mu} \frac{\partial g}{\partial q^\mu} . \quad (4.134)$$

The invariance (4.132) 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  $\omega^{ij}$ ,

$$[f, g] = -[g, f] . \quad (4.135)$$

### 4.19.1 Poisson brackets of phase-space coordinates

The Poisson brackets of the phase-space coordinates and momenta themselves satisfy

$$[z^i, z^j] = \omega^{ij} . \quad (4.136)$$

Explicitly in terms of the generalized coordinates and momenta  $q^\mu$  and  $p_\mu$ ,

$$[q^\mu, p_\nu] = \delta_\nu^\mu , \quad [q^\mu, q^\nu] = 0 , \quad [p_\mu, p_\nu] = 0 . \quad (4.137)$$

Reinterpreting equations (4.137) as operator equations provides a path from classical to quantum mechanics.

## 4.20 (Super-)Hamiltonian as a generator of evolution

The Poisson bracket of a function  $f(z^i)$  with the Hamiltonian  $H$  is

$$[f, H] = \frac{\partial f}{\partial q^\mu} \frac{\partial H}{\partial p_\mu} - \frac{\partial f}{\partial p_\mu} \frac{\partial H}{\partial q^\mu} . \quad (4.138)$$

Inserting Hamilton's equations (4.72) implies

$$[f, H] = \frac{\partial f}{\partial q^\mu} \frac{dq^\mu}{d\lambda} + \frac{\partial f}{\partial p_\mu} \frac{dp_\mu}{d\lambda} = \frac{df}{d\lambda} . \quad (4.139)$$

That is, the evolution of a function  $f(q^\mu, p_\mu)$  of generalized coordinates and momenta is its Poisson bracket with the Hamiltonian  $H$ ,

$$\frac{df}{d\lambda} = [f, H] . \quad (4.140)$$

Equation (4.140) 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  $\lambda$  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  $q^\alpha$  and  $p_\alpha$ . Equation (4.140) becomes in the conventional case

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + [f, H] . \quad (4.141)$$

## 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  $\epsilon$  an infinitesimal parameter,

$$G = q^\mu p'_\mu + \epsilon g(q, p') . \quad (4.142)$$

The resulting canonical transformation is, from equation (4.119),

$$q'^\mu = \frac{\partial G}{\partial p'_\mu} = q^\mu + \epsilon \frac{\partial g}{\partial p'_\mu} , \quad p'_\mu = \frac{\partial G}{\partial q^\mu} = p'_\mu + \epsilon \frac{\partial g}{\partial q^\mu} . \quad (4.143)$$

Because  $\epsilon$  is infinitesimal, the term  $\epsilon \partial g / \partial p'_\mu$  can be replaced by  $\epsilon \partial g / \partial p_\mu$  to linear order, yielding

$$q'^\mu = q^\mu + \epsilon \frac{\partial g}{\partial p_\mu} , \quad p'_\mu = p_\mu - \epsilon \frac{\partial g}{\partial q^\mu} . \quad (4.144)$$

Equations (4.144) imply that the changes  $\delta p_\mu$  and  $\delta q^\mu$  in the coordinates and momenta under an infinitesimal canonical transformation (4.142) is their Poisson bracket with  $g$ ,

$$\delta p_\mu = \epsilon [p_\mu, g] , \quad \delta q^\mu = \epsilon [q^\mu, g] . \quad (4.145)$$

As a particular example, the evolution of the system under an infinitesimal change  $\delta\lambda$  in the parameter  $\lambda$  is, in accordance with the evolutionary equation (4.140), generated by a canonical transformation with  $g$  in equation (4.142) set equal to the Hamiltonian  $H$ ,

$$\delta p_\mu = \delta\lambda [p_\mu, H] , \quad \delta q^\mu = \delta\lambda [q^\mu, H] . \quad (4.146)$$

## 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

$$V \equiv \int dV \equiv \int dz^1 \dots dz^{2N} \equiv \int dq^1 \dots dq^N dp_1 \dots dp_N , \quad (4.147)$$

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  $|\partial z'^i / \partial z^j|$ ,

$$dV' = \left| \frac{\partial z'^i}{\partial z^j} \right| dV . \quad (4.148)$$

But equation (4.131) implies that

$$|\omega^{ij}| = \left| \frac{\partial z'^i}{\partial z^k} \right| |\omega^{kl}| \left| \frac{\partial z'^j}{\partial z^l} \right| , \quad (4.149)$$

so the Jacobian must be 1 in absolute magnitude,

$$\left| \frac{\partial z'^i}{\partial z^j} \right| = \pm 1 . \quad (4.150)$$

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 (4.140), a function  $f(z^i)$  is an integral of motion if and only if its Poisson brackets with the Hamiltonian vanishes,

$$[f, H] = 0 . \quad (4.151)$$

As a particular example, the antisymmetry of the Poisson bracket implies that the Poisson bracket of the Hamiltonian with itself is zero,

$$[H, H] = 0 , \quad (4.152)$$

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,

$$[[f, g], H] = - [[g, H], f] - [[H, f], g] = 0 , \quad (4.153)$$

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 called a **Poisson algebra**.

**Concept question 4.6 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  $\pi_t$  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  $\lambda$  is set equal to the time  $t$ , which is therefore no longer an independent variable, and the generalized momentum  $\pi_t$ , 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 ingoing and outgoing 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.

# 5

---

## 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 attributed 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).
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.

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. At present, no gravitational waves have been definitely detected from anything. In the future, gravitational wave astronomy should eventually detect the merger of two black holes. If the waveforms of merging black holes are consistent with the predictions of general relativity, it will provide a far more stringent test of strong field general relativity than has been possible to date.
8. Although gravitational waves have yet to be detected directly, their existence has been 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.

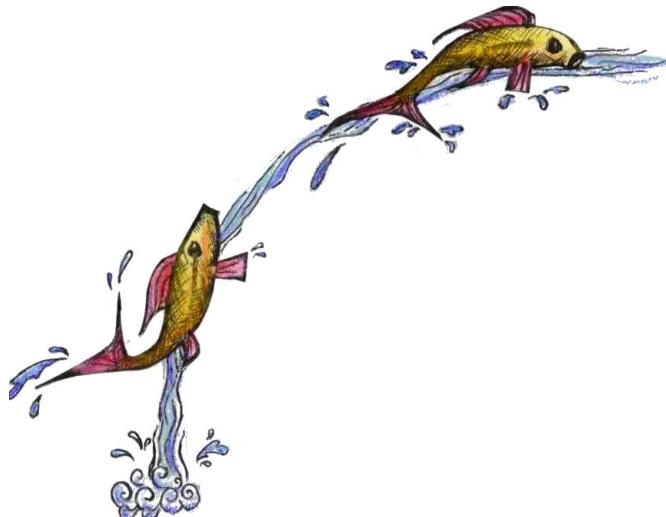


Figure 6.1 The fish upstream can make way against the current, but the fish downstream is swept to the bottom of the waterfall (Art by Wildrose Hamilton). This painting appeared on the cover of the June 2008 issue of the American Journal of Physics (Hamilton and Lisle, 2008). A similar depiction appeared 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 parametrized 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.

# 7

---

## 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 realization 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

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\sigma^2, \quad (7.1)$$

where  $d\sigma^2$  (this is the Landau & Lifshitz notation) is the metric of a unit 2-sphere,

$$d\sigma^2 = d\theta^2 + \sin^2\theta d\phi^2. \quad (7.2)$$

With units restored, the time-time component  $g_{tt}$  of the Schwarzschild metric is

$$g_{tt} = - \left(1 - \frac{2GM}{c^2r}\right). \quad (7.3)$$

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

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)(dr^2 + r^2 d\theta^2) \quad (7.4)$$

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

$$ds^2 \rightarrow -dt^2 + dr^2 + r^2 d\theta^2 \quad \text{as } r \rightarrow \infty . \quad (7.6)$$

**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

$$ds^2 = - \left( \frac{1 - M/2R}{1 + M/2R} \right)^2 dt^2 + (1 + M/2R)^4 (dR^2 + R^2 d\theta^2) . \quad (7.7)$$

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  $\phi$  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  $t_{\text{ff}}$ , 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, becomes singular at its centre,  $r = 0$ , 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  $\Gamma_{\lambda\mu\nu}^{\lambda}$ , equation (2.62). Of the 40 distinct connections, 9 should be non-vanishing. Compute all the components of the Riemann tensor  $R_{\kappa\lambda\mu\nu}$ , equation (2.109). There should be 6 distinct non-zero components. Compute all the components of the Ricci tensor  $R_{\kappa\mu}$ , equation (2.118). 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

$$ds^2 = A(t, r) dt^2 + B(t, r) dt dr + C(t, r) dr^2 + D(t, r) do^2 , \quad (7.14)$$

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  $r_s$ :

$$ds^2 = - \left(1 - \frac{r_s}{r}\right) dt^2 \quad \begin{cases} < 0 & \text{if } r > r_s , \\ = 0 & \text{if } r = r_s , \\ > 0 & \text{if } r < r_s . \end{cases} \quad (7.16)$$

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 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 (2.98) 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 infallor 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 artifact of looking at the tensor through the distorting lens of the coordinate basis vectors  $e_\mu$ , 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,  $\pm 1$ , and  $\pm 2$ . 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.

---

**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

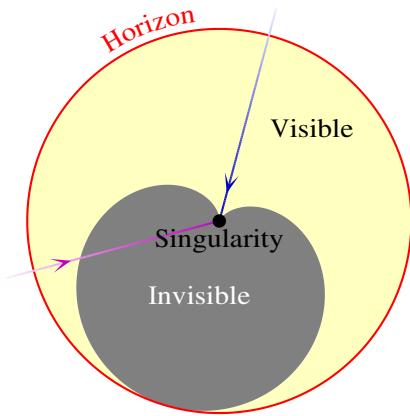


Figure 7.1 The light (yellow) shaded region shows the region visible to an infaller (blue) who falls radially to the singularity of a Schwarzschild black hole; the dark (grey) shaded region shows the region that remains invisible to the infaller. If another infaller (purple) falls along a different radial 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 points, the singularity cannot be a point.

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  $\phi$  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 Exercise 7.8 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 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 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.

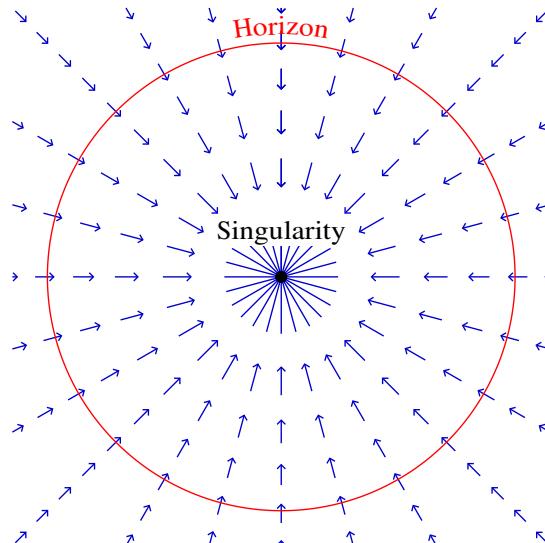


Figure 7.2 The Gullstrand-Painlevé metric for the Schwarzschild geometry encodes locally inertial frames (tetrads) that free-fall radially into the black hole at the Newtonian escape velocity  $\beta$ , equation (7.28). The infall velocity is less than the speed of light outside the horizon, equal to the speed of light at the horizon, and faster than light inside the horizon. The infall velocity tends to infinity at the central singularity.

The Gullstrand-Painlevé metric is

$$ds^2 = -dt_{\text{ff}}^2 + (dr - \beta dt_{\text{ff}})^2 + r^2 d\theta^2 . \quad (7.27)$$

Here  $\beta$  is the Newtonian escape velocity (with a minus sign because space is falling inward),

$$\boxed{\beta = -\left(\frac{2GM}{r}\right)^{1/2}}, \quad (7.28)$$

and  $t_{\text{ff}}$  is the proper time experienced by an object that free falls radially inward from zero velocity at infinity. The free fall time  $t_{\text{ff}}$  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  $e_{t_{\text{ff}}}$  in Gullstrand-Painlevé coordinates is not orthogonal to the radial axis  $e_r$ , but rather is tilted along the radial axis,  $e_{t_{\text{ff}}} \cdot e_r = g_{t_{\text{ff}}r} = -\beta$ .

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  $\beta$  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

$$\Delta(r) = 1 - \frac{2M}{r} . \quad (7.34)$$

- 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  $\tau$  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. **Sketch.** Sketch 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. **Photon sphere.** There is a radius where photons can orbit in circular orbits. What is the radius of this orbit? [Hint: Photons can be taken as the limit of a massive particle whose energy per unit mass  $E$  vastly exceeds its rest mass energy per unit mass, which is 1.]
11. **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  $\phi$  evolves according to  $d\phi/dt = u^\phi/u^t = L\Delta/(Er^2)$ .]

### Exercise 7.7 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

$$u^t = \frac{E}{\Delta}, \quad u^\phi = \frac{L}{r^2}, \quad u^r = \left[ E^2 - \left( 1 + \frac{L^2}{r^2} \Delta \right) \right]^{1/2}. \quad (7.41)$$

2. Letting  $x \equiv 1/r$ , show that

$$\phi = \int \frac{L \, dx}{[(E^2 - 1) + 2Mx - L^2x^2 + 2ML^2x^3]^{1/2}}. \quad (7.42)$$

[Hint: This is a straightforward application of equations (7.41). 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 (7.42) must take the form

$$\phi = \int \frac{dx}{[(x - x_+)(x_- - x)(a - 2Mx)]^{1/2}}, \quad (7.43)$$

where

$$a \equiv 1 - 2M(x_- + x_+). \quad (7.44)$$

[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 (7.42) and (7.43)].

4. By the transformation

$$x = x_+ + (x_- - x_+)y \quad (7.45)$$

bring the integral (7.43) to the form

$$\phi = \int \frac{dy}{[y(1-y)(q-py)]^{1/2}}, \quad (7.46)$$

where

$$p = 2M(x_- - x_+) , \quad q = 1 - 2M(x_- + 2x_+) . \quad (7.47)$$

5. Argue that the angle  $\phi$  integrated around a full period, from apoapsis at  $y = 0$  to periapsis at  $y = 1$  and back, is

$$\phi = \frac{4}{q^{1/2}} K(p/q) , \quad (7.48)$$

where  $K(m)$  is the complete elliptic integral of the first kind, one definition of which is

$$K(m) \equiv \frac{1}{2} \int_0^1 \frac{dy}{[y(1-y)(1-my)]^{1/2}} . \quad (7.49)$$

6. The Taylor series expansion of the complete elliptic integral is

$$K(m) = \frac{\pi}{2} \left( 1 + \frac{m}{4} + \dots \right) . \quad (7.50)$$

Argue that to linear order in mass  $M$ , the angle around a full period is

$$\phi = 2\pi + 3\pi M(x_- + x_+) . \quad (7.51)$$

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 7.8 A body cannot remain rigid as it approaches the Schwarzschild singularity.** You have already found from Exercise 7.6 that the azimuthal angle  $\phi$  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

$$\phi = \int \frac{L dr}{\sqrt{(E^2 - m^2)r^4 + L^2r^2\Delta}} . \quad (7.52)$$

1. Define  $J \equiv L/E$  to be the angular momentum per unit energy. Argue that for photons, which are massless,

$$\phi = \int \frac{J dr}{\sqrt{r^4 - J^2r^2\Delta}} . \quad (7.53)$$

2. Argue that inside the horizon ( $\Delta < 0$ ) the largest possible rate of change  $d\phi/dr$  of the azimuthal angle  $\phi$  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

$$r = r_s \sin^2(\phi/2) . \quad (7.54)$$

4. Parametrize the  $J = \infty$  null geodesic satisfying equation (7.54) by  $\{x, y\} \equiv \{r \cos \phi, r \sin \phi\}$ . Show that

$$\frac{dy}{dx} = \tan(3\phi/2) . \quad (7.55)$$

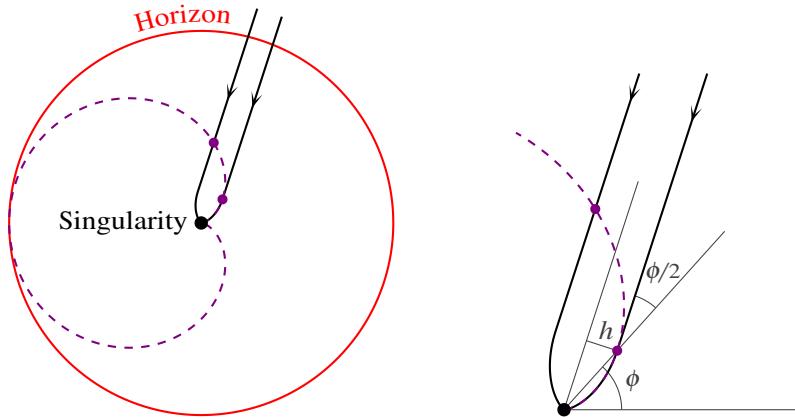


Figure 7.3 The arrowed lines, which are initially parallel, represent the worldtube of a body that remains as rigid as possible (having constant cross-sectional radius  $h$ ) as it falls to the singularity at the centre of a Schwarzschild black hole. (The blow-up at right shows some details.) The dashed (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.

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

$$h \leq r^{3/2}/r_s^{1/2} . \quad (7.56)$$

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  $\phi$  away is at the edge of their past light cone, along the  $J = \infty$  null geodesic.

### 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  $r, \theta, \phi$  of the 3D Schwarzschild spatial geometry, adjoin a fourth spatial coordinate  $w$ . The metric of 4D Euclidean space in the coordinates  $w, r, \theta, \phi$ , is

$$dl^2 = dw^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2. \quad (7.57)$$

The spatial Schwarzschild geometry is represented by a 3D surface embedded in the 4D Euclidean geometry,

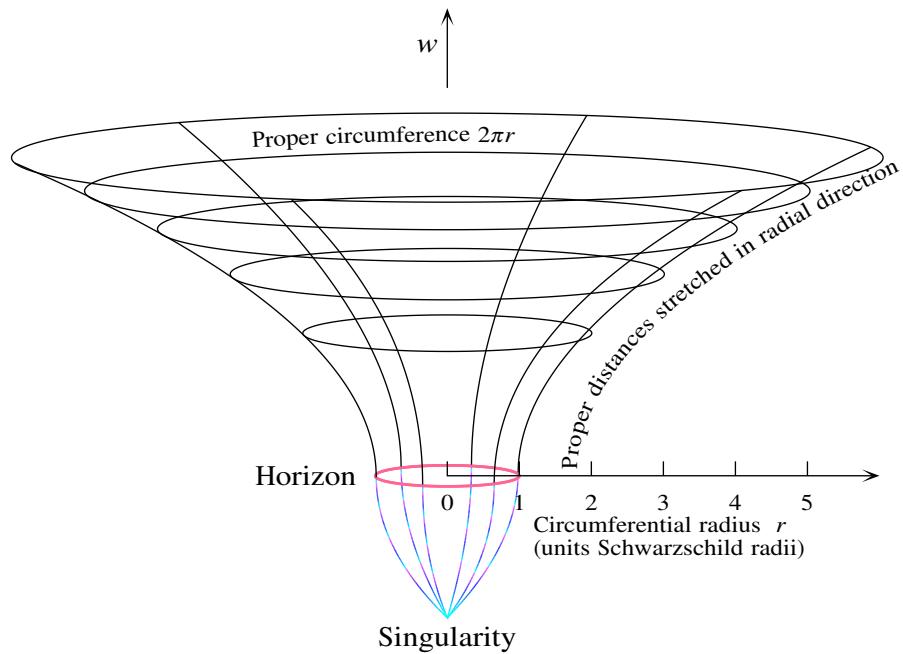


Figure 7.4 Embedding diagram of the Schwarzschild geometry. The 2-dimensional surface represents the 3-dimensional Schwarzschild geometry at a fixed instant of Schwarzschild time  $t$ . Each circle represents a sphere, of proper circumference  $2\pi r$ , as measured by observers at rest in the geometry. 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.

such that the proper distance  $dl$  in the radial direction satisfies equation (7.17), that is

$$dl^2 = \frac{dr^2}{1 - r_s/r} = dw^2 + dr^2 . \quad (7.58)$$

Equation (7.58) rearranges to

$$dw = \frac{dr}{\sqrt{r/r_s - 1}} , \quad (7.59)$$

which integrates to

$$w = 2\sqrt{\frac{r}{r_s} - 1} . \quad (7.60)$$

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,

$$dl^2 = -dw^2 + dr^2 + r^2 do^2 . \quad (7.61)$$

The embedded surface inside the horizon satisfies

$$w = -2\sqrt{1 - \frac{r}{r_s}} , \quad (7.62)$$

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 (7.60) and (7.62) 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

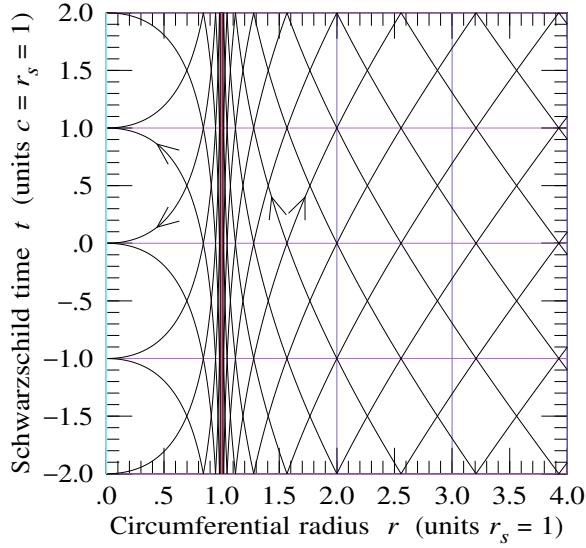


Figure 7.5 Spacetime diagram of the Schwarzschild geometry, in Schwarzschild coordinates. The horizontal axis is the circumferential radius  $r$ , while the vertical axis is Schwarzschild time  $t$ . The horizon (pink) is at one Schwarzschild radius,  $r = r_s$ . The singularity (cyan) is at zero radius,  $r = 0$ . 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 artifact of the Schwarzschild coordinate system.

null geodesics are not at  $45^\circ$ , 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

$$ds^2 = - \left(1 - \frac{r_s}{r}\right) dt^2 + \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 = 0 . \quad (7.63)$$

Radial null geodesics thus follow

$$\frac{dr}{dt} = \pm \left(1 - \frac{r_s}{r}\right) , \quad (7.64)$$

in which the  $\pm$  sign is  $+$  for outgoing,  $-$  for infalling rays. Equation (7.64) 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 artifact 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.

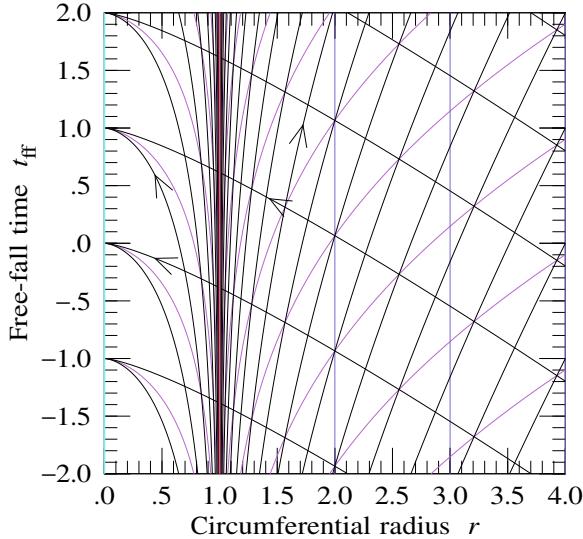


Figure 7.6 Gullstrand-Painlevé, or free-fall, spacetime diagram, in units  $r_s = c = 1$ . In this spacetime diagram the time coordinate is the Gullstrand-Painlevé time  $t_{\text{ff}}$ , which is the proper time of observers who free-fall radially from zero velocity at infinity. The radial coordinate  $r$  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.

### 7.15 Gullstrand-Painlevé spacetime diagram

Figure 7.6 shows a spacetime diagram of the Schwarzschild geometry in Gullstrand-Painlevé (1921) coordinates  $t_{\text{ff}}$  and  $r$  in place of Schwarzschild coordinates  $t$  and  $r$ . As the spacetime diagram shows, in Gullstrand-Painlevé 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 (7.64). Equation (7.64) integrates to

$$t = \pm (r + r_s \ln|r - r_s|) , \quad (7.65)$$

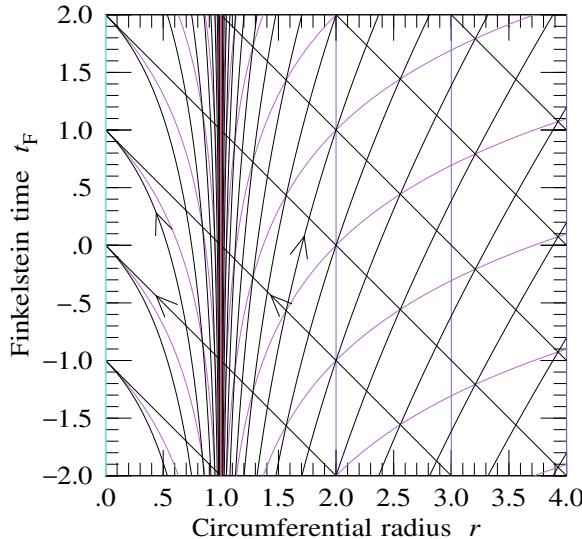


Figure 7.7 Finkelstein spacetime diagram, in units  $r_s = c = 1$ . Here the time coordinate is taken to be the Finkelstein time coordinate  $t_F$ , equation (7.66). The Finkelstein time coordinate  $t_F$  is constructed so that radially infalling light rays are at  $45^\circ$ .

which shows that Schwarzschild time  $t$  approaches  $\pm\infty$  logarithmically as null rays approach the horizon. Finkelstein defined his time coordinate  $t_F$  by

$$t_F \equiv t + r_s \ln |r - r_s| , \quad (7.66)$$

which has the property that infalling null rays follow

$$t_F + r = \text{constant} . \quad (7.67)$$

In other words, on a spacetime diagram in Finkelstein coordinates, Figure 7.7, radially infalling light rays move at  $45^\circ$ , 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  $45^\circ$  in a spacetime diagram, it was natural to look for coordinates in which outgoing as well as infalling light rays are at  $45^\circ$ . 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

$$r^* \equiv \int \frac{dr}{1 - 2M/r} = r + 2M \ln |r - 2M| . \quad (7.68)$$

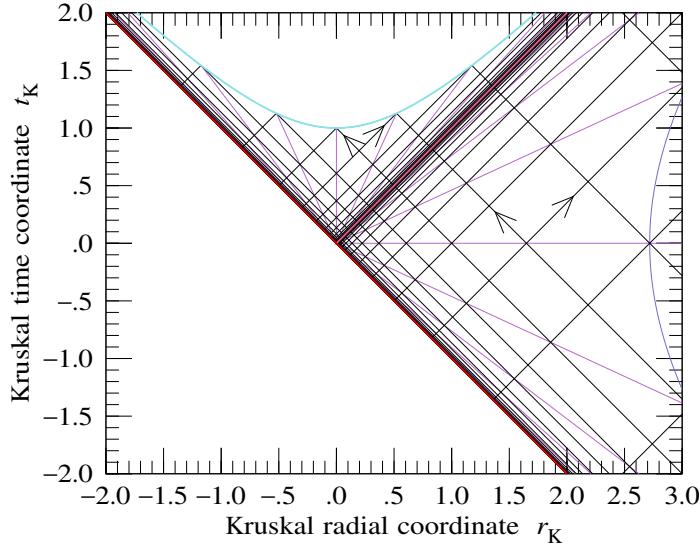


Figure 7.8 Kruskal-Szekeres spacetime diagram, in units  $r_s = c = 1$ . Kruskal-Szekeres coordinates are arranged such that not only infalling, but also outgoing null rays move at  $45^\circ$  on the spacetime diagram. The Kruskal-Szekeres spacetime diagram reveals the causal structure of the Schwarzschild 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) 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, while lines of constant circumferential radius  $r$  (blue) are drawn spaced uniformly by  $1/4$ .

Then radially infalling and outgoing null rays follow

$$\begin{aligned} r^* + t &= \text{constant} && \text{infalling ,} \\ r^* - t &= \text{constant} && \text{outgoing .} \end{aligned} \quad (7.69)$$

In a spacetime diagram in coordinates  $t$  and  $r^*$ , infalling and outgoing light rays are indeed at  $45^\circ$ . Unfortunately the metric in these coordinates is still singular at the horizon  $r = 2M$ :

$$ds^2 = \left(1 - \frac{2M}{r}\right) (-dt^2 + dr^{*2}) + r^2 d\theta^2 . \quad (7.70)$$

The singularity at the horizon can be eliminated by the following transformation into Kruskal-Szekeres

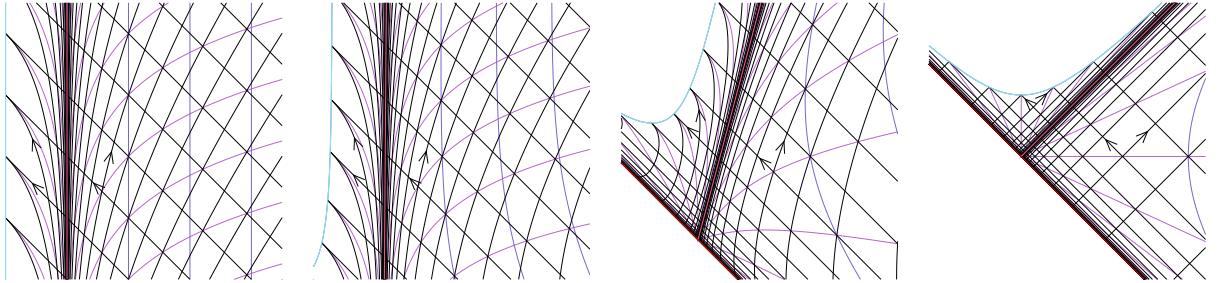


Figure 7.9 From left to right, the Finkelstein spacetime diagram, Figure 7.7, morphs into the Kruskal-Szekeres spacetime diagram, Figure 7.8. The morph illustrates how the antihorizon, or past horizon (red), emerges from the depths of  $t = -\infty$ . Like the horizon, the antihorizon is a null surface, thus appearing at  $45^\circ$  in the Kruskal-Szekeres spacetime diagram.

coordinates  $t_K$  and  $r_K$ :

$$\begin{aligned} r_K + t_K &= 2M \exp\left(\frac{r^* + t}{4M}\right) , \\ r_K - t_K &= \pm 2M \exp\left(\frac{r^* - t}{4M}\right) , \end{aligned} \quad (7.71)$$

where the  $\pm$  sign in the last equation is  $+$  outside the horizon,  $-$  inside the horizon. The Kruskal-Szekeres metric is

$$ds^2 = \frac{8M}{r} e^{-r/2M} (-dt_K^2 + dr_K^2) + r^2 d\Omega^2 , \quad (7.72)$$

which is non-singular at the horizon. The Schwarzschild radial coordinate  $r$ , which appears in the factors  $(8M/r)e^{-r/2M}$  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  $r = 2M$ , 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

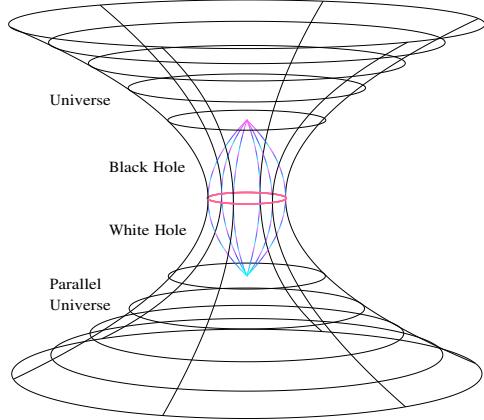


Figure 7.10 Embedding diagram of the analytically extended Schwarzschild geometry. The analytically extended geometry is constructed by gluing together two copies of the Schwarzschild geometry along the antihorizon. The extended geometry contains a Universe, a Parallel Universe, a Black Hole, and a White Hole.

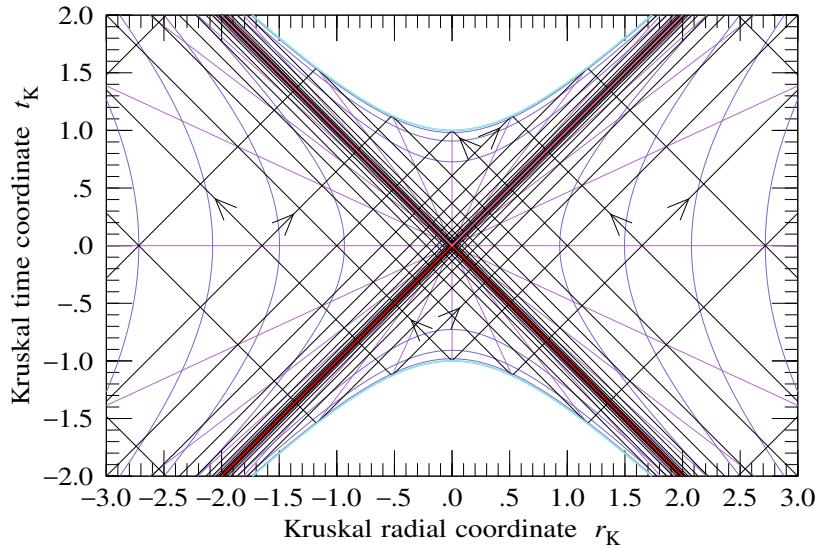


Figure 7.11 Analytically extended Kruskal-Szekeres spacetime diagram, in units  $r_s = c = 1$ . The analytically extended horizon and antihorizon (crossing pink/red lines at  $45^\circ$ ) divide the spacetime into 4 regions, a Universe region at right, a Black Hole region bounded by the singularity at 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 artifact 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.

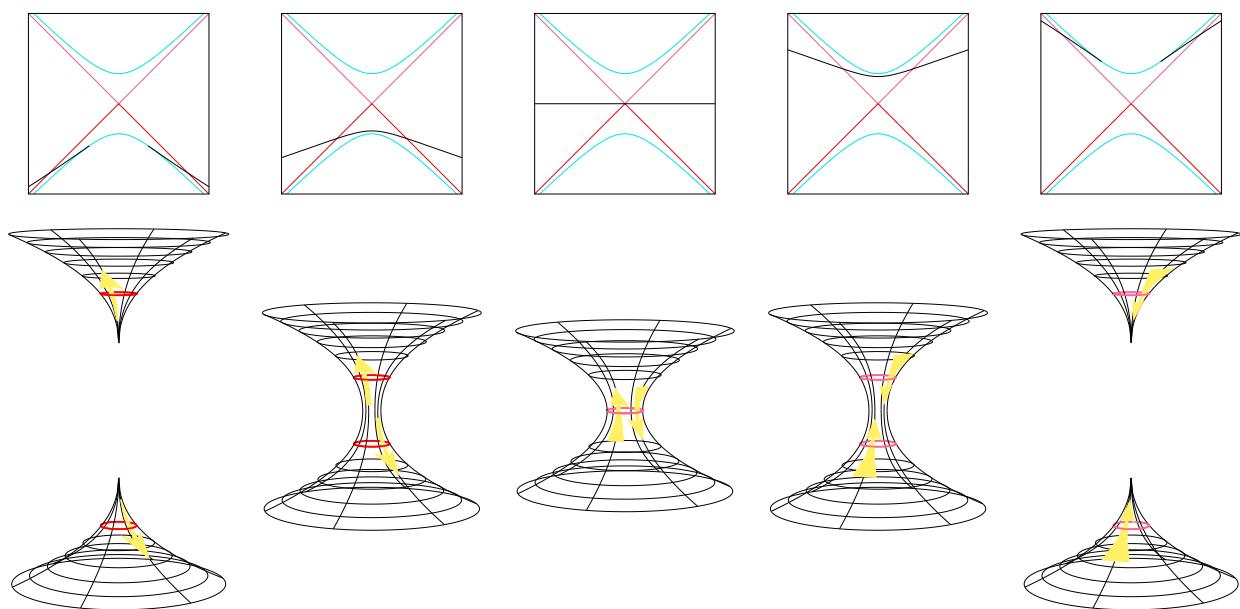


Figure 7.12 Sequence of embedding diagrams of spatial slices of the analytically extended Schwarzschild geometry, progressing in time from left to right. Two white holes merge, form an Einstein-Rosen bridge, then fall apart into two black holes. The wormhole formed by the Einstein-Rosen bridge is non-traversable. 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.

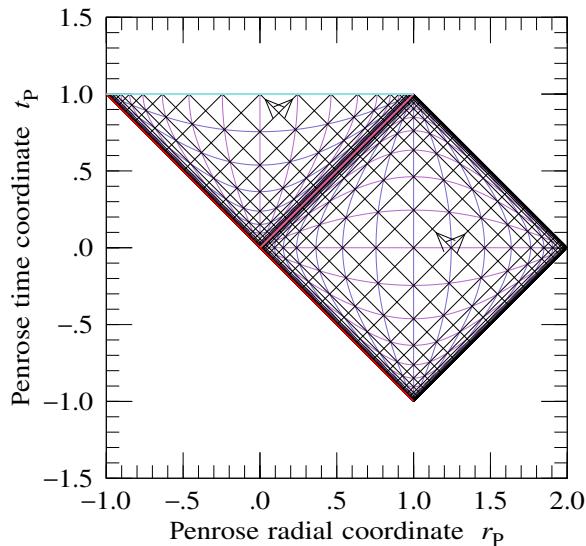


Figure 7.13 Penrose spacetime diagram, in units  $r_s = c = 1$ . The Penrose coordinates  $t_P$  and  $r_P$  here are defined by equations (7.73) and (7.74). Lines of constant Schwarzschild time  $t$  (violet), and infalling and outgoing null lines (black) are spaced uniformly at intervals of 1 (units  $r_s = c = 1$ ), while lines of constant circumferential radius  $r$  (blue) are spaced uniformly in the tortoise coordinate  $r^*$ , equation (7.68), so that the intersections of  $t$  and  $r$  lines are also intersections of infalling and outgoing null lines.

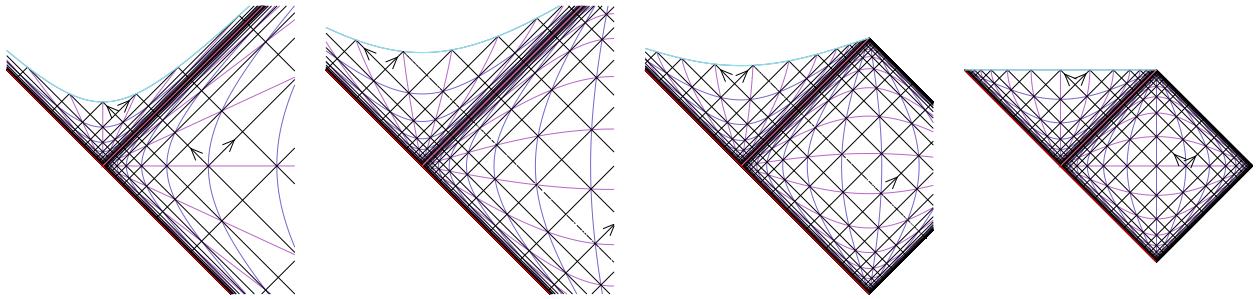


Figure 7.14 From left to right, the Kruskal-Szekeres spacetime diagram, Figure 7.8, morphs into the Penrose spacetime diagram, Figure 7.13.

## 7.20 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  $45^\circ$ , Penrose wanted to bring regions at infinity (in time or

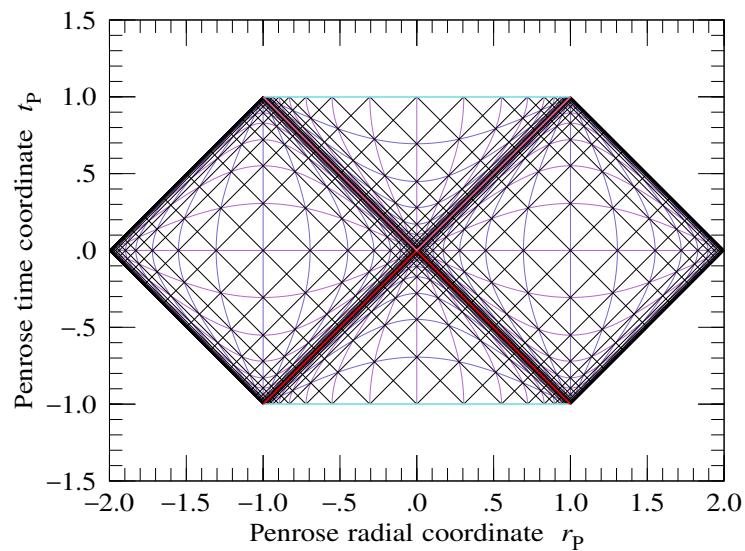


Figure 7.15 Penrose spacetime diagram of the analytically extended Schwarzschild geometry. This is the analytically extended version of Figure 7.13.

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

$$r_P \pm t_P = f(r_K \pm t_K) \quad (7.73)$$

for which  $f(z)$  is finite as  $z \rightarrow \pm\infty$ . The transformation (7.73) brings spatial and temporal infinity to finite values of the coordinates, while keeping infalling and outgoing light rays at  $45^\circ$  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

$$f(z) = \frac{2}{\pi} \tan^{-1} z . \quad (7.74)$$

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.

## 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  $45^\circ$  edges of the Penrose diagram at infinite 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

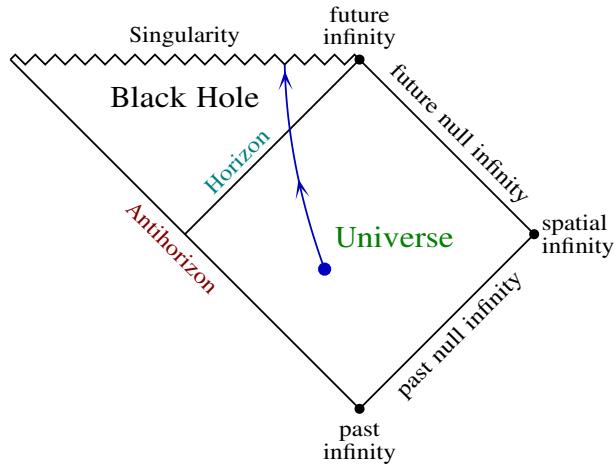


Figure 7.16 Penrose diagram of the Schwarzschild geometry, labelled with the Universe and Black Hole regions, and their various boundaries. The (blue) line at less than  $45^\circ$  from vertical is a possible trajectory of a person who falls through the horizon from the Universe into the Black Hole. Once inside the horizon, the infaller cannot avoid the Singularity.

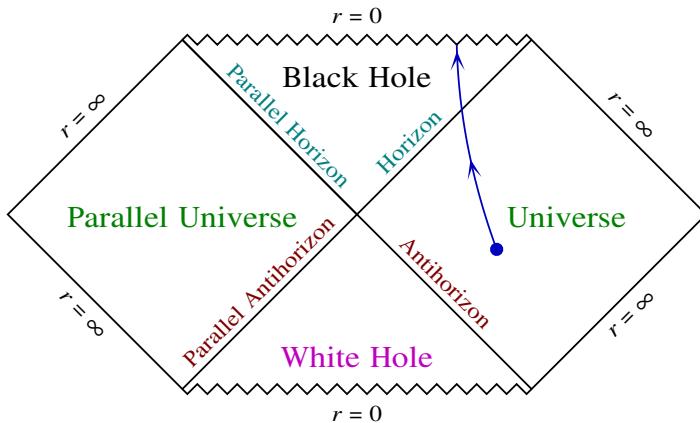


Figure 7.17 Penrose diagram of the analytically extended Schwarzschild geometry.

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 7.9 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  $\beta$  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 collapsed with the star would find themself 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, barreling 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

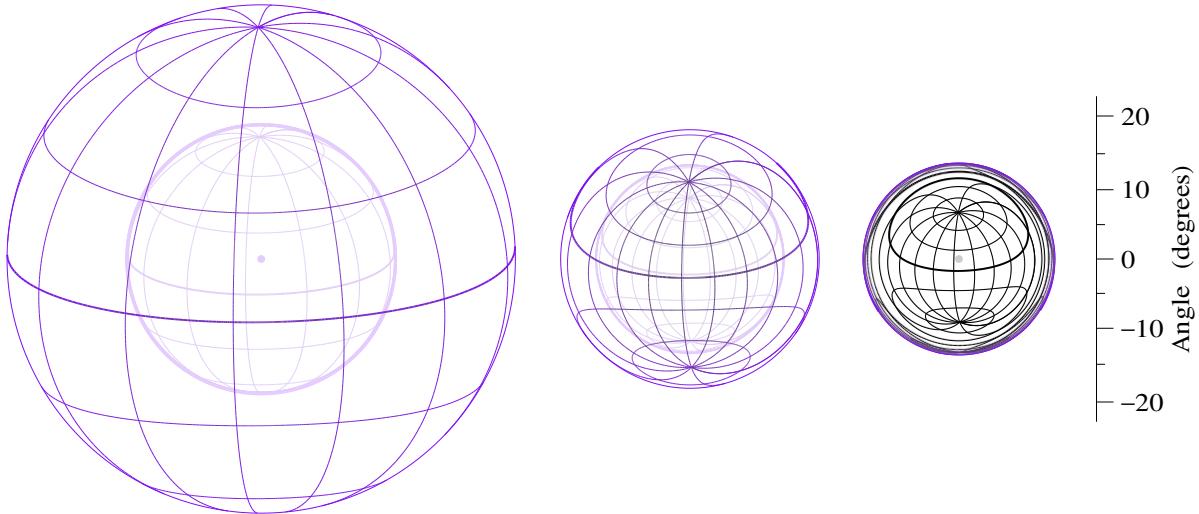


Figure 7.18 Three frames in the collapse of a uniform density, pressureless, spherical star from zero velocity at infinity (Oppenheimer and Snyder, 1939), as seen by an outside observer at rest at a radius of 10 Schwarzschild radii. The frames are spaced by 10 units of Schwarzschild time ( $c = r_s = 1$ ). 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.

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  $45^\circ$ .

## 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,

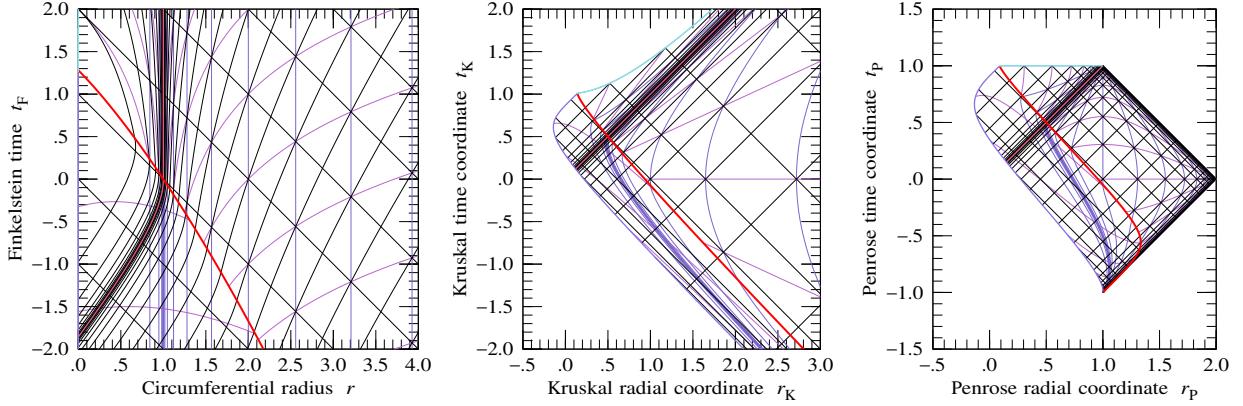


Figure 7.19 Finkelstein, Kruskal-Szekeres, and Penrose spacetime diagrams of the Oppenheimer-Snyder of a pressureless, spherical star. The thick (red) line is the surface of the collapsing star. The geometry outside the surface of the star is Schwarzschild, and the spacetime diagrams there look like 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 (7.68), so before collapse appear bunched around the radius  $r = r_s$  that after collapse becomes the horizon radius. The thick (pink) line at  $45^\circ$  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.

an apparent horizon is a place where radially moving null geodesics remain at rest in radius  $r$ ,

$$\frac{dr}{d\lambda} = 0 . \quad (7.75)$$

## 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 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

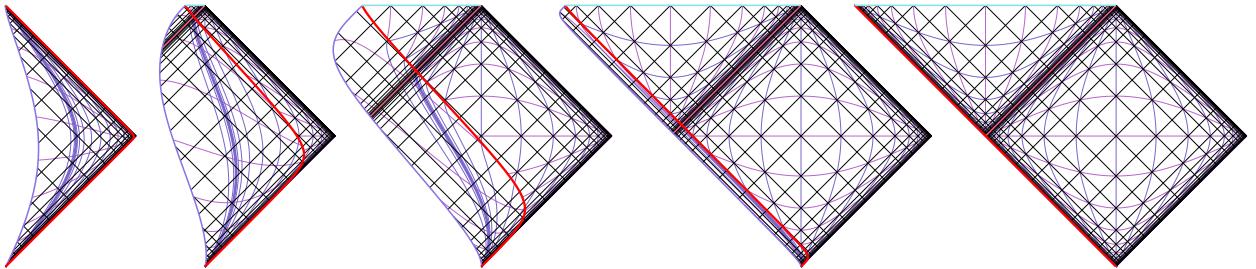


Figure 7.20 Sequence of Penrose diagrams illustrating the Oppenheimer-Snyder collapse of a pressureless, spherical star to a Schwarzschild black hole, progressing in time of collapse from left to right. On the left, the collapse is to the future of an observer at the centre of the diagram; on the right, 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 (7.73) and (7.74).

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 (7.73) with the function  $f(z)$  given by equation (7.74). 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 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

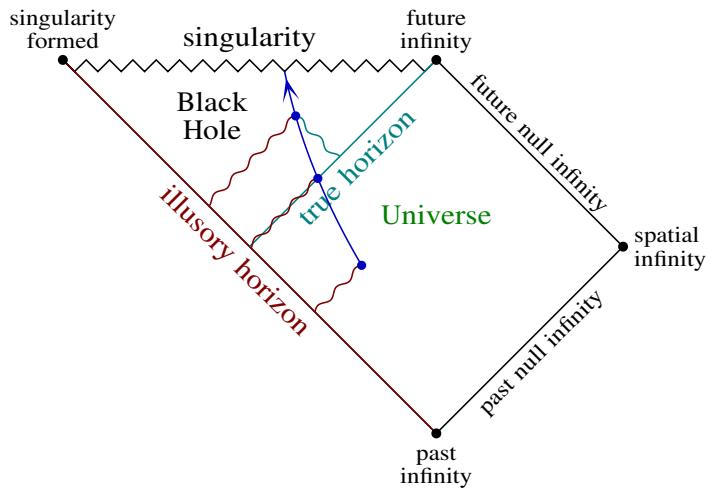


Figure 7.21 Penrose diagram of a collapsed spherical star at late times. The Penrose diagram looks essentially identical to the Penrose diagram 7.16 of the Schwarzschild geometry, except that the antihorizon is replaced by the illusory horizon. The wiggly lines show the paths of outgoing light rays from 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.

spacetime diagram. An infaller inside the horizon must of course follow a worldline at less than  $45^\circ$  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.

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 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 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. The shell collapses from zero velocity at infinity. As usual in this book, the frames are accurately ray-traced.

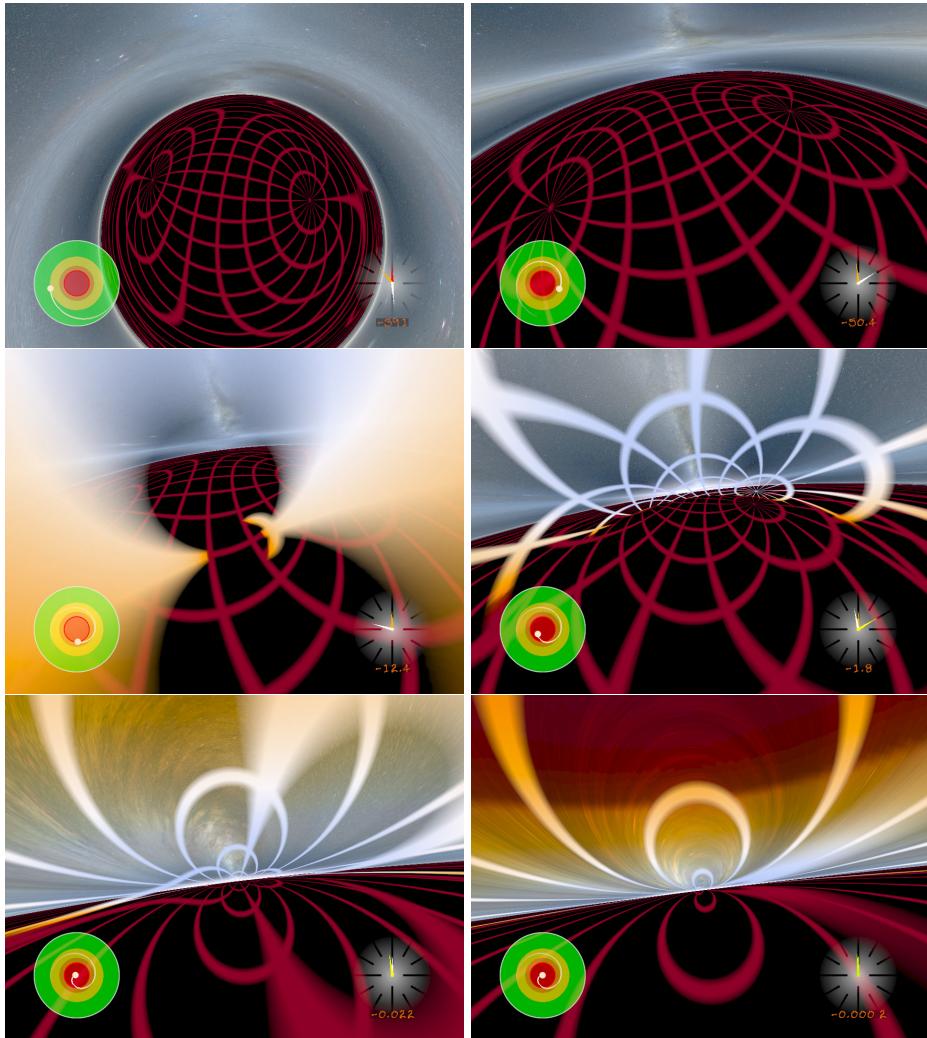


Figure 7.22 Six frames from a visualization of the view seen by an observer who free-falls into a Schwarzschild black hole. The infaller is on a geodesic with energy per unit mass  $E = 1$ , and angular momentum per unit mass  $L = 1.96 r_s$ . From left to right and top to bottom, the observer is at 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).

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 7.10 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. 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

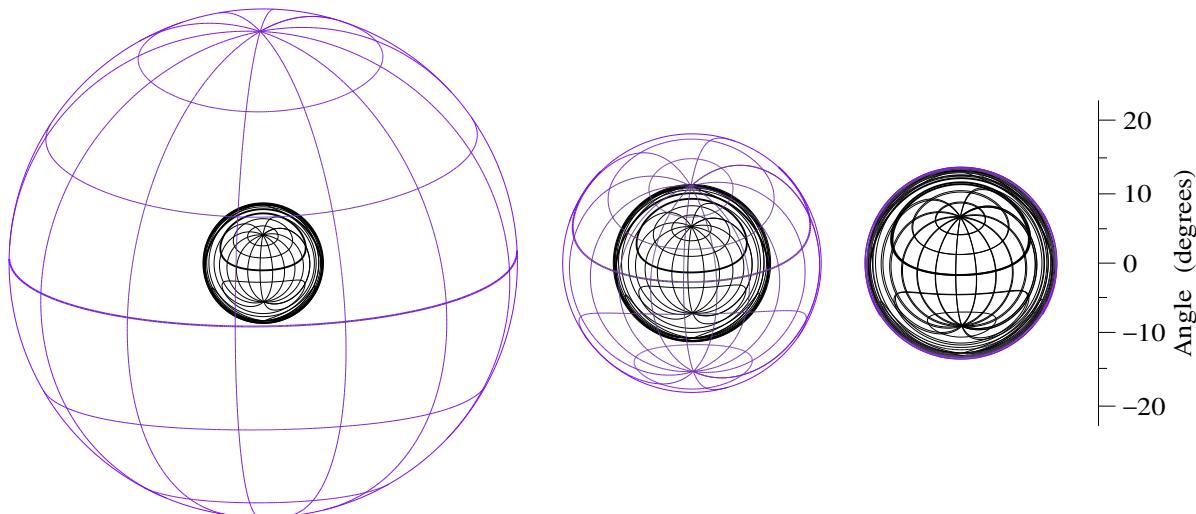


Figure 7.23 Three frames in the collapse of a thin spherical shell of matter on to a pre-existing Schwarzschild black hole, as seen by an outside observer at rest at a radius of 10 Schwarzschild radii (Schwarzschild radius of the final black hole). The frames are spaced by 10 units of Schwarzschild time ( $c = r_s = 1$ ). 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.

the shell). The apparent horizon before collapse follows an outgoing null ( $45^\circ$ ) 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.

### 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

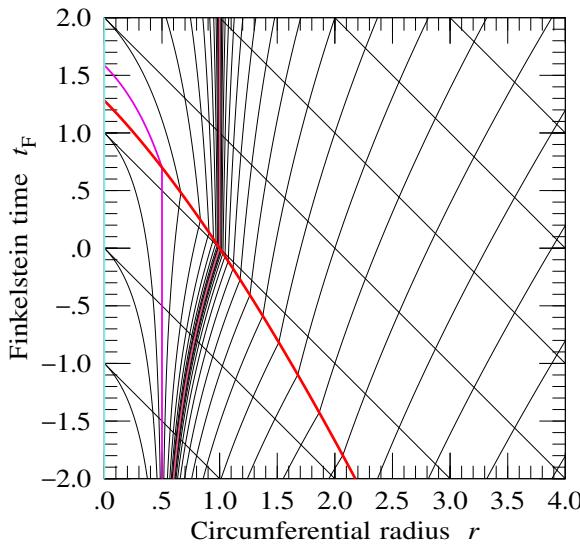


Figure 7.24 Finkelstein spacetime diagram of a thin spherical shell of matter collapsing on to a pre-existing Schwarzschild black hole, in units  $r_s$  of the Schwarzschild radius of the final black hole. The mass of the (red) shell equals that of the pre-existing black hole, so the black hole doubles 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.

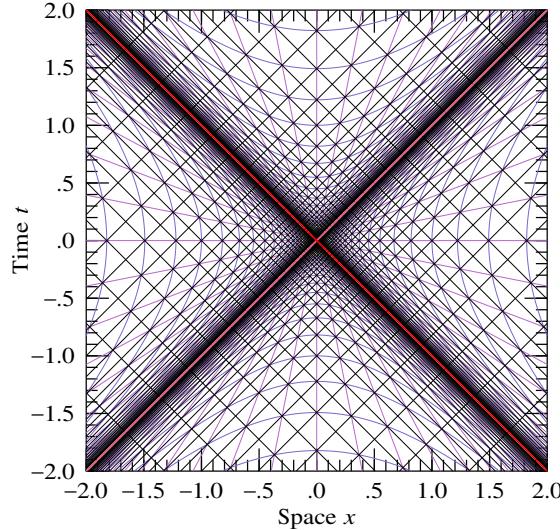


Figure 7.25 Rindler diagram, which is a Minkowski spacetime diagram showing lines of constant Rindler coordinates  $\alpha$  and  $l$ , equations (7.76) and (7.78). The Rindler lines are uniformly spaced by 0.2 in  $\alpha$  and  $\ln l$ . The spacetime diagram resembles that of the analytically extended 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.

space

$$\{t, x\} = l \{\sinh \alpha, \cosh \alpha\} \quad (7.76)$$

with fixed  $l$  and varying  $\alpha$ . 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 2.8,

$$ds^2 = -l^2 d\alpha^2 + dl^2 + dy^2 + dz^2. \quad (7.77)$$

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.

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  $\alpha$  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  $45^\circ$  passing through the origin in the Rindler diagram 7.25. The Rindler metric (7.77) 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

$$\{t, x\} = l \{\cosh \alpha, \sinh \alpha\} , \quad (7.78)$$

where the timelike coordinate  $l$  is positive in the upper quadrant, negative in the lower quadrant.

The null ( $45^\circ$ ) 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 (7.76) 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  $\alpha$  and  $l$  are analogues of the Schwarzschild coordinates  $t$  and  $r$ . The Schwarzschild and Rindler time coordinates  $t$  and  $\alpha$  are both Killing coordinates, §7.32. Lines of constant Schwarzschild and Rindler time  $t$  and  $\alpha$  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.

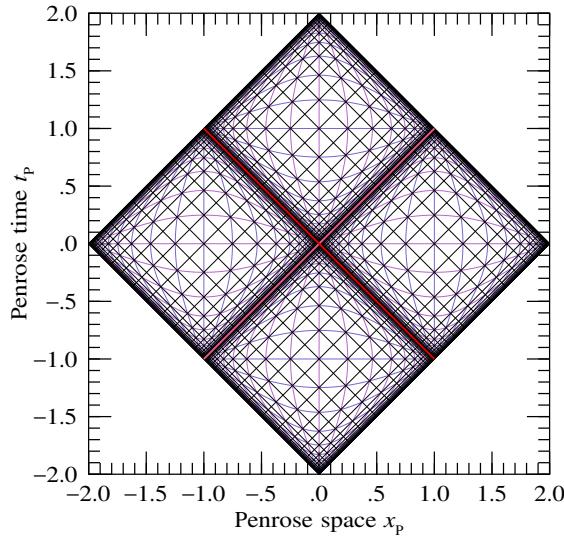


Figure 7.26 Penrose diagram of Rindler space. This is the Penrose diagram of Minkowski space corresponding to the Rindler diagram 7.25. The Penrose coordinates  $t_p$  and  $x_p$  are related to Minkowski coordinates  $t$  and  $x$  by equations (7.79). The Rindler lines are uniformly spaced by 0.4 in  $\alpha$  and  $\ln l$ . The Penrose diagram resembles that of the analytically extended Schwarzschild geometry, Figure 7.15, but without singularities.

### 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  $\alpha$  and  $l$ . Penrose time and space coordinates  $t_P$  and  $x_P$  can be defined by any conformal transformation

$$t_P \pm x_P \equiv f(t \pm x) \quad (7.79)$$

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

$$f(z) \equiv \text{sign}(z) + \frac{2}{\pi} \tan^{-1} \left( \frac{z - z^{-1}}{2} \right). \quad (7.80)$$

The choice (7.80) is inspired by the form (10.162) 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.

**Concept question 7.11 Spherical Rindler space.** The Rinder line-element (7.77) 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

$$ds^2 = -r_R^2 d\alpha^2 + dr_R^2 + r^2 do^2, \quad (7.81)$$

where all Rindler observers accelerate in the radial direction with  $\{t, r\} = r_R \{\sinh \alpha, \cosh \alpha\}$ ? **Answer.** The spherical Rindler line-element (7.81) 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 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.

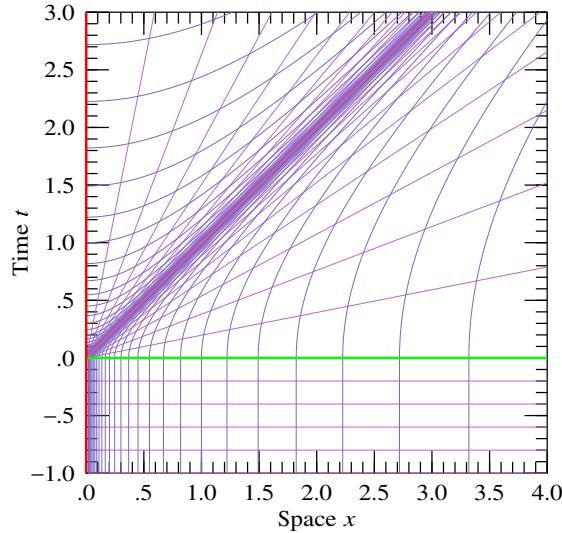


Figure 7.27 Minkowski spacetime diagram, showing worldlines of observers who start at rest, then begin accelerating uniformly, as Rindler observers, at  $t = 0$ . At  $t \leq 0$ , the lines are lines of constant Minkowski time and space  $t$  and  $x$ , while at  $t \geq 0$ , the lines are lines of constant Rindler time and space  $\alpha$  and  $l$ , equations (7.76) and (7.78). The Rindler lines are uniformly spaced by 0.2 in  $\alpha$  and  $\ln l$ . The null line starting at the origin  $\{t, x\} = \{0, 0\}$  extending upward at  $45^\circ$  from vertical is a future horizon for the Rindler observers.

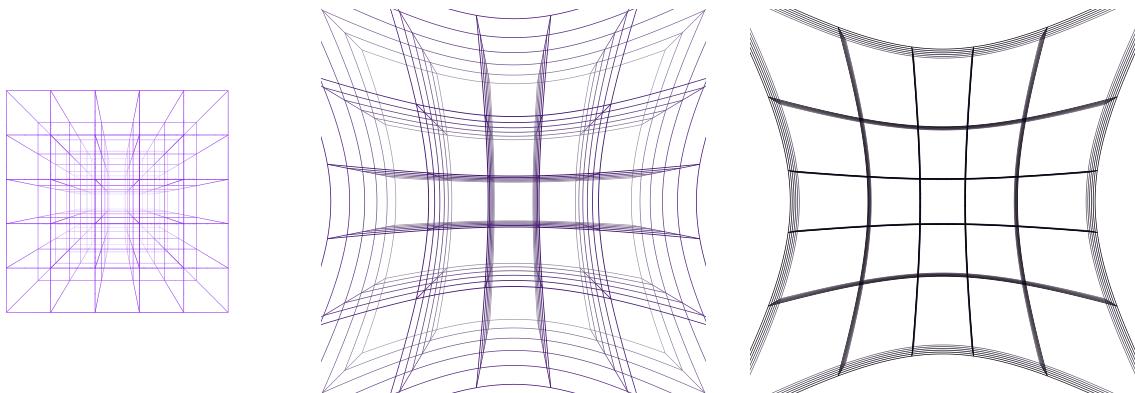


Figure 7.28 Three frames in the appearance of Minkowski space as seen by a uniformly accelerating observer, a Rindler observer. Minkowski space is represented by a unit box at rest, centred at the origin. The box is drawn as a  $5 \times 5 \times 5$  lattice. The Rindler observer starts at rest at unit distance from the origin, and watches rearward while accelerating at unit acceleration away from the box. The field of view is  $120^\circ$  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.

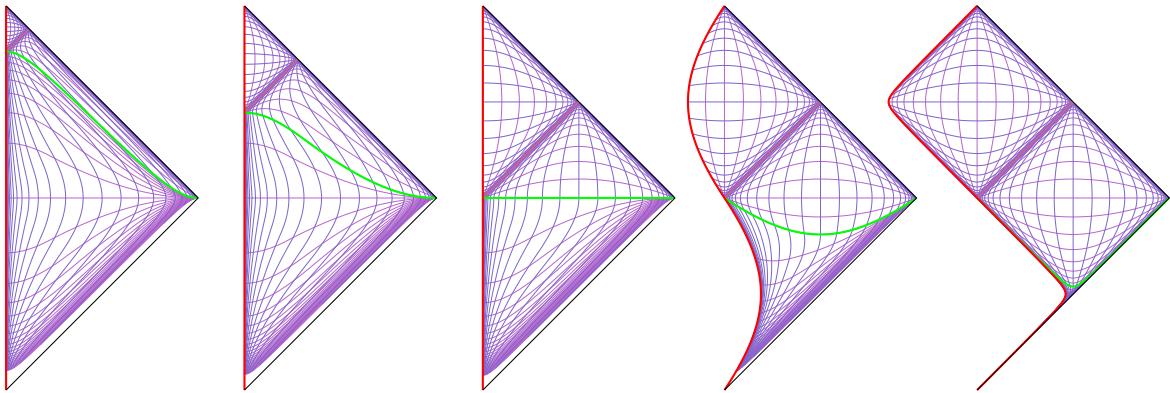


Figure 7.29 Sequence of Penrose diagrams of the Minkowski space shown in Figure 7.27, progressing in time from left to right. The left edge of each diagram is the surface at  $x = 0$ . The diagrams at left are in the frames of observers who are at rest relative to each other. In the middle diagram, the 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  $\alpha$  and  $\ln l$ . This sequence of Penrose diagrams resembles that of the Oppenheimer-Snyder collapse of a star shown in Figure 7.20.

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 an 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 7.12 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 calculated, not sketched, with Penrose coordinates given by equations (7.79). 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  $45^\circ$  off the left edge of the diagram at  $x = 0$ . Whatever the case, as seen in Exercise 7.12, light emitted from  $x < 0$  appears to a Rindler observer asymptotically to dim, redshift, and freeze at the observer's illusory horizon.

---

**Exercise 7.12 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  $\alpha$ , 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

$$l = \frac{1}{2}(y^2 + z^2 - 1). \quad (7.82)$$

The Rindler observer sees their illusory horizon as a parabola with themself at the focus, the origin.

2. **Redshift.** Show further that the Rindler observer sees points on the illusory horizon redshifting exponentially, at rate  $e^\alpha$ .

#### 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

$$l = \gamma(x - x_0) + \gamma v \sqrt{(x - x_0)^2 + y^2 + z^2}, \quad (7.83)$$

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 (7.83) factors to

$$l = \gamma(x - x_0) \left[ 1 - v \sqrt{1 + (y^2 + z^2)/(x - x_0)^2} \right], \quad (7.84)$$

which rearranges to

$$\begin{aligned} l &= \frac{\gamma(x - x_0)[1 - v^2 - v^2(y^2 + z^2)/(x - x_0)^2]}{1 + v^2\sqrt{1 + (y^2 + z^2)/(x - x_0)^2}} \\ &= \frac{1}{1 + v^2\sqrt{1 + (y^2 + z^2)/(x - x_0)^2}} \left[ \frac{x - x_0}{\gamma} - \frac{v^2(y^2 + z^2)\gamma}{x - x_0} \right]. \end{aligned} \quad (7.85)$$

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 (7.85) reduces to

$$l \approx -\frac{1}{2} + \frac{1}{2}(y^2 + z^2), \quad (7.86)$$

yielding equation (7.82) 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,

$$\frac{E_{\text{em}}}{E_{\text{obs}}} = \frac{\sqrt{(x - x_0)^2 + y^2 + z^2}}{\sqrt{l^2 + y^2 + z^2}}. \quad (7.87)$$

A point  $\{l, y, z\}$  on the Rindler observer's illusory horizon appears fixed to the observer,  $l$  satisfying equation (7.86). The only quantity on the right hand side of equation (7.87) that varies with the Rindler observer's time  $\alpha$  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

$$\frac{E_{\text{em}}}{E_{\text{obs}}} \propto x_0 \propto e^\alpha. \quad (7.88)$$

The redshift factor of a point on the Rindler observer's illusory horizon thus increases exponentially with Rindler time  $\alpha$ .

**Exercise 7.13 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  $y$ - $z$  plane orthogonal to the Rindler observer's boost plane  $t$ - $x$ . For a Rindler observer who starts accelerating at a finite time, the illusory horizon after  $\alpha$  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  $\sim e^{2\alpha}$ .

---

## 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  $\xi^{\mu}$  at each point of the spacetime is then defined by

$$\xi^{\mu} \frac{\partial}{\partial x^{\mu}} = \frac{\partial}{\partial t}, \quad (7.89)$$

so that in Schwarzschild coordinates  $\{t, r, \theta, \phi\}$

$$\xi^{\mu} = \{1, 0, 0, 0\}. \quad (7.90)$$

In coordinate-independent notation, the Killing vector is

$$\xi = e_{\mu} \xi^{\mu} = e_t. \quad (7.91)$$

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 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  $e_t$  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

$$\xi_{\mu} \xi^{\mu} = e_t \cdot e_t = g_{tt} = -\left(1 - \frac{2M}{r}\right). \quad (7.92)$$

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  $\phi$ , implying that  $\phi$  is a Killing coordinate. The associated Killing vector at each point of the spacetime is

$$e_{\phi} \quad (7.93)$$

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:

$$iL_x = y \frac{\partial}{\partial z} - z \frac{\partial}{\partial y} = -\sin \phi \frac{\partial}{\partial \theta} - \cot \theta \cos \phi \frac{\partial}{\partial \phi}, \quad (7.94a)$$

$$iL_y = z \frac{\partial}{\partial x} - x \frac{\partial}{\partial z} = \cos \phi \frac{\partial}{\partial \theta} - \cot \theta \sin \phi \frac{\partial}{\partial \phi}, \quad (7.94b)$$

$$iL_z = x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} = \frac{\partial}{\partial \phi}. \quad (7.94c)$$

The 3 rotational Killing vectors are correspondingly:

$$\text{rotation about } x\text{-axis: } -\sin \phi e_\theta - \cot \theta \cos \phi e_\phi, \quad (7.95a)$$

$$\text{rotation about } y\text{-axis: } \cos \phi e_\theta - \cot \theta \sin \phi e_\phi, \quad (7.95b)$$

$$\text{rotation about } z\text{-axis: } e_\phi. \quad (7.95c)$$

The 3 Killing vectors span the 2-dimensional surface of the unit sphere, and are therefore not linearly independent. Specifically, they satisfy

$$xL_x + yL_y + zL_z = 0. \quad (7.96)$$

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,  $iL_x g_{\phi\phi} = 2r^2 \cos \phi \sin \theta \cos \theta$  does not vanish. This example shows that a more powerful

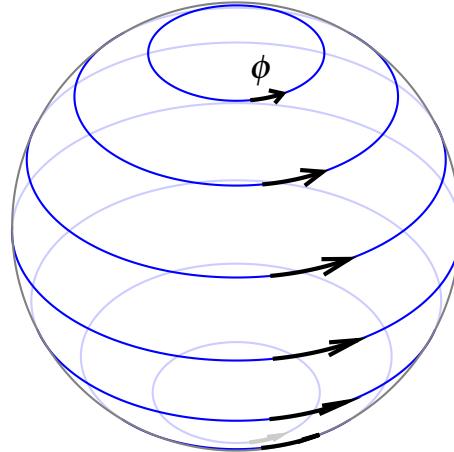


Figure 7.30 The Killing vector field associated with rotation of a 2-sphere about an axis.

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

$$(-\sin\phi \mathbf{e}_\theta - \cot\theta \cos\phi \mathbf{e}_\phi)^2 + (\cos\phi \mathbf{e}_\theta - \cot\theta \sin\phi \mathbf{e}_\phi)^2 + \mathbf{e}_\phi^2 = g_{\theta\theta} + (\cot^2\theta + 1)g_{\phi\phi} = -2r^2 . \quad (7.97)$$

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  $\phi$ . Then the covariant momentum  $p_\phi$  of a particle along a geodesic is a constant of motion, equation (4.50),

$$p_\phi = \text{constant} . \quad (7.98)$$

Equivalently

$$\xi^\nu p_\nu = \text{constant} , \quad (7.99)$$

where  $\xi^\nu$  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  $\xi^\nu$  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

$$\frac{d\xi^\nu p_\nu}{d\lambda} = 0 . \quad (7.100)$$

But this is equivalent to

$$0 = p^\mu \mathring{D}_\mu (\xi^\nu p_\nu) = p^\mu p^\nu \mathring{D}_\mu \xi_\nu = \tfrac{1}{2} p^\mu p^\nu (\mathring{D}_\mu \xi_\nu + \mathring{D}_\nu \xi_\mu) , \quad (7.101)$$

the  $\mathring{\phantom{D}}$  atop  $\mathring{D}_\mu$  serving as a reminder that this is the torsion-free covariant derivative, §2.12. The second equality of equations (7.101) 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 (7.101) to be true for all geodesics is that

$$\boxed{\mathring{D}_\mu \xi_\nu + \mathring{D}_\nu \xi_\mu = 0} , \quad (7.102)$$

which is **Killing's equation**. This equation is the desired necessary and sufficient condition for  $\xi^\nu$  to be a Killing vector. It is a generally covariant equation, valid in any coordinate system. Equation (7.102) can

also be written as the statement that the Lie derivative of the metric, equation (7.132), along the Killing direction  $\xi^\nu$  vanishes,

$$\mathcal{L}_\xi g_{\mu\nu} = 0 . \quad (7.103)$$

### 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 (7.99), for a Killing tensor

$$\xi^{\mu\nu} p_\mu p_\nu = \text{constant} . \quad (7.104)$$

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

$$g^{\mu\nu} p_\mu p_\nu = -m^2 = \text{constant} . \quad (7.105)$$

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 (7.100). A necessary and sufficient condition for this to be true is Killing's equation

$$\mathring{D}_{(\kappa} \xi_{\mu\nu)} = 0 , \quad (7.106)$$

where the parentheses denote symmetrization over all indices.

### 7.34 Lie derivative

It was remarked above that Killing's equation (7.102) can be recast as the statement that the Lie derivative of the metric along the Killing vector vanishes, equation (7.103). 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 (7.107). The change in coordinates should be understood as leaving the spacetime itself and physical quantities within it unchanged.

Let the coordinates  $x^\mu$  be changed by an infinitesimal amount  $\epsilon$  with a prescribed shape  $\xi^\mu(x)$  as a function of spacetime,

$$x^\mu \rightarrow x'^\mu = x^\mu + \epsilon \xi^\mu . \quad (7.107)$$

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 (7.107) is given by  $\epsilon$  times minus its Lie derivative, denoted  $\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots}$ ,

$$A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) \rightarrow A'^{\kappa\lambda\dots}_{\mu\nu\dots}(x) = A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) - \epsilon \mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} . \quad (7.108)$$

Equivalently,

$$\mathcal{L}_\xi A_{\mu\nu\dots}^{\kappa\lambda\dots} = \lim_{\epsilon \rightarrow 0} \frac{A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) - A'_{\mu\nu\dots}^{\kappa\lambda\dots}(x)}{\epsilon} . \quad (7.109)$$

The reason for the minus sign in the definition (7.108) of the Lie derivative is that, as will be seen below, equation (7.129), 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  $\xi^\kappa$ ,

$$\mathcal{L}_\xi A_{\dots} = \xi^\kappa \frac{\partial A_{\dots}}{\partial x^\kappa} + \dots . \quad (7.110)$$

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 7.14 What use is a Lie derivative? Answer.** The general rule to remember is that the change in any object under a coordinate transformation is, by construction, (minus) its Lie derivative. A prominent application of the Lie derivative is in general relativistic perturbation theory, Chapter 25, 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, §16.10.2. 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 (7.103).

### 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  $x'$ ,

$$DA \propto A(x') - A(x)_{\text{parallel-transported}} . \quad (7.111)$$

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  $x'$ ,

$$\mathcal{L}_\xi A \propto A(x') - A'(x')_{\text{parallel-transported}} . \quad (7.112)$$

Concept question 7.14 discusses the physical justification for this mathematical artifice.

### 7.34.2 Lie derivative of a coordinate scalar

Under a coordinate transformation (7.107), a coordinate-frame scalar  $\Phi(x)$  remains unchanged

$$\Phi(x) \rightarrow \Phi'(x') = \Phi(x). \quad (7.113)$$

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  $\Phi'$  at  $x$  is related to that at  $x'$  by

$$\Phi'(x) = \Phi'(x' - \epsilon\xi) = \Phi'(x') - \epsilon\xi^{\kappa} \frac{\partial\Phi'}{\partial x^{\kappa}}. \quad (7.114)$$

Since  $\epsilon$  is a small quantity, and  $\Phi'$  differs from  $\Phi$  by a small quantity, the last term  $\epsilon\xi^{\kappa} \partial\Phi'/\partial x^{\kappa}$  in equation (7.118) can be replaced by  $\epsilon\xi^{\kappa} \partial\Phi/\partial x^{\kappa}$  to linear order in  $\epsilon$ . Putting equations (7.113) and (7.114) together shows that the coordinate scalar  $\Phi$  changes under a coordinate transformation (7.107) as

$$\Phi(x) \rightarrow \Phi'(x) = \Phi(x) - \epsilon\mathcal{L}_{\xi}\Phi, \quad (7.115)$$

where  $\mathcal{L}_{\xi}\Phi$  is the Lie derivative of the scalar  $\Phi$ ,

$\mathcal{L}_{\xi}\Phi = \xi^{\kappa} \frac{\partial\Phi}{\partial x^{\kappa}}$

(7.116)

### 7.34.3 Lie derivative of a contravariant coordinate vector

A similar argument applies to coordinate vectors. Under an infinitesimal coordinate transformation (7.107), a contravariant coordinate 4-vector  $A^{\mu}(x)$  transforms in the usual way as

$$A^{\mu}(x) \rightarrow A'^{\mu}(x') = A^{\kappa}(x) \frac{\partial x'^{\mu}}{\partial x^{\kappa}} = A^{\mu}(x) + \epsilon A^{\kappa}(x) \frac{\partial\mu}{\partial x^{\kappa}}. \quad (7.117)$$

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  $A'^{\mu}$  at  $x$  is related to that at  $x'$  by

$$A'^{\mu}(x) = A'^{\mu}(x' - \epsilon\xi) = A'^{\mu}(x') - \epsilon\xi^{\kappa} \frac{\partial A'^{\mu}}{\partial x^{\kappa}}. \quad (7.118)$$

The last term  $\epsilon\xi^{\kappa} \partial A'^{\mu}/\partial x^{\kappa}$  in equation (7.118) can be replaced by  $\epsilon\xi^{\kappa} \partial A^{\mu}/\partial x^{\kappa}$  to linear order in the infinitesimal parameter  $\epsilon$ . Putting equations (7.117) and (7.118) together shows that the coordinate 4-vector  $A^{\mu}$  changes under a coordinate transformation (7.107) as

$$A^{\mu}(x) \rightarrow A'^{\mu}(x) = A^{\mu}(x) - \epsilon\mathcal{L}_{\xi}A^{\mu}, \quad (7.119)$$

where  $\mathcal{L}_\xi A^\mu$  is the Lie derivative of the contravariant vector  $A^\mu$ ,

$$\boxed{\mathcal{L}_\xi A^\mu = \xi^\kappa \frac{\partial A^\mu}{\partial x^\kappa} - A^\kappa \frac{\partial \xi^\mu}{\partial x^\kappa}} . \quad (7.120)$$

The ordinary partial derivatives in equation (7.120) 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)

$$\mathcal{L}_\xi A^\mu = \xi^\kappa \mathring{D}_\kappa A^\mu - A^\kappa \mathring{D}_\kappa \xi^\mu \quad \text{a coordinate vector .} \quad (7.121)$$

Equation (7.121) 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  $A^\mu$  in terms of torsion-full covariant derivatives  $D_\kappa$  is

$$\mathcal{L}_\xi A^\mu = \xi^\kappa D_\kappa A^\mu - A^\kappa D_\kappa \xi^\mu + A^\kappa \xi^\lambda S_{\kappa\lambda}^\mu , \quad (7.122)$$

where  $S_{\kappa\lambda}^\mu$  is the torsion.

**Exercise 7.15 Equivalence of expressions for the Lie derivative.** Confirm that equations (7.120), (7.121), and (7.122) are all equivalent.

### 7.34.4 Lie bracket

If  $A^\mu$  and  $B^\mu$  are two contravariant coordinate vectors, then the Lie derivative with respect to  $A^\mu$  of  $B^\mu$  is minus the Lie derivative with respect to  $B^\mu$  of  $A^\mu$ ,

$$\mathcal{L}_A B^\mu = A^\kappa \frac{\partial B^\mu}{\partial x^\kappa} - B^\kappa \frac{\partial A^\mu}{\partial x^\kappa} = -\mathcal{L}_B A^\mu . \quad (7.123)$$

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

$$\boxed{[\mathbf{A}, \mathbf{B}] \equiv \mathcal{L}_A \mathbf{B} = e_\mu \mathcal{L}_A B^\mu = -[\mathbf{B}, \mathbf{A}]} . \quad (7.124)$$

### 7.34.5 Lie derivative of a covariant coordinate vector

Under a coordinate transformation (7.107), a covariant coordinate 4-vector  $A_\mu(x)$  transforms in the usual way as

$$A_\mu(x) \rightarrow A'_\mu(x') = A_\kappa(x) \frac{\partial x^\kappa}{\partial x'^\mu} = A_\mu(x) - \epsilon A_\kappa(x) \frac{\partial \xi^\kappa}{\partial x^\mu} . \quad (7.125)$$

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  $A'_{\mu}$  at  $x$  is related to that at  $x'$  by

$$A'_{\mu}(x) = A'_{\mu}(x' - \epsilon\xi) = A'_{\mu}(x') - \epsilon\xi^{\kappa} \frac{\partial A'_{\mu}}{\partial x^{\kappa}}, \quad (7.126)$$

and again, the last term  $\epsilon\xi^{\kappa}\partial A'_{\mu}/\partial x^{\kappa}$  in equation (7.126) can be replaced by  $\epsilon\xi^{\kappa}\partial A_{\mu}/\partial x^{\kappa}$  to linear order in  $\epsilon$ . Putting equations (7.117) and (7.118) together shows that the covariant coordinate 4-vector  $A_{\mu}$  changes under a coordinate transformation (7.107) as

$$A_{\mu}(x) \rightarrow A'_{\mu}(x) = A_{\mu}(x) - \epsilon\mathcal{L}_{\xi}A_{\mu}, \quad (7.127)$$

where  $\mathcal{L}_{\xi}A_{\mu}$  is the Lie derivative of the covariant vector  $A_{\mu}$ ,

$$\boxed{\mathcal{L}_{\xi}A_{\mu} = \xi^{\kappa} \frac{\partial A_{\mu}}{\partial x^{\kappa}} + A_{\kappa} \frac{\partial \xi^{\kappa}}{\partial x^{\mu}}}. \quad (7.128)$$

### 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

$$\boxed{\mathcal{L}_{\xi}A_{\mu\nu\dots}^{\kappa\lambda\dots} \equiv \xi^{\pi} \frac{\partial A_{\mu\nu\dots}^{\kappa\lambda\dots}}{\partial x^{\pi}} + A_{\pi\nu\dots}^{\kappa\lambda\dots} \frac{\partial \xi^{\pi}}{\partial x^{\mu}} + A_{\mu\pi\dots}^{\kappa\lambda\dots} \frac{\partial \xi^{\pi}}{\partial x^{\nu}} \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} \frac{\partial \xi^{\kappa}}{\partial x^{\pi}} - A_{\mu\nu\dots}^{\kappa\pi\dots} \frac{\partial \xi^{\lambda}}{\partial x^{\pi}}} \quad \text{a coordinate tensor ,} \quad (7.129)$$

with an overall  $\partial A$  term, and a  $+\partial\xi$  term for each covariant index and a  $-\partial\xi$  term for each contravariant index. Equivalently, in terms of torsion-free covariant derivatives,

$$\mathcal{L}_{\xi}A_{\mu\nu\dots}^{\kappa\lambda\dots} = \xi^{\pi} \mathring{D}_{\pi}A_{\mu\nu\dots}^{\kappa\lambda\dots} + A_{\pi\nu\dots}^{\kappa\lambda\dots} \mathring{D}_{\mu}\xi^{\pi} + A_{\mu\pi\dots}^{\kappa\lambda\dots} \mathring{D}_{\nu}\xi^{\pi} \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} \mathring{D}_{\pi}\xi^{\kappa} - A_{\mu\nu\dots}^{\kappa\pi\dots} \mathring{D}_{\pi}\xi^{\lambda} \quad \text{a coordinate tensor .} \quad (7.130)$$

Equivalently, in terms of torsion-full covariant derivatives,

$$\begin{aligned} \mathcal{L}_{\xi}A_{\mu\nu\dots}^{\kappa\lambda\dots} &= \xi^{\pi} D_{\pi}A_{\mu\nu\dots}^{\kappa\lambda\dots} + A_{\pi\nu\dots}^{\kappa\lambda\dots} D_{\mu}\xi^{\pi} + A_{\mu\pi\dots}^{\kappa\lambda\dots} D_{\nu}\xi^{\pi} \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} D_{\pi}\xi^{\kappa} - A_{\mu\nu\dots}^{\kappa\pi\dots} D_{\pi}\xi^{\lambda} \dots \\ &\quad + (A_{\pi\nu\dots}^{\kappa\lambda\dots} S_{\mu\rho}^{\pi} + A_{\mu\pi\dots}^{\kappa\lambda\dots} S_{\nu\rho}^{\pi} \dots - A_{\mu\nu\dots}^{\pi\lambda\dots} S_{\pi\rho}^{\kappa} - A_{\mu\nu\dots}^{\kappa\pi\dots} S_{\pi\rho}^{\lambda}) \xi^{\rho} \quad \text{a coordinate tensor .} \end{aligned} \quad (7.131)$$

**Exercise 7.16 Lie derivative of the metric.** What is the Lie derivative of the metric tensor  $g_{\mu\nu}$  along the direction  $\xi^{\kappa}$ ?

**Solution.** The Lie derivative of the metric  $g_{\mu\nu}$  along  $\xi^{\kappa}$  is

$$\begin{aligned} \mathcal{L}_{\xi}g_{\mu\nu} &= \xi^{\kappa} \frac{\partial g_{\mu\nu}}{\partial x^{\kappa}} + g_{\kappa\nu} \frac{\partial \xi^{\kappa}}{\partial x^{\mu}} + g_{\mu\kappa} \frac{\partial \xi^{\kappa}}{\partial x^{\nu}} \\ &= \frac{\partial \xi_{\nu}}{\partial x^{\mu}} + \frac{\partial \xi_{\mu}}{\partial x^{\nu}} - 2\mathring{\Gamma}_{\kappa\mu\nu}\xi^{\kappa} \\ &= \mathring{D}_{\mu}\xi_{\nu} + \mathring{D}_{\nu}\xi_{\mu}, \end{aligned} \quad (7.132)$$

where  $\mathring{\Gamma}_{\kappa\mu\nu}$  is the torsion-free coordinate-frame connection, equation (2.62), and  $\mathring{D}_\mu$  is the torsion-free covariant derivative.

**Exercise 7.17 Lie derivative of the inverse metric.** Show that the Lie derivative of a Kronecker delta is zero,

$$\mathcal{L}_\xi \delta_\mu^\kappa = 0 . \quad (7.133)$$

Show that the Lie derivative of the inverse metric tensor  $g^{\kappa\lambda}$  is

$$\mathcal{L}_\xi g^{\kappa\lambda} = -g^{\kappa\mu} g^{\lambda\nu} \mathcal{L}_\xi g_{\mu\nu} = -(\mathring{D}^\kappa \xi^\lambda + \mathring{D}^\lambda \xi^\kappa) . \quad (7.134)$$


---

# 8

---

## 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  $c = G = 1$ ,

$$\boxed{ds^2 = -\Delta dt^2 + \Delta^{-1}dr^2 + r^2 d\theta^2}, \quad (8.1)$$

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  $r$ ,

$$\int_r^\infty \frac{E^2}{8\pi} 4\pi r^2 dr = \int_r^\infty \frac{Q^2}{8\pi r^4} 4\pi r^2 dr = \frac{Q^2}{2r}. \quad (8.4)$$

This seems 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

$$G_{\mu}^{\nu} = \begin{pmatrix} G_t^t & 0 & 0 & 0 \\ 0 & G_r^r & 0 & 0 \\ 0 & 0 & G_{\theta}^{\theta} & 0 \\ 0 & 0 & 0 & G_{\phi}^{\phi} \end{pmatrix} = 8\pi \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p_r & 0 & 0 \\ 0 & 0 & p_{\perp} & 0 \\ 0 & 0 & 0 & p_{\perp} \end{pmatrix} = \frac{Q^2}{r^4} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (8.5)$$

The trick of writing one index up and the other down on the Einstein tensor  $G_{\mu}^{\nu}$  partially cancels the distorting effect of the metric, yielding the proper energy density  $\rho$ , the proper radial pressure  $p_r$ , and transverse pressure  $p_{\perp}$ , up to factors of  $\pm 1$ . A more systematic way to extract proper quantities is to work in the tetrad formalism.

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 as 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  $\Delta$ , 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

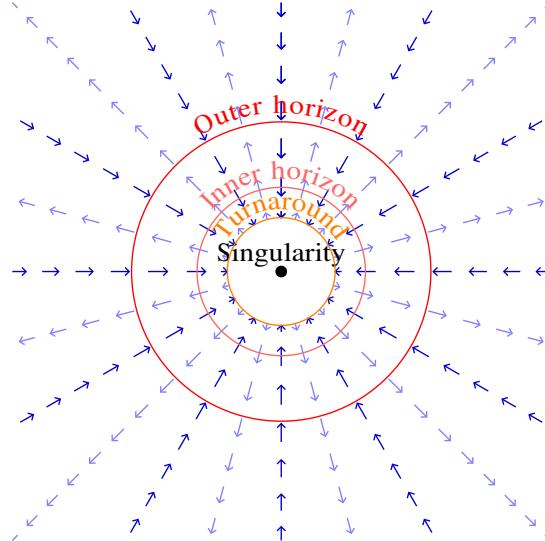


Figure 8.1 Depiction of the Gullstrand-Painlevé metric for the Reissner-Nordström geometry, for a black hole of charge  $Q = 0.96M$ . The Gullstrand-Painlevé line-element defines locally inertial frames attached to observers who free-fall radially from zero velocity at infinity. Frames fall at less than 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,

$$ds^2 = -dt_{\text{ff}}^2 + (dr - \beta dt_{\text{ff}})^2 + r^2 d\theta^2 . \quad (8.11)$$

The velocity  $\beta$  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  $\beta$  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  $\beta$  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  $\beta$  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 artifact 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.

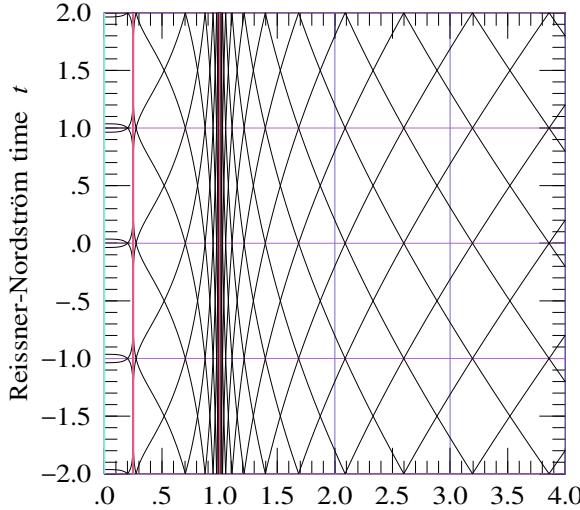


Figure 8.2 Spacetime diagram of the Reissner-Nordström geometry, in Reissner-Nordström coordinates, for a black hole of charge  $Q = 0.8M$ , plotted in units of the outer horizon radius  $r_+$  of the black hole. The geometry has two horizons (pink), an outer horizon, and an inner horizon at  $r_- = 0.25r_+$ . 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 artifact 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

$$t_F + r = t + r^* , \quad (8.19)$$

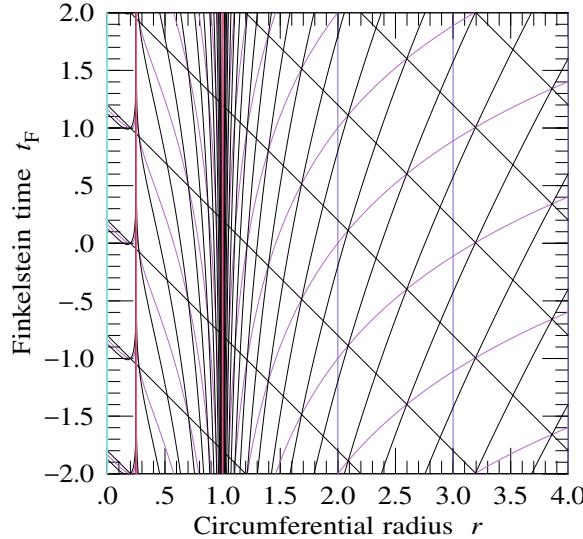


Figure 8.3 Finkelstein spacetime diagram of the Reissner-Nordström geometry, for a black hole of charge  $Q = 0.8M$ , plotted in units of the outer horizon radius  $r_+$  of the black hole. The Finkelstein time coordinate  $t_F$  is constructed so that radially infalling light rays are at  $45^\circ$ .

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

$$ds^2 = \Delta (-dt^2 + dr^{*2}) + r^2 d\Omega^2 . \quad (8.20)$$

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

$$r_K + t_K \equiv f(r^* + t) , \quad (8.21a)$$

$$r_K - t_K \equiv sf(r^* - t) + 2nk , \quad (8.21b)$$

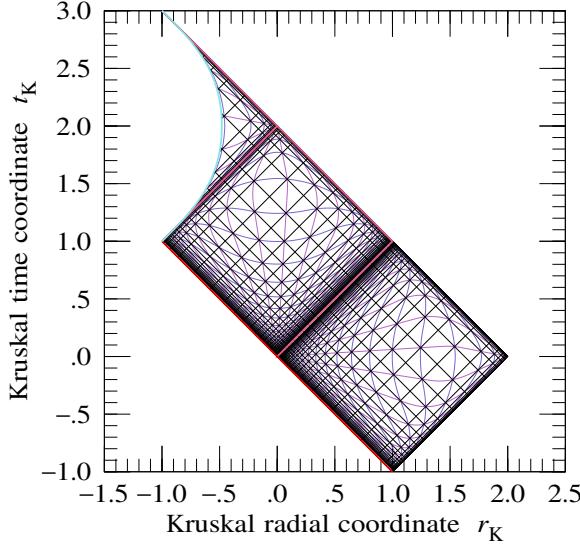


Figure 8.4 Kruskal spacetime diagram of the Reissner-Nordström geometry, plotted in units of  $k$ , equation (8.23), for a black hole of charge  $Q = 0.96M$ . The Kruskal coordinates  $t_K$  and  $r_K$  are defined by equations (8.21), and are constructed so that radially infalling and outgoing light rays are at  $45^\circ$ . 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.

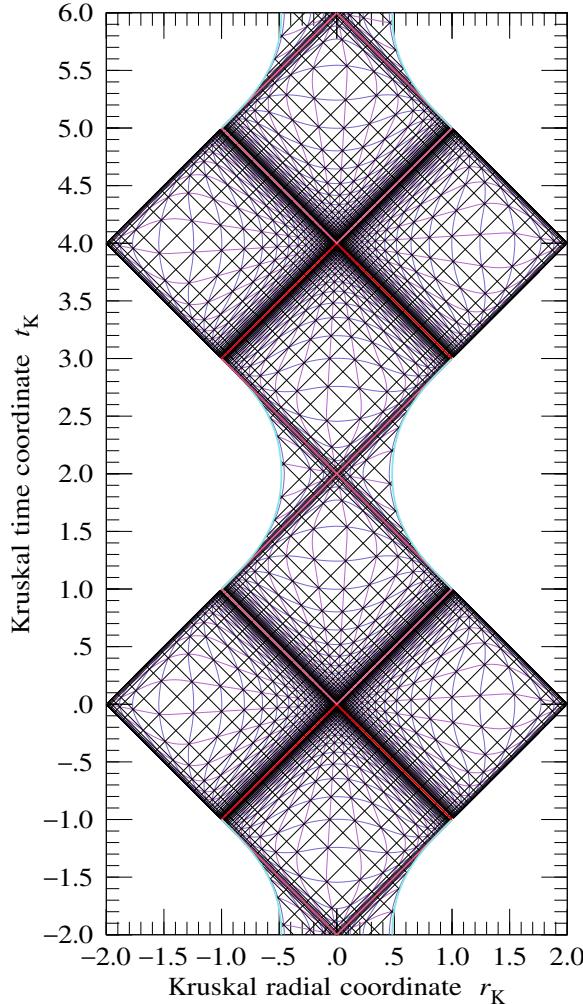


Figure 8.5 Kruskal spacetime diagram of the analytically extended Reissner-Nordström geometry, plotted in units of  $k$ , equation (8.23), for a black hole of charge  $Q = 0.96M$ .

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 → wormhole → 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

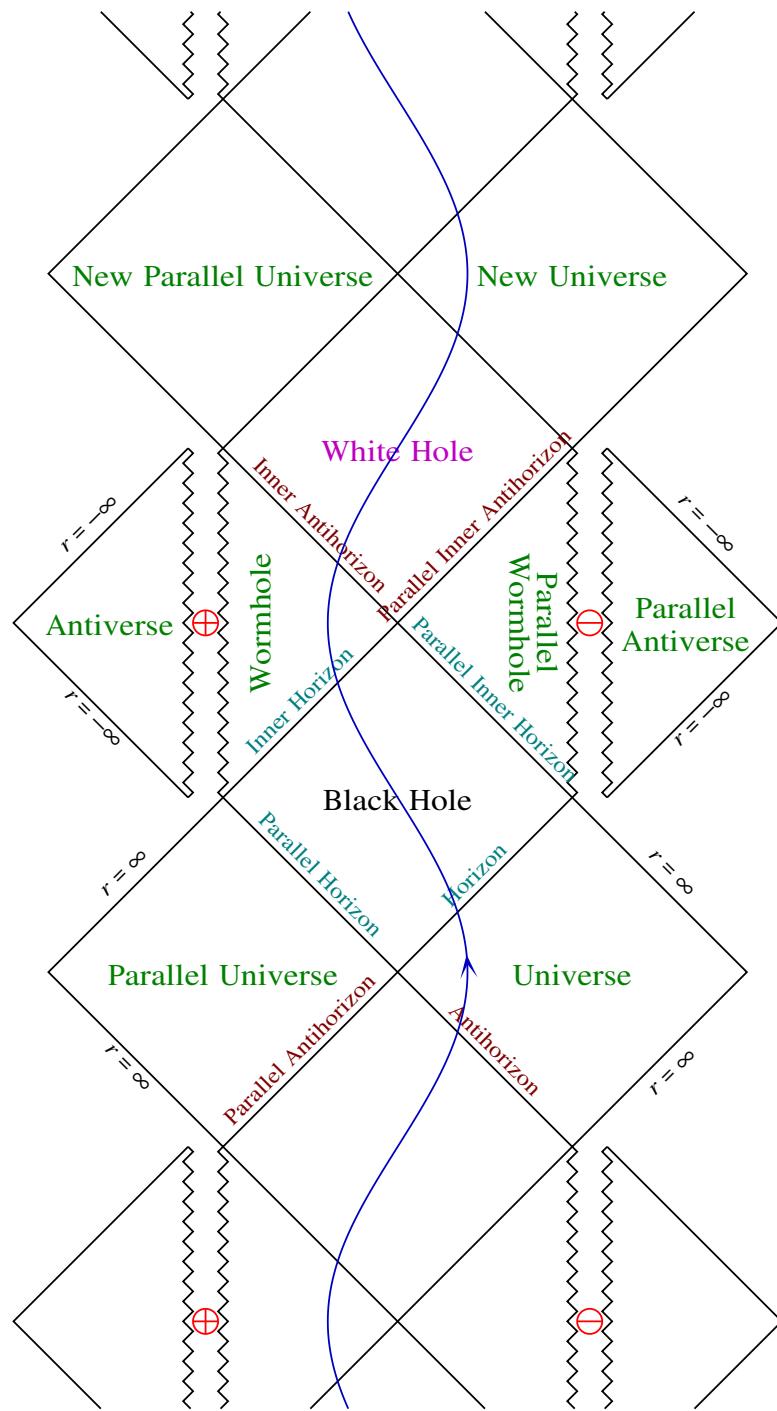


Figure 8.6 Penrose diagram of the analytically extended Reissner-Nordström geometry.

Wormhole, and away from the Parallel Wormhole. All observes 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 Ingoing, outgoing

The black hole in the Reissner-Nordström geometry is bounded at its inner edge by 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 ingoing and outgoing. 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 increasing<sup>1</sup>. Persons with positive energy are **ingoing**, while persons with negative energy are **outgoing**. Both ingoing and outgoing persons fall inward, to smaller radii, but ingoing persons think that the inward direction is towards the Wormhole, while outgoing persons think that the inward direction is in the opposite direction, towards the Parallel Wormhole. Ingoing persons fall through the ingoing inner horizon, while outgoing persons fall through the outgoing 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

<sup>1</sup> 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

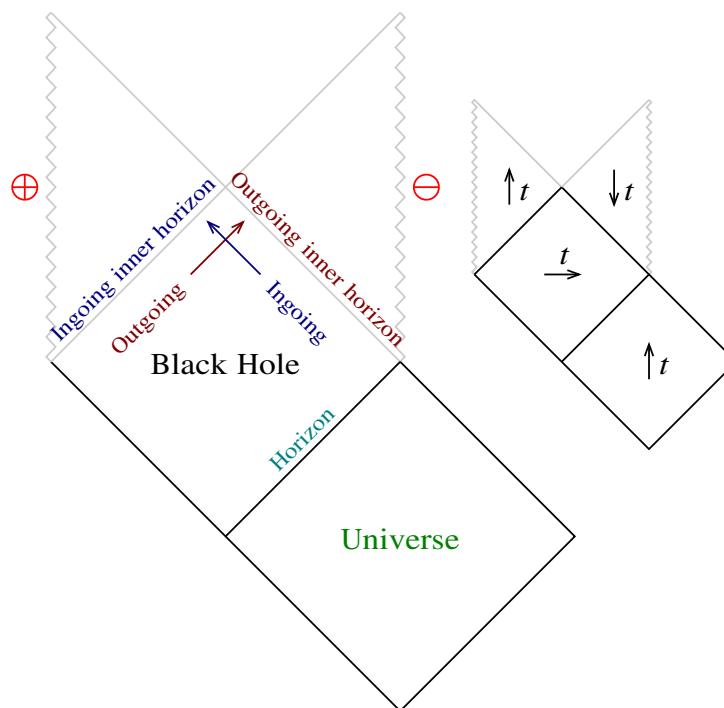


Figure 8.7 Penrose diagram illustrating why the Reissner-Nordström geometry is subject to the inflationary instability. Ingoing and outgoing streams just outside the inner horizon must pass through separate ingoing and outgoing inner horizons into causally separated pieces of spacetime where the timelike time coordinate  $t$  goes in opposite directions. To accomplish this, the ingoing and outgoing 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 ingoing and outgoing 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 ingoing and outgoing 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, ingoing and outgoing streams must drop through separate ingoing and outgoing 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, ingoing and outgoing 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, ingoing and outgoing streams resolve their differences by exceeding the speed of light relative to each other, and passing into causally separated regions. As the ingoing and outgoing streams drop through their respective inner horizons, they each see the other stream infinitely blueshifted.

In reality however, this cannot occur: ingoing and outgoing streams cannot exceed the speed of light relative to each other. Instead, as the ingoing and outgoing 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 ingoing and outgoing 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  $v_t = \pm 1$  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

$$u_\mu v^\mu = \frac{\text{something}}{\Delta} . \quad (8.25)$$

Argue that the blueshift diverges at the horizon for ingoing observers observing outgoing photons, and for outgoing observers observing ingoing photons.

---

## 8.14 The X point

The point in the Reissner-Nordström geometry where the ingoing and outgoing 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 ingoing and outgoing streams, coming from opposite directions, infinitely blueshifted.

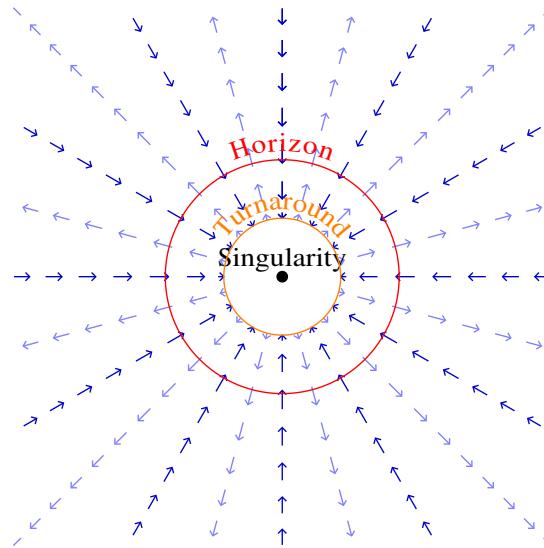


Figure 8.8 Depiction of the Gullstrand-Painlevé metric for the extremal Reissner-Nordström geometry, with  $Q = M$ . In the extremal geometry, the inner and outer horizons are at the same radius, so there is only one horizon.

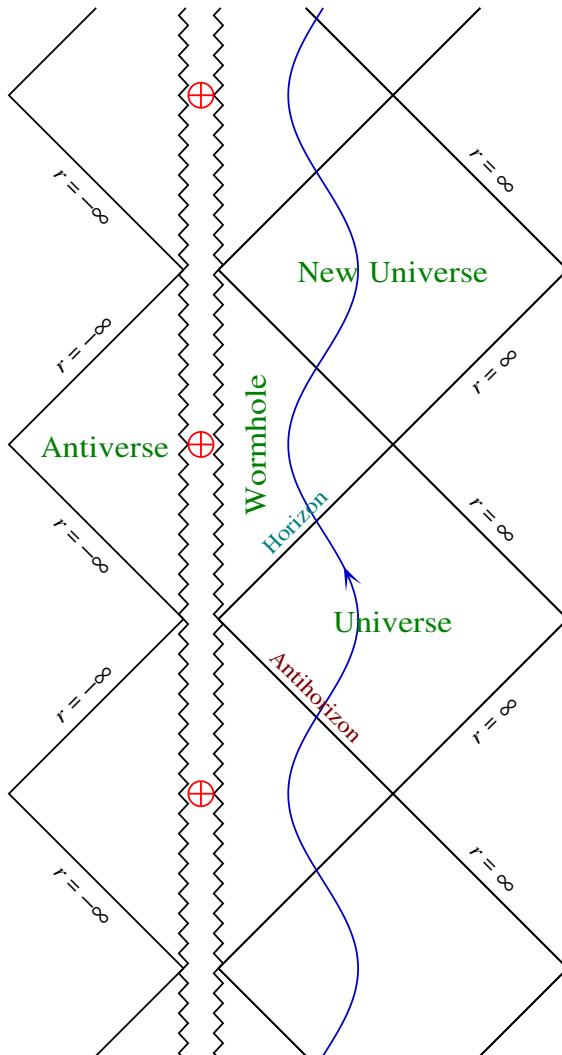


Figure 8.9 Penrose diagram of the extremal Reissner-Nordström geometry.

### 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,

$$Q = M , \quad (8.26)$$

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  $\beta$  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,

$$Q > M , \quad (8.27)$$

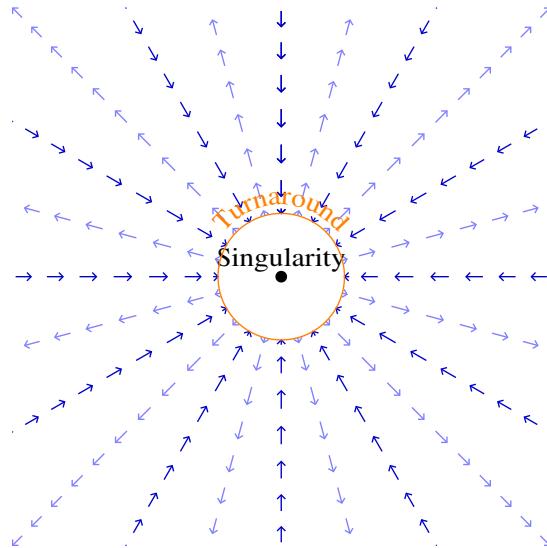


Figure 8.10 Depiction of the Gullstrand-Painlevé metric for a super-extremal Reissner-Nordström geometry, with  $Q = 1.04M$ . The super-extremal geometry has no horizons.

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-mass ratio of a proton is

$$\frac{e}{m_p} \approx 10^{18} \quad (8.28)$$

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. 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  $Q$

$$Q^2 < 0. \quad (8.29)$$

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 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.

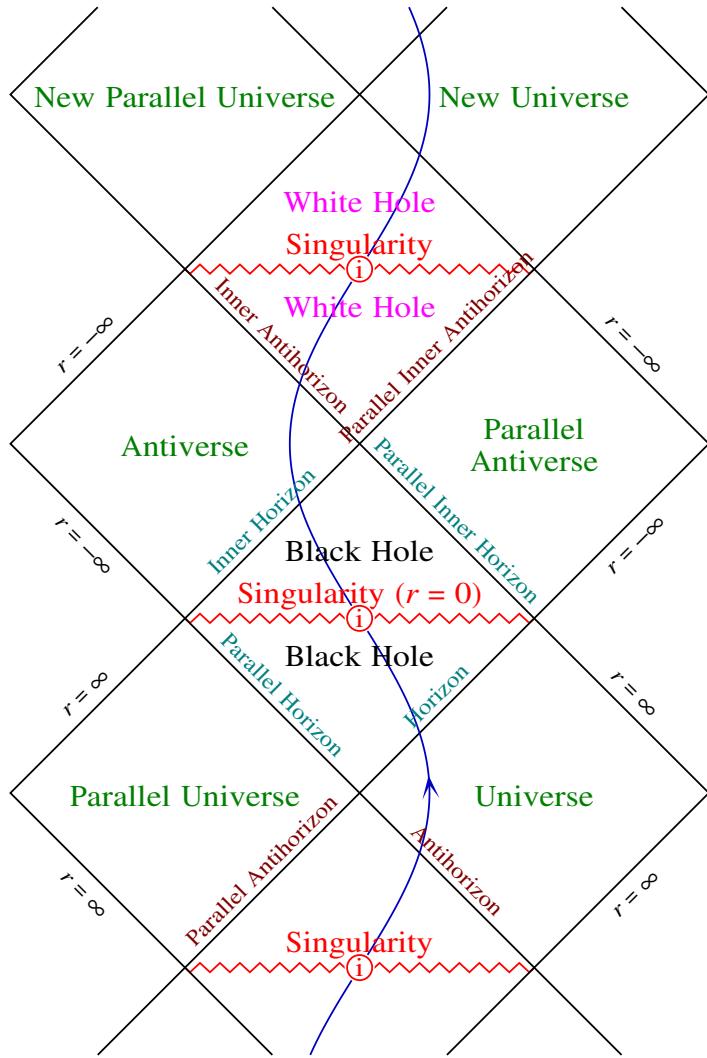


Figure 8.11 Penrose diagram of the Reissner-Nordström geometry with imaginary charge  $Q$ . If charge were imaginary, then electromagnetic energy would be negative, which is completely unphysical. But the metric is well-defined, and the spacetime is fun.

# 9

---

## 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 realized 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  $\rho$  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  $\Delta$  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 ellipsoidal coordinates.

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 (26.62), 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 (26.79) and (26.86), the scalar  $\Psi$ ,  $\Phi$  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 becomes apparent only in the tetrad frame, and are obscure in the coordinate frame.

## 9.2 Oblate spheroidal coordinates

Boyer-Lindquist coordinates  $r, \theta, \phi$  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.

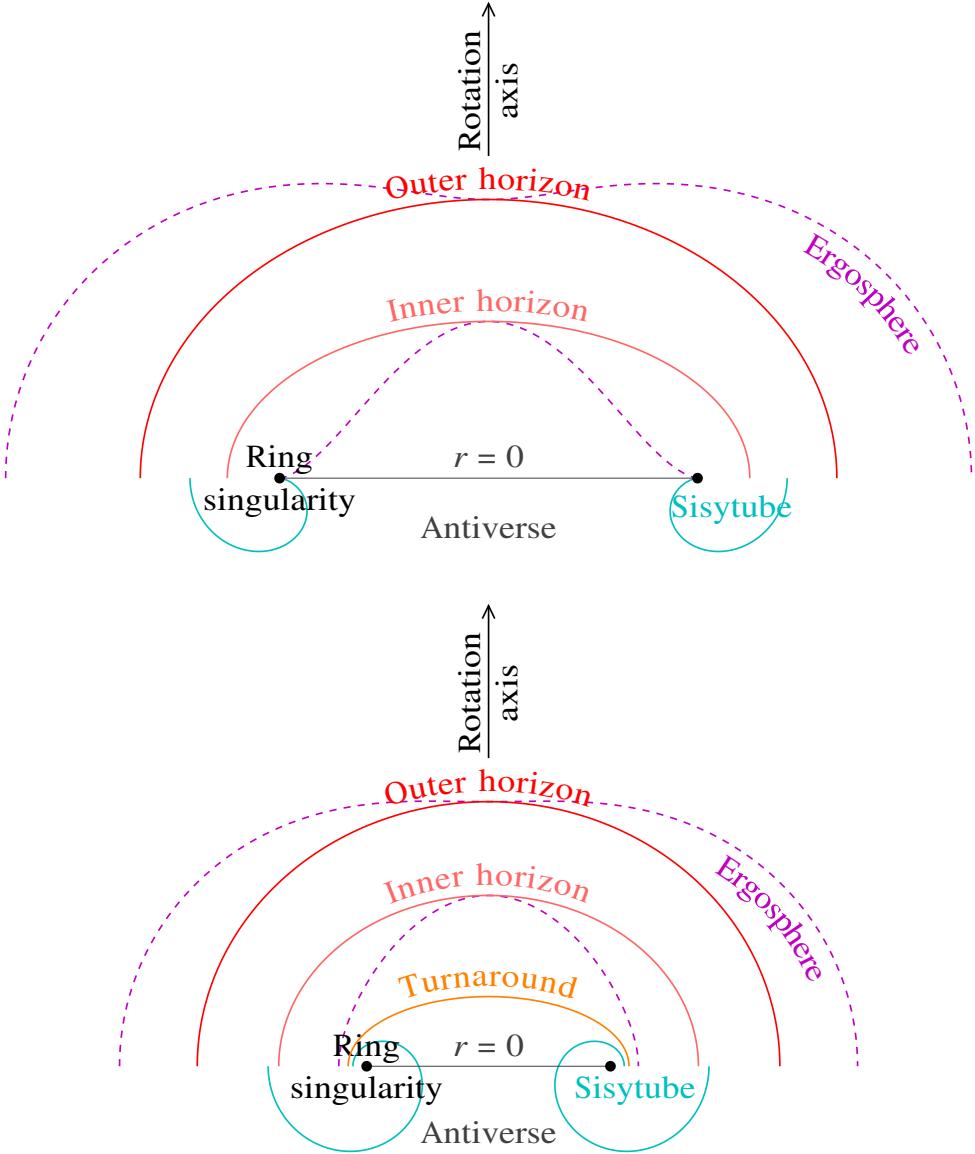


Figure 9.1 Spatial geometry of (upper) a Kerr black hole with spin parameter  $a = 0.96M$ , and (lower) a Kerr-Newman black hole with charge  $Q = 0.8M$  and spin parameter  $a = 0.56M$ . The upper half of each diagram shows  $r \geq 0$ , while the lower half shows  $r \leq 0$ , the Antiverse. The outer 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  $\phi$ . 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  $e_t$  and  $e_\phi$  in Boyer-Lindquist coordinates are Killing vectors. It follows that their scalar products

$$\begin{aligned} e_t \cdot e_t = g_{tt} &= -\frac{1}{\rho^2} (R^2 \Delta - a^2 \sin^2 \theta) , \\ e_t \cdot e_\phi = g_{t\phi} &= -\frac{a R^2 \sin^2 \theta}{\rho^2} (1 - \Delta) , \\ e_\phi \cdot e_\phi = g_{\phi\phi} &= \frac{R^2 \sin^2 \theta}{\rho^2} (R^2 - a^2 \sin^2 \theta \Delta) , \end{aligned} \quad (9.10)$$

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  $\phi$  are arranged further to satisfy the condition that  $e_t$  and  $e_\phi$  are each orthogonal to both  $e_r$  and  $e_\theta$ .

### 9.4 Ring singularity

The Kerr-Newman geometry contains a **ring singularity** where the Weyl tensor (9.26) diverges,  $\rho = 0$ , or equivalently at

$$r = 0 \text{ and } \theta = \pi/2 . \quad (9.11)$$

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  $\rho$  in a Kerr black hole.

### 9.5 Horizons

The horizon of a Kerr-Newmann 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  $\theta$ , but it moves in time  $t$  and azimuthal angle  $\phi$ . 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

$$0 = v_\mu v^\mu = g_{\mu\nu} v^\mu v^\nu = g_{tt} (v^t)^2 + 2 g_{t\phi} v^t v^\phi + g_{\phi\phi} (v^\phi)^2 . \quad (9.12)$$

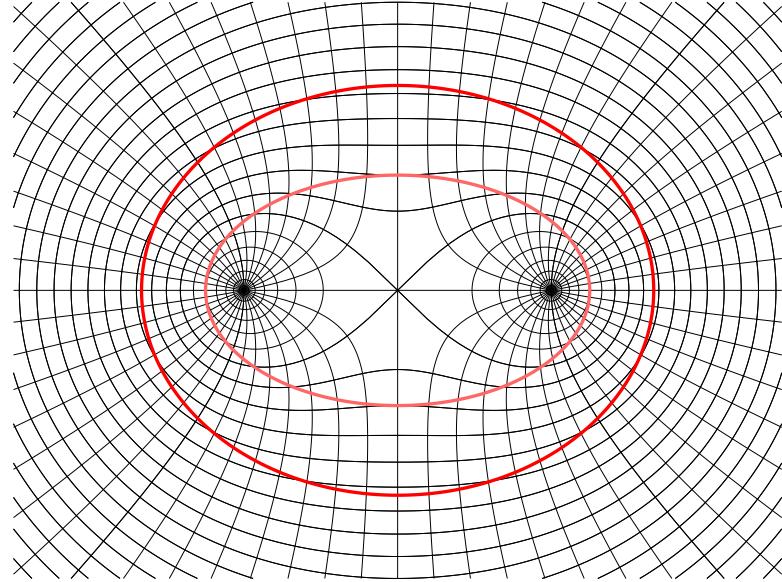


Figure 9.2 *Not* a mouse's eye view of a snake coming down its mousehole, uhoh. Contours of constant  $\rho$  and their covariant normals  $\partial\rho/\partial x^\mu$  in a spatial cross-section of a Kerr black hole of spin parameter  $a = 0.96M$ , in Boyer-Lindquist coordinates. The thicker contours 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  $2 \times 2$  matrix of metric coefficients in  $t$  and  $\phi$  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  $\Delta$  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  $\theta$ , whereas if  $\Delta < 0$ , then no such geodesics exist. The boundary

$$\Delta = 0 \quad (9.14)$$

defines the location of horizons. With  $\Delta$  given by equation (9.3), equation (9.14) gives **outer** and **inner horizons** at

$$r_\pm = M \pm \sqrt{M^2 - Q^2 - a^2} . \quad (9.15)$$

Between the horizons  $\Delta$  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  $\pi$ .

## 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  $t\text{-}\phi$  coordinates opens up to the point that  $\phi$  as well as  $t$  is a timelike coordinate. In the Wormhole, the direction of increasing proper time along  $t$  is  $t$  increasing, and along  $\phi$  is  $\phi$  decreasing, which is retrograde. In the Parallel Wormhole, the direction of increasing proper time along  $t$  is  $t$  decreasing, and along  $\phi$  is  $\phi$  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  $\phi$  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 (1968a), 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

$$M^2 = Q^2 + a^2 . \quad (9.24)$$

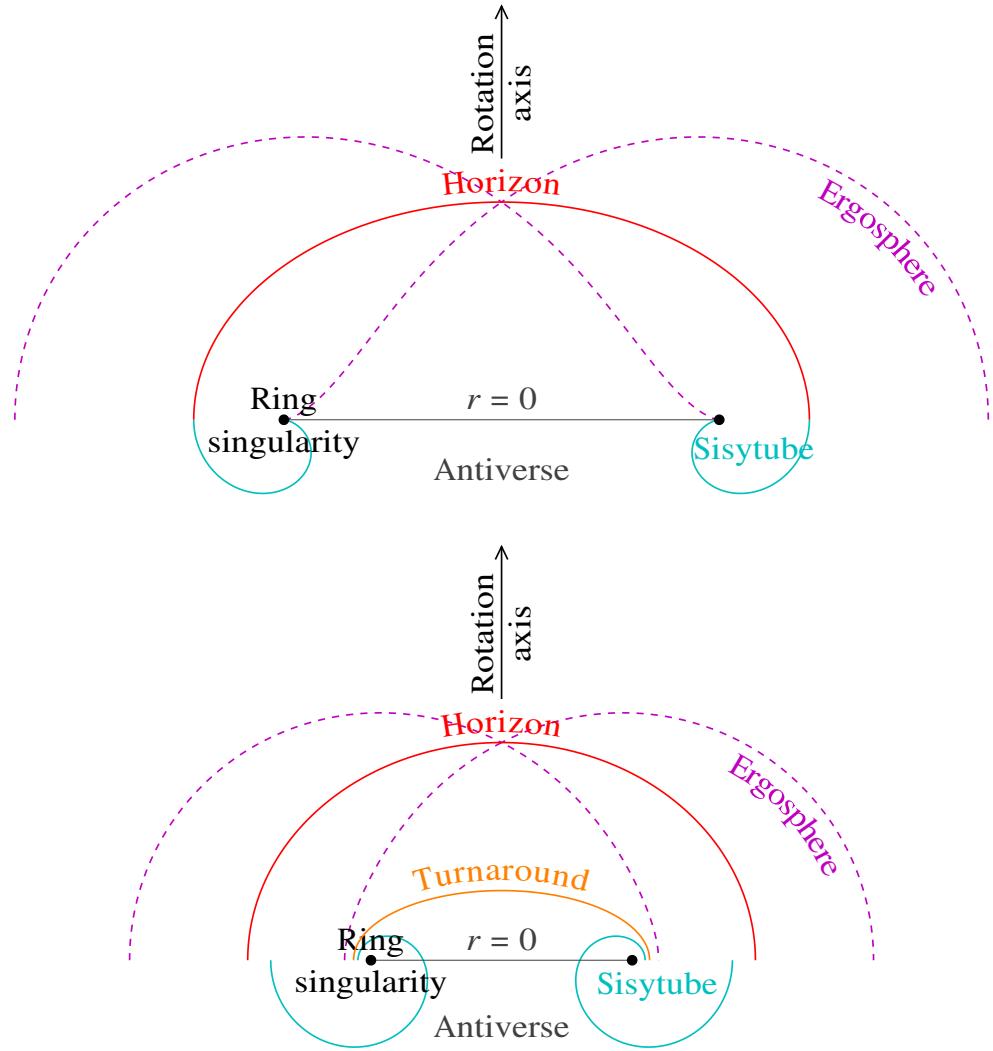


Figure 9.3 Spatial geometry of (upper) an extremal ( $a = M$ ) Kerr black hole, and (lower) an extremal Kerr-Newman black hole with charge  $Q = 0.8M$  and spin parameter  $a = 0.6M$ .

Figure 9.3 illustrates the structure of an extremal Kerr (uncharged) black hole, and an extremal Kerr-Newman (charged) black hole.

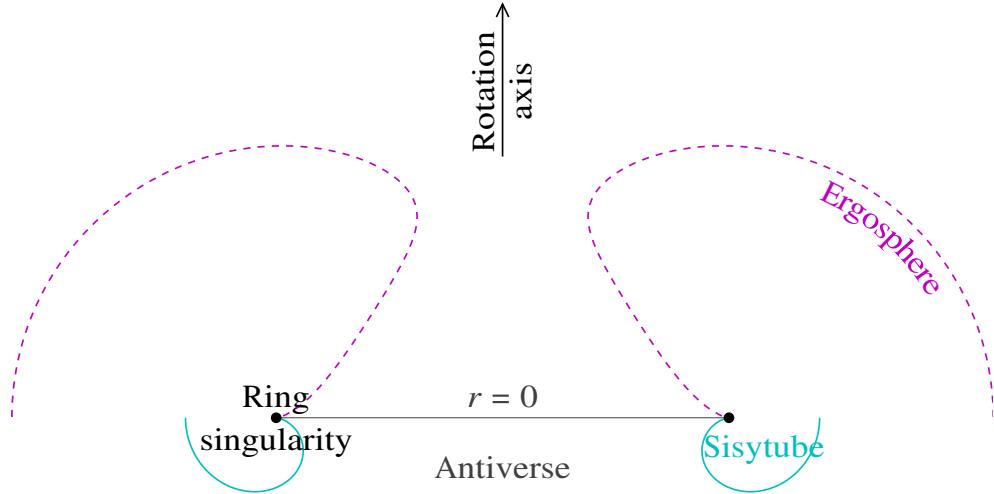


Figure 9.4 Spatial geometry of a super-extremal Kerr black hole with spin parameter  $a = 1.04M$ . A super-extremal black hole has no horizons.

### 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 un-hidden 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  $\rho$ , the proper radial pressure  $p_r$ , and the proper transverse pressure  $p_\perp$  in that frame are (do not confuse the notation  $\rho$  for proper density with the radial parameter  $\rho$ , 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,  $\pm 1$ , and  $\pm 2$ . 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  $A_{\mu}$  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 ingoing and outgoing 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 ingoing and outgoing 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 ingoing and outgoing null geodesics follow

$$\begin{aligned} r^* + t &= \text{constant} && \text{ingoing}, \\ r^* - t &= \text{constant} && \text{outgoing}. \end{aligned} \quad (9.31)$$

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  $t_{\text{ff}}$  and azimuthal angle  $\phi_{\text{ff}}$  are related to the Boyer-Lindquist time  $t$  and azimuthal angle  $\phi$  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  $t_{\text{ff}}$  is the proper time experienced by persons who free-fall from rest at infinity, with zero angular momentum. They follow trajectories of fixed  $\theta$  and  $\phi_{\text{ff}}$ , 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

$$u^{t_{\text{ff}}} = 1, \quad u^r = \frac{R^2 \beta}{\rho^2}, \quad u^{\theta} = 0, \quad u^{\phi_{\text{ff}}} = 0. \quad (9.34)$$

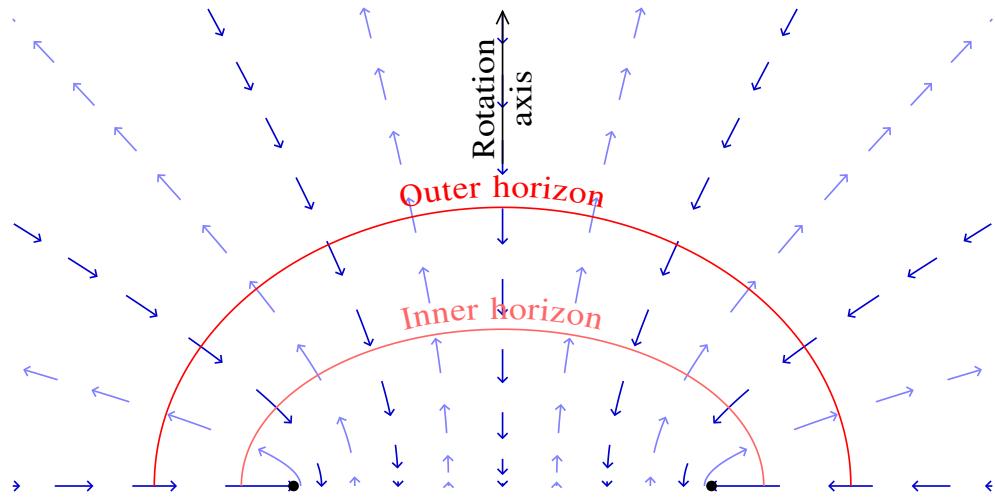


Figure 9.5 Spatial geometry of a Kerr black hole with spin parameter  $a = 0.96M$ . The arrows show the velocity  $\beta$  in the Doran metric. The flow follows lines of constant  $\theta$ , which form nested hyperboloids orthogonal to and confocal with the nested spheroids of constant  $r$ .

For the Kerr-Newman geometry, the velocity  $\beta$  is

$$\beta = \mp \frac{\sqrt{2Mr - Q^2}}{R} \quad (9.35)$$

where the  $\mp$  sign is  $-$  (infalling) for black hole solutions, and  $+$  (outfalling) for white hole solutions.

Horizons occur where the magnitude of the velocity  $\beta$  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

$$\beta = i \frac{\rho}{a \sin \theta} . \quad (9.39)$$

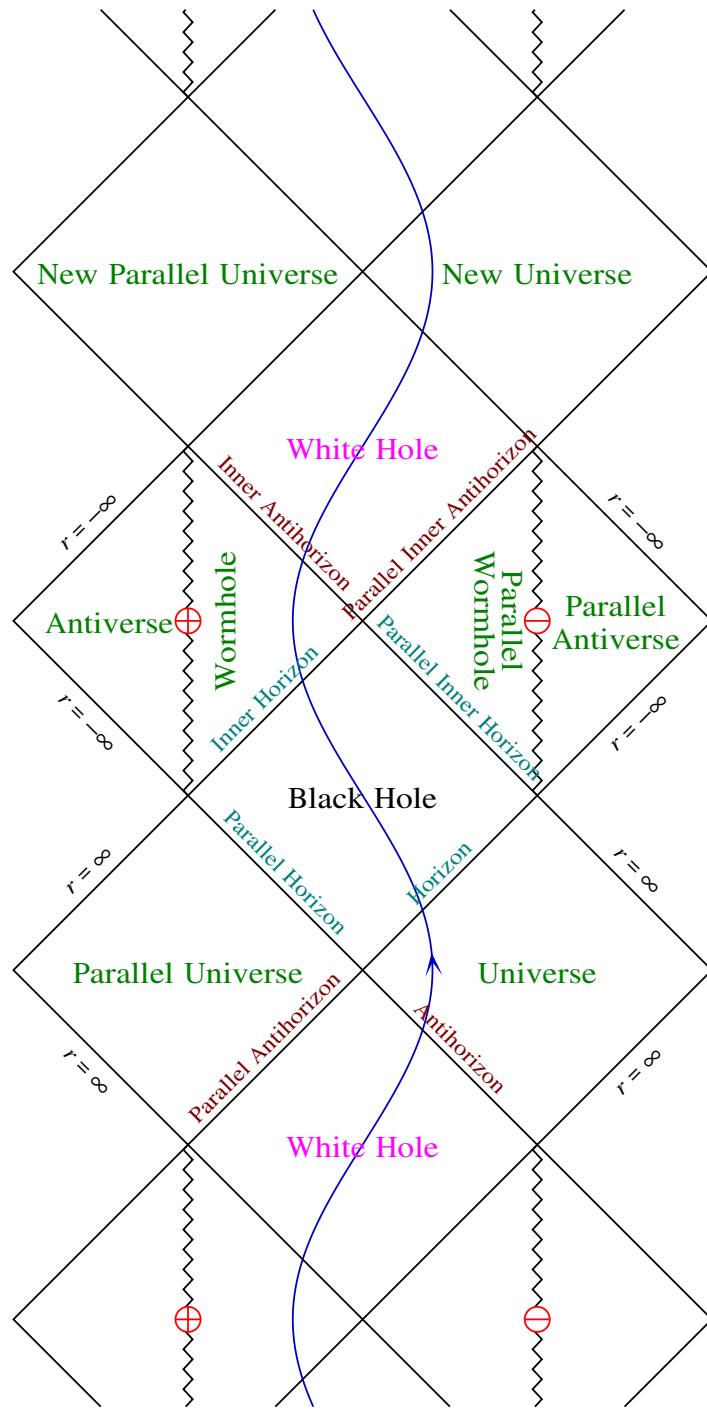


Figure 9.6 Penrose diagram of the Kerr-Newman geometry. The diagram is similar to that of the Reissner-Nordström geometry, except that it is possible to pass through the disk at  $r = 0$  from the Wormhole region into the Antiverse region. This Penrose diagram, which represents a slice at fixed  $\theta$  and  $\phi$ , 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 non-relativistic 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?

# 10

---

## Homogeneous, Isotropic Cosmology

### 10.1 Observational basis

Since 1998, observations have converged on a **Standard “ $\Lambda$ 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 ( $\Lambda$ ), 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 non-relativistic 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 Betoule (2014). 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

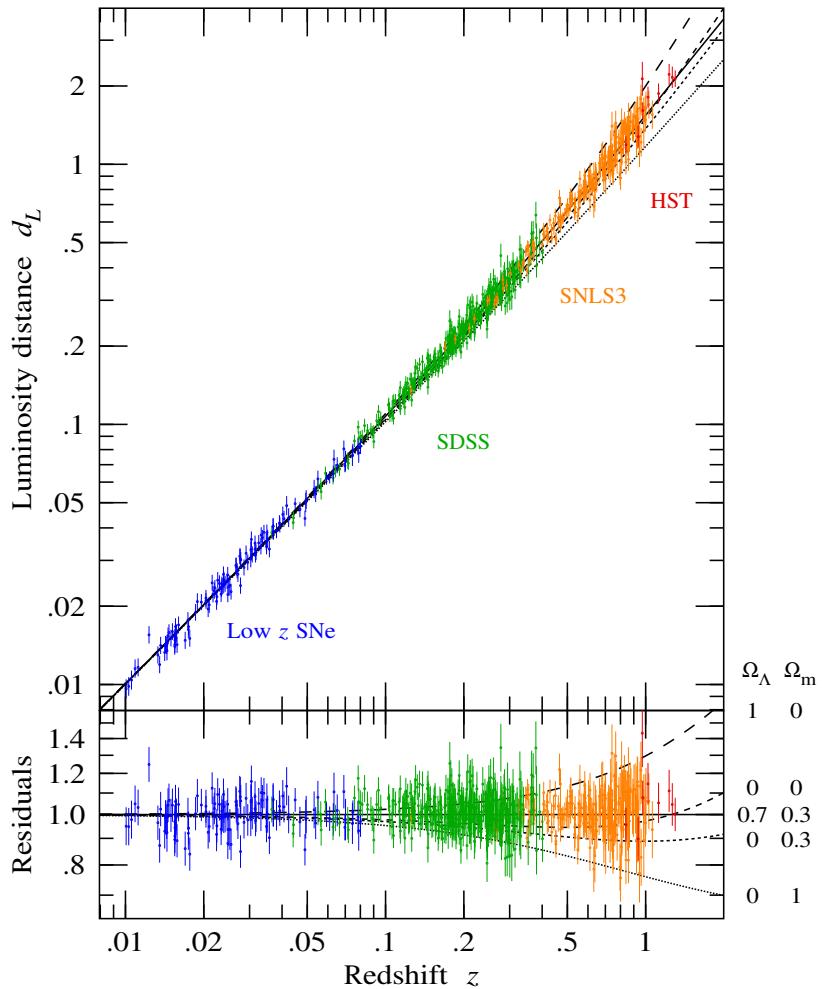


Figure 10.1 Hubble diagram of 740 Type Ia supernovae from a compilation of surveys, from data tabulated by Betoule (2014). The vertical axis is the luminosity distance  $d_L$  in units of the present-day Hubble distance  $c/H_0$ . The bottom panel shows residuals. The various smooth curves are 5 theoretical model Hubble diagrams, with parameters as indicated. The solid line is a flat  $\Lambda$ CDM model with  $\Omega_\Lambda = 0.7$  and  $\Omega_m = 0.3$ .

amount of  $^{56}\text{Ni}$  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, 1999), and the High- $z$  Supernova Search team (Riess and Filippenko, 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  $\Lambda$ .

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 meters in size, tuned to a wavelength of  $\sim 7\text{ cm}$ , that they had built to detect radio waves. They were unaware that Robert Dicke's group at Princeton had already realized 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 realized 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 with temperature (Fixsen, 2009)

$$T_0 = 2.72548 \pm 0.00057 \text{ K} . \quad (10.1)$$

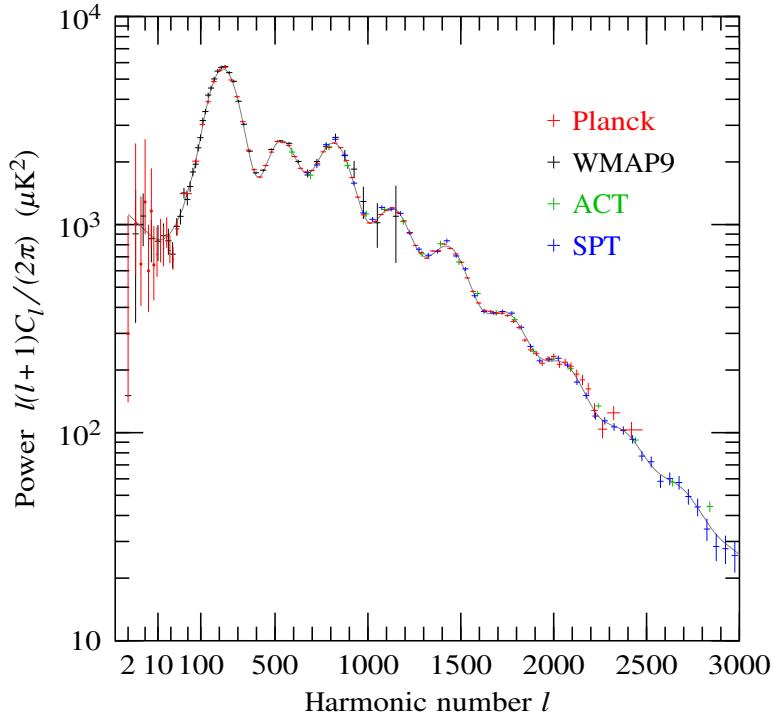


Figure 10.2 Power spectrum of fluctuations in the CMB from observations with the Planck satellite (Ade, 2013), WMAP (Hinshaw et al., 2012), the Atacama Cosmology Telescope (Das and Sherwin, 2011), and the South Pole Telescope (Keisler and Reichardt, 2011). The plot is logarithmic in harmonic number  $l$  up to 100, linear thereafter. The fit is a best-fit flat  $\Lambda$ CDM model.

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)

$$v = 369.1 \text{ km s}^{-1} \quad \text{in Galactic coordinate direction } \{l, b\} = \{263^\circ 99 \pm 0.14, 48^\circ 26 \pm 0.03\}. \quad (10.2)$$

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 10.2. 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.

Fluctuations in the CMB are expected to be polarized at some level. 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

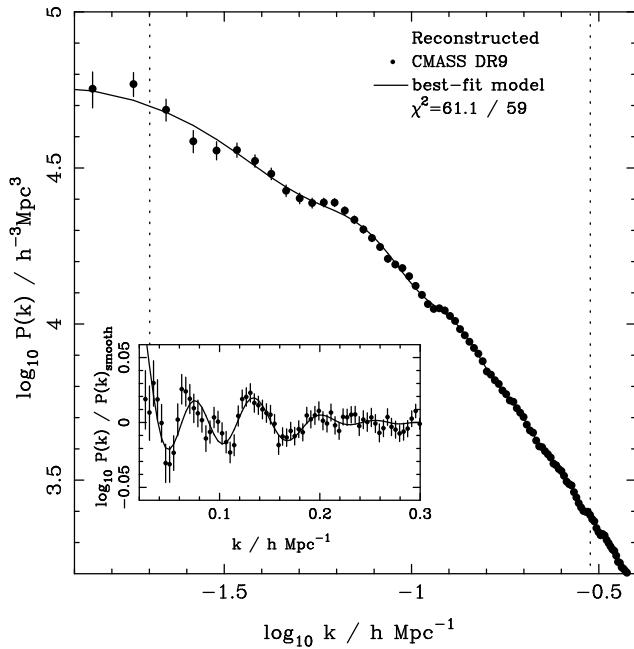


Figure 10.3 Power spectrum of galaxies from the Baryon Oscillation Spectroscopic Survey (BOSS), which is part of the Sloan Digital Sky Survey III (SDSS-III) (Anderson, 2013). The survey contains 264,283 massive galaxies, covering 3,275 square degrees. The fitted curve is a flat  $\Lambda$ CDM model multiplied by a polynomial. The inset shows the spectrum with the smooth component divided out to bring out the baryon acoustic oscillations (BAO).

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  $T-E$  cross-power spectrum has been measured by the WMAP satellite, and is interpreted as arising from scattering of CMB photons by ionized gas intervening between recombination and us.

#### 10.1.4 The clustering of galaxies

The clustering of galaxies shows a power spectrum in good agreement with the Standard Model, Figure 10.3.

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) anal-

ogous to those in the CMB. Observations from large galaxy surveys have been able to detect the predicted BAO, Figure 10.3. 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 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,  $^3\text{He}$ , 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** 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) and Friedmann (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), Robertson (1936a), and Robertson (1936b) 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 10.4 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 10.4) 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

$$dl^2 = dr_{\parallel}^2 + r^2 do^2 , \quad (10.7)$$

where  $do^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$  is the metric of a unit 2-sphere.

Introduce the angle  $\chi$  illustrated in the diagram. Evidently

$$\begin{aligned} r_{\parallel} &= R\chi, \\ r &= R \sin \chi. \end{aligned} \quad (10.8)$$

In terms of the angle  $\chi$ , the spatial metric is

$$dl^2 = R^2 (d\chi^2 + \sin^2\chi do^2), \quad (10.9)$$

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 10.4 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  $\chi$ , the metric can be expressed in terms of the circumferential radius  $r$ . It follows from equations (10.8) that

$$r_{\parallel} = R \arcsin(r/R), \quad (10.10)$$

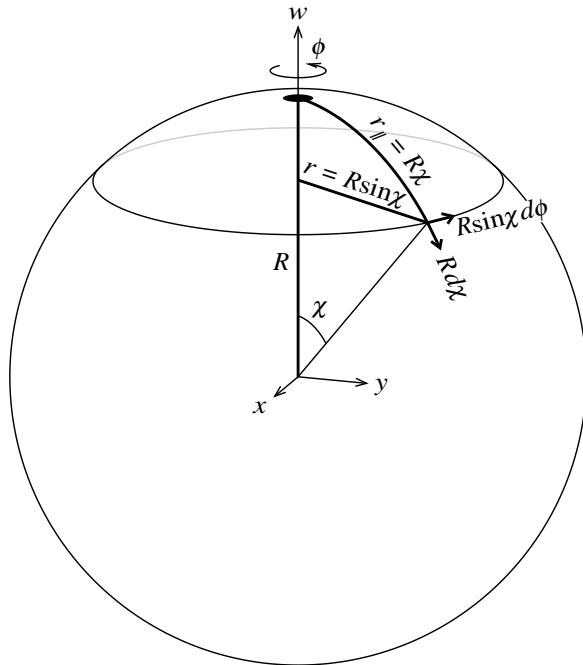


Figure 10.4 Embedding diagram of the FLRW geometry.

whence

$$\begin{aligned} dr_{\parallel} &= \frac{dr}{\sqrt{1 - r^2/R^2}} \\ &= \frac{dr}{\sqrt{1 - Kr^2}} , \end{aligned} \quad (10.11)$$

where  $K$  is the curvature

$$K \equiv \frac{1}{R^2} . \quad (10.12)$$

In terms of  $r$ , the spatial FLRW metric is then

$$\boxed{dl^2 = \frac{dr^2}{1 - Kr^2} + r^2 do^2} . \quad (10.13)$$

The embedding diagram Figure 10.4 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:

$$K \left\{ \begin{array}{lll} > 0 & \text{closed} & R \text{ real} , \\ = 0 & \text{flat} & R \rightarrow \infty , \\ < 0 & \text{open} & R \text{ imaginary} . \end{array} \right. \quad (10.14)$$

## 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

$$dl^2 = a^2 \left( \frac{dx^2}{1 - \kappa x^2} + x^2 do^2 \right),$$

(10.18)

where the curvature constant  $\kappa$ , 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

$$dl^2 = a^2 \left( dx_{\parallel}^2 + x^2 do^2 \right), \quad (10.20)$$

where

$$x = \begin{cases} \frac{\sin(\kappa^{1/2} x_{\parallel})}{\kappa^{1/2}} & \kappa > 0 \text{ closed}, \\ x_{\parallel} & \kappa = 0 \text{ flat}, \\ \frac{\sinh(|\kappa|^{1/2} x_{\parallel})}{|\kappa|^{1/2}} & \kappa < 0 \text{ open}. \end{cases} \quad (10.21)$$

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

$$dl^2 = a^2 \left( d\chi^2 + x^2 do^2 \right), \quad (10.22)$$

where

$$x = \begin{cases} \sin(\chi) & \kappa = 1 \text{ closed}, \\ \chi & \kappa = 0 \text{ flat}, \\ \sinh(\chi) & \kappa = -1 \text{ open}. \end{cases} \quad (10.23)$$

## 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

$$dl^2 = f(r) dr^2 + r^2 do^2. \quad (10.25)$$

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

$$dl^2 = \frac{4a^2}{(1 + \kappa X^2)^2} (dX^2 + X^2 do^2). \quad (10.26)$$

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 10.5. 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.

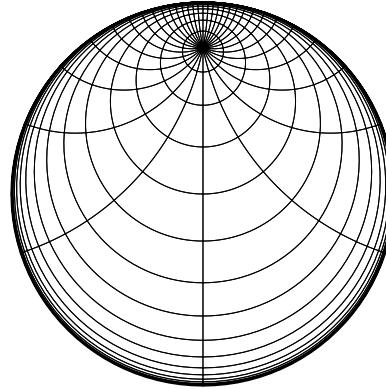


Figure 10.5 The Poincaré disk depicts the geometry of an open FLRW universe in isotropic coordinates. The lines are lines of latitude and longitude relative to a “pole” chosen here to be displaced from the centre of the disk. In isotropic coordinates, geodesics correspond to circles that intersect the 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.

**Solution.**

$$X = \frac{x}{1 + \sqrt{1 - \kappa x^2}} = \frac{1}{\sqrt{\kappa}} \tan\left(\frac{\sqrt{\kappa} x_{||}}{2}\right), \quad x = \frac{2X}{1 + \kappa X^2} = \frac{1}{\sqrt{\kappa}} \sin\left(\sqrt{\kappa} x_{||}\right). \quad (10.27)$$

For an open geometry with  $\kappa = -1$ ,  $X$  goes from 0 to 1 as  $x$  goes from 0 to  $\infty$ .

## 10.7 FLRW metric

The full Friedmann-Lemaître-Robertson-Walker spacetime metric is

$$ds^2 = -dt^2 + a(t)^2 \left( \frac{dx^2}{1 - \kappa x^2} + x^2 d\theta^2 \right), \quad (10.28)$$

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).

## 10.8 Einstein equations for FLRW metric

The Einstein equations for the FLRW metric (10.28) are

$$-G_{\textcolor{brown}{t}}^{\textcolor{brown}{t}} = 3 \left( \frac{\dot{a}^2}{a^2} + \frac{\kappa}{a^2} \right) = 8\pi G\rho, \quad (10.29\text{a})$$

$$G_x^x = G_\theta^\theta = G_\phi^\phi = -\frac{2\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{\kappa}{a^2} = 8\pi Gp, \quad (10.29\text{b})$$

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**

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G\rho}{3} - \frac{\kappa}{a^2}, \quad (10.30\text{a})$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} (\rho + 3p). \quad (10.30\text{b})$$

Friedmann's two equations (10.30) are fundamental to cosmology. The first one relates the curvature  $\kappa$  of the Universe to the expansion rate  $\dot{a}/a$  and the density  $\rho$ . The second one relates the acceleration  $\ddot{a}/a$  to the density  $\rho$  plus 3 times the pressure  $p$ .

## 10.9 Newtonian “derivation” of Friedmann equations

The Friedmann equations can be reproduced with a heuristic Newtonian argument.

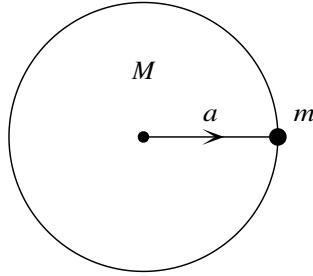


Figure 10.6 Newtonian picture in which the Universe is modeled as a uniform density sphere of radius  $a$  and mass  $M$  that gravitationally attracts a test mass  $m$ .

### 10.9.1 Energy equation

Model a piece of the Universe as a ball of radius  $a$  with uniform density  $\rho$ , 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

$$\frac{1}{2}m\dot{a}^2 - \frac{GMm}{a} = -\frac{\kappa mc^2}{2}, \quad (10.31)$$

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 (10.31) rearranges to

$$\boxed{\frac{\dot{a}^2}{a^2} = \frac{8\pi G\rho}{3} - \frac{\kappa c^2}{a^2}}, \quad (10.32)$$

which reproduces the first Friedmann equation.

### 10.9.2 First law of thermodynamics

For adiabatic expansion, the first law of thermodynamics is

$$dE + p dV = 0. \quad (10.33)$$

With  $E = \rho V$  and  $V = \frac{4}{3}\pi a^3$ , the first law (10.56) becomes

$$d(\rho a^3) + p da^3 = 0, \quad (10.34)$$

or, with the derivative taken with respect to cosmic time  $t$ ,

$$\dot{\rho} + 3(\rho + p)\frac{\dot{a}}{a} = 0. \quad (10.35)$$

Differentiating the first Friedmann equation in the form

$$\dot{a}^2 = \frac{8\pi G\rho a^2}{3} - \kappa c^2 \quad (10.36)$$

gives

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3} (\dot{\rho}a^2 + 2\rho a\dot{a}) , \quad (10.37)$$

and substituting  $\dot{\rho}$  from the first law (10.35) reduces this to

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3} a\dot{a} (-\rho - 3p) . \quad (10.38)$$

Hence

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} (\rho + 3p)} , \quad (10.39)$$

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. 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

$$\boxed{H \equiv \frac{\dot{a}}{a}} . \quad (10.40)$$

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)

$$H_0 = 73.8 \pm 2.4 \text{ km s}^{-1} \text{ Mpc}^{-1} . \quad (10.41)$$

The observed CMB power spectrum, Figure 10.2, 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 31. 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 10.3, 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 (Ade, 2015)

$$H_0 = 67.3 \pm 0.7 \text{ km s}^{-1} \text{ Mpc}^{-1} . \quad (10.42)$$

The Cepheid and CMB measurements (10.41) and (10.42) 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 are 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

$$\boxed{v = H_0 d} . \quad (10.43)$$

Since it takes light time to travel from a distant object, and the Hubble parameter varies in time, the linear relation (10.43) 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

$$h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}} . \quad (10.44)$$

The reciprocal of the Hubble constant gives an approximate estimate of the age of the Universe (c.f. Exercise 10.6),

$$\frac{1}{H_0} = 9.778 h^{-1} \text{ Gyr} = 14.0 h_{0.70}^{-1} \text{ Gyr} . \quad (10.45)$$

Table 10.1 *Cosmic inventory*

Species	Hinshaw et al. (2012)	Ade (2015)
Dark energy ( $\Lambda$ )	$\Omega_\Lambda$	$0.72 \pm 0.01$
Non-baryonic cold dark matter (CDM)	$\Omega_c$	$0.24 \pm 0.01$
Baryonic matter	$\Omega_b$	$0.047 \pm 0.002$
Neutrinos	$\Omega_\nu$	$< 0.02$
Photons (CMB)	$\Omega_\gamma$	$5 \times 10^{-5}$
Total	$\Omega$	$1.003 \pm 0.004$
Curvature	$\Omega_k$	$-0.003 \pm 0.004$

## 10.11 Critical density

The critical density  $\rho_{\text{crit}}$  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

$$\boxed{\rho_{\text{crit}} \equiv \frac{3H^2}{8\pi G}} . \quad (10.46)$$

The critical density  $\rho_{\text{crit}}$ , like the Hubble parameter  $H$ , evolves with time.

## 10.12 Omega

Cosmologists designate the ratio of the actual density  $\rho$  of the Universe to the critical density  $\rho_{\text{crit}}$  by the fateful letter  $\Omega$ , the final letter of the Greek alphabet,

$$\boxed{\Omega \equiv \frac{\rho}{\rho_{\text{crit}}}} . \quad (10.47)$$

With no subscript,  $\Omega$  denotes the total mass-energy density in all forms. A subscript  $x$  on  $\Omega_x$  denotes mass-energy density of type  $x$ .

The curvature density  $\rho_k$ , which is not really a form of mass-energy but it is sometimes convenient to treat it as though it were, is defined by

$$\rho_k \equiv -\frac{3\kappa c^2}{8\pi G a^2} , \quad (10.48)$$

and correspondingly  $\Omega_k \equiv \rho_k / \rho_{\text{crit}}$ . According to the first of Friedmann's equations (10.30), the curvature density  $\Omega_k$  satisfies

$$\boxed{\Omega_k = 1 - \Omega} . \quad (10.49)$$

Note that  $\Omega_k$  has opposite sign from  $\kappa$ , so a closed universe has negative  $\Omega_k$ .

Table 10.1 gives measurements of  $\Omega$  in various species, as reported by Hinshaw et al. (2012) from the final analysis of the CMB power spectrum from WMAP, and by Ade (2015) from the 2015 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 10.2, supernova data, Figure 10.1, galaxy clustering (Baryonic Acoustic Oscillation, or BAO) data, Figure 10.3, and local measurements of the Hubble constant  $H_0$  (Riess et al., 2011). 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  $\Omega_k$ , as illustrated in Figure 20 of Ade (2015). 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 (Ade, 2015),

$$\sum_{\nu} m_{\nu} < 0.2 \text{ eV} . \quad (10.50)$$

**Exercise 10.2 Omega in photons.** Most of the energy density in electromagnetic radiation today is in CMB photons. Calculate  $\Omega_{\gamma}$  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 10.20.

**Solution.** CMB photons have a blackbody spectrum at temperature  $T_0 = 2.725$  K, so their density can be calculated from the blackbody formulae. The present day ratio  $\Omega_{\gamma}$  of the mass-energy density  $\rho_{\gamma}$  of CMB photons to the critical density  $\rho_{\text{crit}}$  is

$$\Omega_{\gamma} \equiv \frac{\rho_{\gamma}}{\rho_{\text{crit}}} = \frac{8\pi G \rho_{\gamma}}{3H_0^2} = \frac{8\pi^3 G (kT_0)^4}{45H_0^2 c^5 \hbar^3} = 2.471 \times 10^{-5} h^{-2} T_{2.725 \text{ K}}^4 = 5.0 \times 10^{-5} h_{0.70}^{-2} T_{2.725 \text{ K}}^4 . \quad (10.51)$$

## 10.13 Types of mass-energy

The energy-momentum tensor  $T_{\mu\nu}$  of a 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)

$$T_{\nu}^{\mu} = \begin{pmatrix} T_t^t & 0 & 0 & 0 \\ 0 & T_r^r & 0 & 0 \\ 0 & 0 & T_{\theta}^{\theta} & 0 \\ 0 & 0 & 0 & T_{\phi}^{\phi} \end{pmatrix} = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} . \quad (10.52)$$

Table 10.2 gives equations of state  $p/\rho$  for generic species of mass-energy, along with  $(\rho + 3p)/\rho$ , which

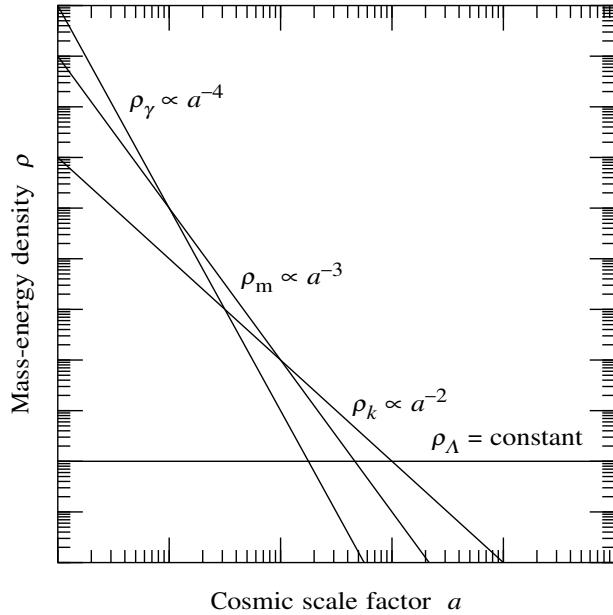


Figure 10.7 Behaviour of the mass-energy density  $\rho$  of various species as a function of cosmic time  $t$ .

determines the gravitational attraction (deceleration) per unit energy, and how the mass-energy varies with cosmic scale factor,  $\rho \propto a^n$ , Exercise 10.3.

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

$$D_{\mu} T^{\mu\nu} = 0. \quad (10.53)$$

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  $\rho$  varies with cosmic scale factor  $a$ . Figure 10.7 illustrates how the energy densities  $\rho$  of various species evolve as a function of scale factor  $a$ .

Table 10.2 *Properties of universes dominated by various species*

Species	$p/\rho$	$(\rho + 3p)/\rho$	$\rho \propto$
Radiation	1/3	2	$a^{-4}$
Matter	0	1	$a^{-3}$
Curvature	“−1/3”	“0”	$a^{-2}$
Vacuum	−1	−2	$a^0$

Vacuum energy is equivalent to a **cosmological constant**. Einstein originally introduced the cosmological constant  $\Lambda$  as a modification to the left hand side of his equations,

$$G_{\kappa\mu} + \Lambda g_{\kappa\mu} = 8\pi G T_{\kappa\mu} . \quad (10.54)$$

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  $\rho_\Lambda$ , satisfying

$$\Lambda = 8\pi G \rho_\Lambda . \quad (10.55)$$

### Exercise 10.3 Mass-energy in a 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):

$$d(\rho a^3) + p da^3 = 0 . \quad (10.56)$$

How does the density  $\rho$  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

$$\rho \propto a^n . \quad (10.57)$$

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 10.4 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?

## 10.14 Redshifting

The spatial translation symmetry of the FLRW metric implies conservation of generalized momentum. As you will show in Exercise 10.5, a particle that moves along a geodesic in the radial direction, so that  $d\theta = d\phi = 0$ , has 4-velocity  $u^\nu$  satisfying

$$u_{x\parallel} = \text{constant} . \quad (10.58)$$

This conservation law implies that the proper momentum  $p_{x\parallel}$  of a radially moving particle decays as

$$p_{x\parallel} \equiv ma \frac{dx\parallel}{d\tau} \propto \frac{1}{a} , \quad (10.59)$$

which is true for both massive and massless particles.

It follows from equation (10.66) that light observed on Earth from a distant object will be redshifted by a factor

$$1 + z = \frac{a_0}{a}, \quad (10.60)$$

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 10.5 Geodesics in the FLRW geometry.** The Friedmann-Lemaître-Robertson-Walker metric of cosmology is

$$ds^2 = -dt^2 + a(t)^2 \left[ dx_{\parallel}^2 + \frac{\sin^2(\kappa^{1/2}x_{\parallel})}{\kappa} (d\theta^2 + \sin^2\theta d\phi^2) \right], \quad (10.61)$$

where  $\kappa$  is a constant, the curvature constant. Note that equation (10.61) is valid for all values of  $\kappa$ , including zero and negative values: there is no need to consider the cases separately.

1. **Conservation of generalized momentum.** Consider a particle moving along a geodesic in the radial direction, so that  $d\theta = d\phi = 0$ . Argue that the Lagrangian equations of motion

$$\frac{d}{d\tau} \frac{\partial L}{\partial u^{\textcolor{brown}{x}_{\parallel}}} = \frac{\partial L}{\partial x_{\parallel}} \quad (10.62)$$

with effective Lagrangian

$$L = \frac{1}{2} g_{\mu\nu} u^{\mu} u^{\nu} \quad (10.63)$$

imply that

$$u_{\textcolor{brown}{x}_{\parallel}} = \text{constant}. \quad (10.64)$$

Argue further from the same Lagrangian equations of motion that the assumption of a radial geodesic is valid because

$$u_{\theta} = u_{\phi} = 0 \quad (10.65)$$

is a consistent solution. [Hint: The metric  $g_{\mu\nu}$  depends on the coordinate  $x_{\parallel}$ . But for radial geodesics with  $u^{\theta} = u^{\phi} = 0$ , 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 (10.64) that the proper momentum  $p_{\textcolor{brown}{x}_{\parallel}}$  of the particle relative to comoving observers (who are at rest in the FLRW metric) evolves as

$$p_{\textcolor{brown}{x}_{\parallel}} \equiv ma \frac{dx_{\parallel}}{d\tau} \propto \frac{1}{a}. \quad (10.66)$$

3. **Redshift.** What relation does your result (10.66) 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 (10.66) 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

$$T \propto \frac{1}{a} . \quad (10.67)$$


---

## 10.15 Evolution of the cosmic scale factor

Given how the energy density  $\rho$  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

$$t = \int \frac{da}{aH} . \quad (10.68)$$

The definition (10.46) of the critical density allows the Hubble parameter  $H$  to be written

$$\frac{H}{H_0} = \sqrt{\frac{\rho_{\text{crit}}}{\rho_{\text{crit}}(a_0)}} . \quad (10.69)$$

The critical density  $\rho_{\text{crit}}$  is itself the sum of the densities  $\rho$  of all species *including* the curvature density,

$$\rho_{\text{crit}} = \rho_k + \sum_{\text{species } x} \rho_x . \quad (10.70)$$

For example, in the case that the density is comprised of radiation, matter, and vacuum, the critical density is

$$\rho_{\text{crit}} = \rho_r + \rho_m + \rho_k + \rho_\Lambda , \quad (10.71)$$

and equation (10.69) is

$$\frac{H(t)}{H_0} = \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda} , \quad (10.72)$$

where  $\Omega_x$  represents its value at the present time. For density comprised of radiation, matter, and vacuum, equation (10.72), the time  $t$ , equation (10.68), is

$$t = \frac{1}{H_0} \int \frac{da}{a\sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda}} , \quad (10.73)$$

which is an elliptic integral of the third kind. The elliptic integral simplifies to elementary functions in some cases relevant to reality, Exercises 10.6 and 10.7.

If one single species in particular dominates the mass-energy density, then equation (10.73) integrates to give the results in Table 10.3.

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	$e^{Ht}$

## 10.16 Age of the Universe

The present age  $t_0$  of the Universe since the Big Bang can be derived from equation (10.73) and cosmological parameters, Table 10.1. Ade (2015) give the age of the Universe to be

$$t_0 = 13.80 \pm 0.02 \text{ Gyr} . \quad (10.74)$$

### Exercise 10.6 Age of a 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

$$t = \frac{2}{3H_0\sqrt{\Omega_\Lambda}} \operatorname{asinh} \sqrt{\frac{\Omega_\Lambda a^3}{\Omega_m}} . \quad (10.75)$$

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 10.7 Age of a 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

$$t = \frac{2\Omega_r^{3/2}}{3H_0\Omega_m^2} \frac{\hat{a}^2(2 + \sqrt{1 + \hat{a}})}{(1 + \sqrt{1 + \hat{a}})^2} , \quad (10.76)$$

where  $\hat{a}$  is the cosmic scale factor scaled to 1 at matter-radiation equality,

$$\hat{a} \equiv \frac{a}{a_{\text{eq}}} = \frac{\Omega_m a}{\Omega_r} . \quad (10.77)$$

You may well find a formula different from (10.76), 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 (10.76) 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**  $\eta$  defined by (with units  $c$  temporarily restored)

$$a d\eta \equiv c dt , \quad (10.78)$$

with respect to which the FLRW metric is

$$\boxed{ds^2 = a(\eta)^2 \left( -d\eta^2 + dx_{\parallel}^2 + x^2 do^2 \right)} , \quad (10.79)$$

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 (10.79), the cosmic scale factor  $a$  is the conformal factor.

Conformal time  $\eta$  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

$$\frac{dx_{\parallel}}{d\eta} = -1 . \quad (10.80)$$

---

**Exercise 10.8 Relation between conformal time and cosmic scale factor.** What is the relation between conformal time  $\eta$  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  $\eta$  is related to cosmic scale factor  $a$  by (units  $c = 1$ )

$$\eta = \int \frac{da}{a^2 H} . \quad (10.81)$$

For  $p/\rho = w = \text{constant}$ , a possible choice of integration constant for  $\eta$  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

$$\eta = \frac{2}{(1+3w)aH} \propto \pm a^{(1+3w)/2} , \quad (10.82)$$

in which the sign is positive for  $w > -1/3$ , negative for  $w < -1/3$ , ensuring that conformal time  $\eta$  always increases with cosmic scale factor  $a$ . For the special case of a curvature-dominated universe,  $w = -1/3$ ,

$$\eta = \frac{\ln a}{aH} \propto \ln a , \quad (10.83)$$

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  $\eta_{\text{em}}$  is observed at conformal time  $\eta_{\text{obs}}$  given by

$$x_{\parallel} = \eta_{\text{obs}} - \eta_{\text{em}} . \quad (10.84)$$

The comoving geodesic distance  $x_{\parallel}$  to an object is

$$x_{\parallel} = \int_{\eta_{\text{em}}}^{\eta_{\text{obs}}} d\eta = \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{c dt}{a} = \int_{a_{\text{em}}}^{a_{\text{obs}}} \frac{c da}{a^2 H} = \int_0^z \frac{c dz}{H} , \quad (10.85)$$

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 (10.85) gives

$$x_{\parallel} = \frac{c}{H_0} \int_{1/(1+z)}^1 \frac{da}{a^2 \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda}} , \quad (10.86)$$

which is an elliptical integral of the first kind. Given the geodesic comoving distance  $x_{\parallel}$ , the circumferential comoving distance  $x$  then follows as

$$x = \frac{\sinh(\sqrt{\Omega_k} H_0 x_{\parallel}/c)}{\sqrt{\Omega_k} H_0/c} . \quad (10.87)$$

To second order in redshift  $z$ ,

$$x \approx x_{\parallel} \approx \frac{c}{H_0} [z - z^2 (\Omega_r + \frac{3}{4}\Omega_m + \frac{1}{2}\Omega_k) + \dots] . \quad (10.88)$$

The geodesic and circumferential distances  $x_{\parallel}$  and  $x$  differ at order  $z^3$ .

Figure 10.8 illustrates the relation between the comoving geodesic and circumferential distances  $x_{\parallel}$  and  $x$ , equations (10.86) and (10.87), and redshift  $z$ , equation (10.60), in three cosmological models, including the standard flat  $\Lambda$ 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

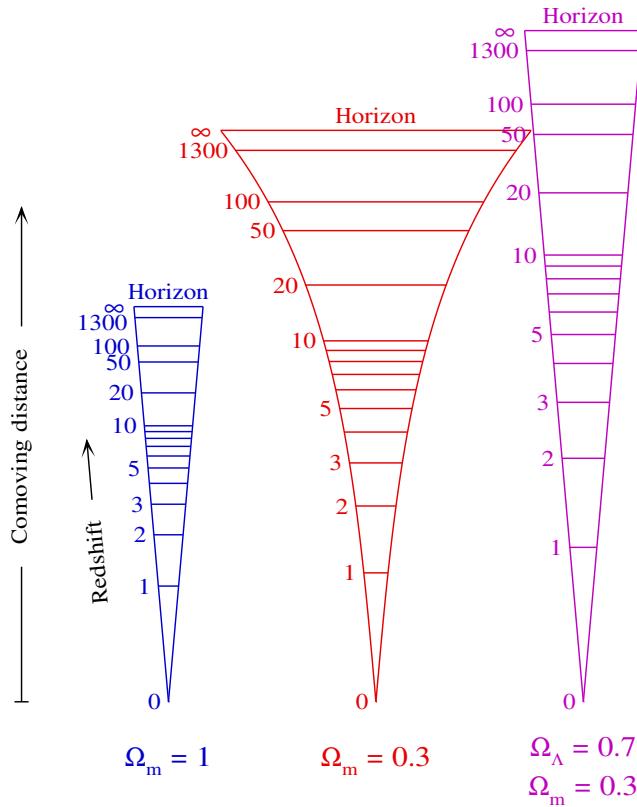


Figure 10.8 In this diagram, each wedge represents a cone of fixed opening angle, with the observer (us) at the point of the cone, at zero redshift. The wedges show the relation between physical sizes, namely the comoving distances  $x_{\parallel}$  in the radial (vertical) and  $x$  in the transverse (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  $\Lambda$ CDM universe.

the intrinsic luminosity  $L$  of the object (energy per unit time emitted by the object in its rest frame) divided by  $4\pi d_L^2$ ,

$$F = \frac{L}{4\pi d_L^2} . \quad (10.89)$$

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  $\lambda_{\text{em}}$  is observed at the redshifted wavelength  $\lambda_{\text{obs}} = (1 + z)\lambda_{\text{em}}$ . The luminosity distance (10.89) 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

$$F = \frac{L}{(1+z)^2 4\pi x^2} , \quad (10.90)$$

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 (10.90) 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 (10.89) and (10.90) imply that the luminosity distance  $d_L$  is related to the circumferential distance  $x$  and the redshift  $z$  by

$$d_L = (1+z)x . \quad (10.91)$$

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

$$m = m_{\text{Vega}} - 2.5 \log(F/F_{\text{Vega}}) . \quad (10.92)$$

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

$$m - M = 5 \log(d_L/10 \text{ pc}) , \quad (10.93)$$

where  $d_L$  is the luminosity distance. The difference  $m - M$  is called the **distance modulus**.

**Exercise 10.9 Hubble diagram.** Draw a theoretical Hubble diagram, a plot of luminosity distance  $d_L$  versus redshift  $z$ , for universes with various values of  $\Omega_\Lambda$  and  $\Omega_m$ . 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 3,000K 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. Ade (2015) give the redshift of last scattering, where the photon-electron scattering (Thomson) optical depth was 1,

$$z_* = 1089.9 \pm 0.2 . \quad (10.94)$$

Hinshaw et al. (2012, supplementary data) give the age of the Universe at recombination,

$$t_* = 376,000 \pm 4,000\text{ yr} . \quad (10.95)$$

## 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 (10.85) gives the geodesic distance between us at redshift zero and the horizon as

$$x_{||}(\text{horizon}) = \int_0^\infty \frac{c dz}{H} . \quad (10.96)$$

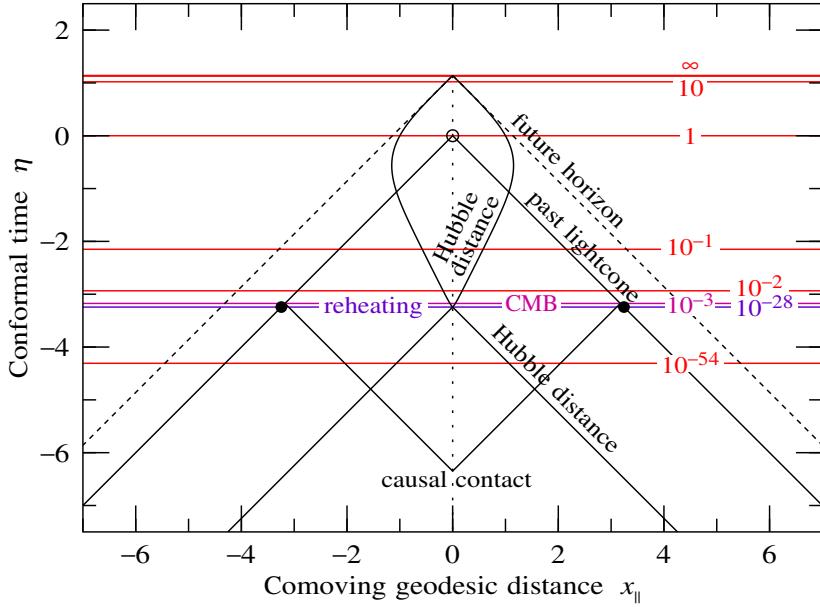


Figure 10.9 Spacetime diagram of a FLRW Universe in conformal coordinates  $\eta$  and  $x_{\parallel}$ , in units of the present day Hubble distance  $c/H_0$ . The unfilled circle marks our position, which is taken to be the origin of the conformal coordinates. In conformal coordinates, light moves at  $45^\circ$  on the spacetime diagram. The diagram is drawn for a flat  $\Lambda$ 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 Ade (2015). 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 standard  $\Lambda$ 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,

$$x_{\parallel}(\text{horizon}) = \int_0^{\text{huge}} \frac{c dz}{H} . \quad (10.97)$$

Figure 10.9 shows a spacetime diagram of a FLRW Universe with cosmological parameters consistent with those of Hinshaw et al. (2012). In this model, the comoving horizon distance to reheating is

$$x_{\parallel}(\text{horizon}) = 3.333 c/H_0 = 14.5 \text{ Gpc} = 4.72 \text{ Glyr} . \quad (10.98)$$

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.

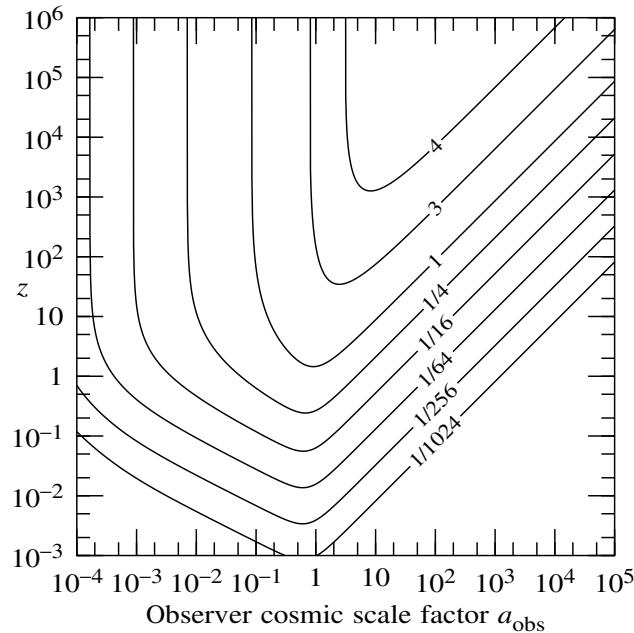


Figure 10.10 Redshift  $z$  of objects at fixed comoving distances as a function of the epoch  $a_{\text{obs}}$  at which an observer observes them. The label on each line is the comoving distance  $x_{\parallel}$  in units of  $c/H_0$ . The diagram is drawn for a flat  $\Lambda$ CDM model with  $\Omega_{\Lambda} = 0.7$ ,  $\Omega_m = 0.3$ , 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 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 10.9.

A quantity that cosmologists sometimes refer to loosely as the horizon is the **Hubble distance**, defined to be

$$\text{Hubble distance} \equiv \frac{c}{H} , \quad (10.99)$$

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  $\Lambda$ 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 10.10 illustrates the evolution of the observed redshifts of objects at fixed comoving distances. In the past decelerating phase, 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 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 at an energy of around  $m_{\text{GUT}} \sim 10^{16} \text{ GeV}$ , not much less than the Planck energy of  $m_P \sim 10^{19} \text{ GeV}$ . 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 10.11. 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 10.10. If that were the case, then how come the temperature at points in the cosmic microwave background more than a degree apart, indeed even  $180^\circ$  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, provided that the inflationary expansion continued over a sufficient number *e-folds*, Exercise 10.11. Guth’s solution is illustrated in the spacetime diagram in Figure 10.9.

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 10.12.

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 10.10 Horizon size at recombination.**

1. **Comoving horizon distance.** Assume for simplicity a flat, matter-dominated Universe. From equation (10.96), 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 10.11 The horizon problem.**

1. **Expansion factor.** The temperature of the CMB today is  $T_0 \approx 3$  K. By approximately what factor has the Universe expanded since the temperature was some initial high temperature, say the GUT temperature  $T_i \approx 10^{29}$  K, or the Planck temperature  $T_i \approx 10^{32}$  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  $x_H \propto a^?$ .]
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 10.12 Relation between horizon and flatness problems.** Show that Friedmann’s equa-

tion (10.30a) can be written in the form

$$\Omega - 1 = \kappa x_H^2 , \quad (10.100)$$

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 10.11 shows the evolution of the cosmic scale factor  $a$  as a function of time  $t$  predicted by the standard flat  $\Lambda$ CDM model, coupled with a plausible depiction of the early inflationary epoch. The parameters of the model are the same as those for Figure 10.9. In the model, the Universe starts with an inflationary phase, and transitions instantaneously at reheating to a radiation-dominated phases. 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 (10.68), and some relevant analytic results are in Exercises 10.6 and 10.7.

Figure 10.11 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 10.11 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 the number necessary to solve the horizon problem. Figure 10.11 does not attempt to extrapolate to what might possibly have happened prior to inflation, a problem that is the topic of much speculation by theoretical physicists.

Figure 10.12 shows the mass-energy density  $\rho$  as a function of time  $t$  for the same flat  $\Lambda$ CDM model as shown in Figure 10.11. Since the Universe here is taken to be flat, the density equals the critical density at all times, and is proportional to the inverse square of the Hubble distance  $c/H$  plotted in Figure 10.11. 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

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

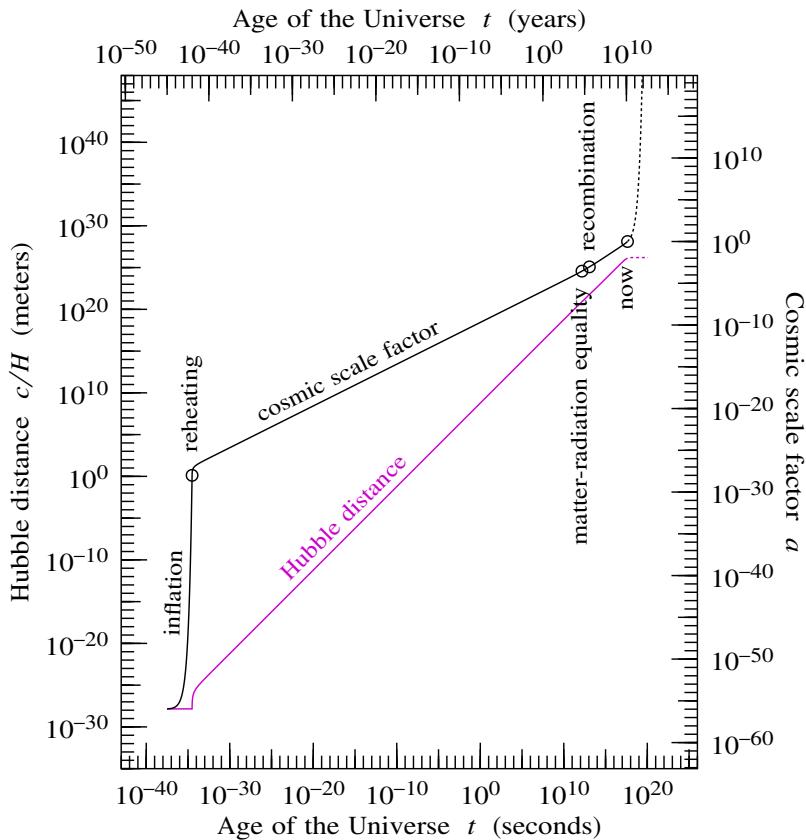


Figure 10.11 Cosmic scale factor  $a$  and Hubble distance  $c/H$  as a function of cosmic time  $t$ , for a flat  $\Lambda$ CDM model with the same parameters as in Figure 10.9. In this model, the Universe began with an inflationary epoch where the density was dominated by constant vacuum energy, the 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  $z_{\text{eq}} \approx 3200$  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  $z_{\text{rec}} \approx 1090$ . 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.

CMB has streamed more or less freely through the Universe since recombination. A thermal distribution of relativistic particles retains its thermal distribution in an expanding FLRW universe (albeit with a changing temperature), Exercise 10.13.

The evolution of the temperature of photons in the Universe can be deduced from conservation of entropy.

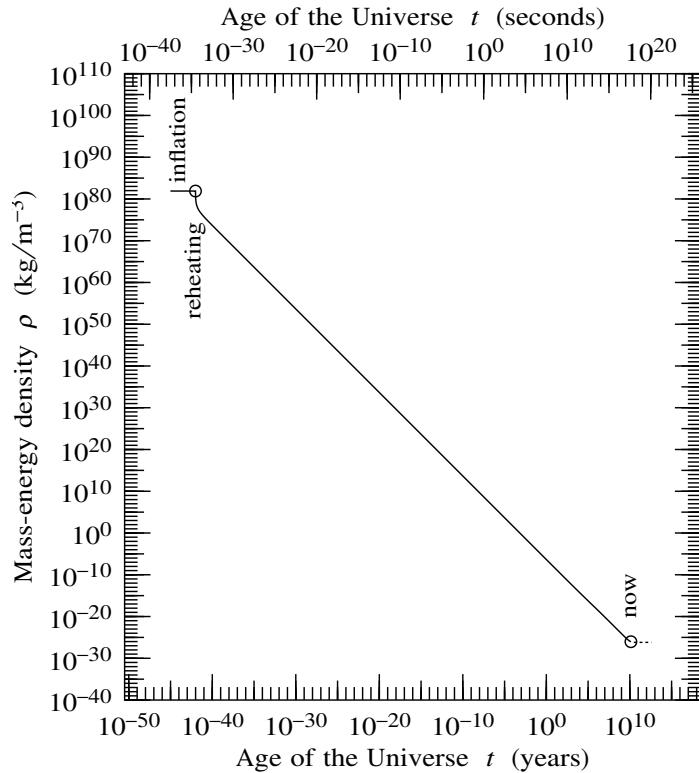


Figure 10.12 Mass-energy density  $\rho$  of the Universe as a function of cosmic time  $t$  corresponding to the evolution of the cosmic scale factor shown in Figure 10.11.

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 ??). Entropy is conserved in a 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, Exercise 10.14.

In the real Universe, entropy increases as a result of fluctuations away from the perfect homogeneity and 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 10.18. 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,

$$\eta_b \equiv \frac{n_b}{n_\gamma} = \frac{\epsilon_\gamma \Omega_b}{m_b \Omega_\gamma} = 6.01 \times 10^{-10} \frac{\Omega_b h^2}{0.022} \left( \frac{T_0}{2.725 \text{ K}} \right)^{-3}, \quad (10.101)$$

where  $\epsilon_\gamma = \pi^4 T_0 / (30\zeta(3)) = 2.701 T_0$  is the mean energy per photon (Exercise 10.15), and  $m_b = 939 \text{ MeV}$  is the approximate mass per baryon. The value  $\Omega_b h^2 = 0.02205 \pm 0.00003$  is as reported by the Planck team (Ade, 2014).

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 10.19)

$$\frac{T}{T_0} = (1+z) \left( \frac{g_{s,0}}{g_s} \right)^{1/3}, \quad (10.102)$$

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  $T \approx 1 \text{ MeV}$ . 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 non-relativistic 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  $m_\nu \lesssim 10^{-4} \text{ eV}$ , equation (10.108), 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 10.20,

$$T_\nu = \left( \frac{4}{11} \right)^{1/3} 2.725 \text{ K} = 1.945 \text{ K}. \quad (10.103)$$

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  $T \sim 1 \text{ MeV}$ .

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 10.16, equation (10.138). 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 (10.149),

$$g_{s,0} = 2 \left( 1 + \frac{7}{8} \frac{4}{11} 3 \right) = \frac{43}{11} = 3.91. \quad (10.104)$$

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 $\gamma$	1					1	$2 \left( 1 + \frac{7}{8} \frac{4}{11} 3 \right) = 3.91$
	neutrinos $\nu_e, \nu_\mu, \nu_\tau$	$\frac{1}{2}$	1	3			3	
$0.5 \text{ MeV} \lesssim T \lesssim 200 \text{ MeV}$	photon $\gamma$	1					1	$2 \left( 1 + \frac{7}{8} 5 \right) = 10.75$
	neutrinos $\nu_e, \nu_\mu, \nu_\tau$	$\frac{1}{2}$	1	3			3	
	electron $e$	$\frac{1}{2}$	2	1			2	
$200 \text{ MeV} \lesssim T \lesssim 100 \text{ GeV}$	photon $\gamma$	1					1	$2 \left( 9 + \frac{7}{8} 25 \right) = 61.75$
	$SU(3)$ gluons	1					8	
	neutrinos $\nu_e, \nu_\mu, \nu_\tau$	$\frac{1}{2}$	1	3			3	
	leptons $e, \mu$	$\frac{1}{2}$	2	2			4	
	quarks $u, d, s$	$\frac{1}{2}$	2	$\frac{3}{2}$	2	3	18	
$T \gtrsim 100 \text{ GeV}$	$SU(2) \times U_Y(1)$ gauge bosons	1					3 + 1	$2 \left( 14 + \frac{7}{8} 45 \right) = 106.75$
	$SU(3)$ gluons	1					8	
	complex Higgs	0					2	
	neutrinos $\nu_e, \nu_\mu, \nu_\tau$	$\frac{1}{2}$	1	3			3	
	leptons $e, \mu, \tau$	$\frac{1}{2}$	2	3			6	
	quarks $u, d, c, s, t, b$	$\frac{1}{2}$	2	3	2	3	36	

A more comprehensive evaluation of  $g_s$  is given by Kolb and Turner (1990, Fig. 3.5). Over the range of energies  $T \lesssim 1 \text{ TeV}$  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  $T \approx 0.5 \text{ MeV}$ , the QCD phase transition from bound nuclei to free quarks and gluons at  $T \approx 200 \text{ MeV}$ , and the electroweak phase transition above which all standard model particles are relativistic at  $T \approx 100 \text{ GeV}$ . There could well be further changes in the number of relativistic species at higher temperatures, for example if supersymmetry becomes unbroken at some energy, but at present no experimental data constrain the possibilities.

Figure 10.13 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 10.11 and 10.12.

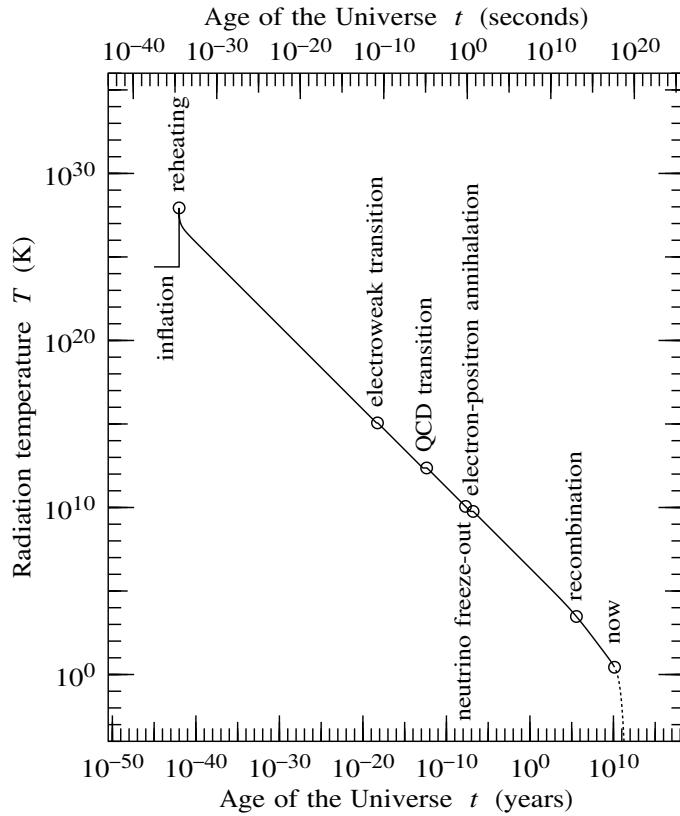


Figure 10.13 Radiation temperature  $T$  of the Universe as a function of cosmic time  $t$  corresponding to the evolution of the cosmic scale factor shown in Figure 10.11. The temperature during inflation was the Hawking temperature, equal to  $H/(2\pi)$  in Planck units. After inflation and reheating, 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 (10.102). In this plot, the effective number  $g_s$  of relativistic particle species has been approximated as changing at three discrete points, electron-positron annihilation, the QCD phase transition, and the electroweak phase transition, Table 10.4.

## 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  $\nu_e$ ,  $\nu_\mu$ , and  $\nu_\tau$ , while the mass eigenstates are denoted  $\nu_1$ ,  $\nu_2$ , and  $\nu_3$ . Oscillation data yield mass squared differences

between the three mass eigenstates (Forero, Tortola, and Valle, 2012)

$$|\Delta m_{21}|^2 = 7.6 \pm 0.2 \times 10^{-5} \text{ eV}^2 \quad \text{solar neutrinos ,} \quad (10.105\text{a})$$

$$|\Delta m_{31}|^2 = 2.4 \pm 0.1 \times 10^{-3} \text{ eV}^2 \quad \text{atmospheric neutrinos .} \quad (10.105\text{b})$$

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

$$m_{\nu_1} \text{ or } m_{\nu_2} \geq \sqrt{7.6 \times 10^{-5} \text{ eV}^2} \approx 0.01 \text{ eV .} \quad (10.106)$$

The squared mass difference between  $m_1$  and  $m_3$  implies that at least one of them must have a mass

$$m_{\nu_1} \text{ or } m_{\nu_3} \geq \sqrt{2.4 \times 10^{-3} \text{ eV}^2} \approx 0.05 \text{ eV .} \quad (10.107)$$

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.

The CNB temperature, equation (10.103), is  $T_\nu = 1.945 \text{ K} = 1.676 \times 10^{-4} \text{ eV}$ . The redshift at which a neutrino of mass  $m_\nu$  becomes non-relativistic is then

$$1 + z_\nu = \frac{m_\nu}{T_\nu} = \frac{m_\nu}{1.676 \times 10^{-4} \text{ eV}} . \quad (10.108)$$

Neutrinos of masses 0.01 eV and 0.05 eV would have become non-relativistic at  $z_\nu \approx 60$  and 300 respectively. Only a neutrino of mass  $\lesssim 10^{-4}$  eV would remain relativistic at the present time.

The masses from neutrino oscillation data suggest that at least two species of cosmological neutrinos are non-relativistic today. If so, then the neutrino density  $\Omega_\nu$  today is related to the sum  $\sum m_\nu$  of neutrino masses by

$$\Omega_\nu = \frac{8\pi G \sum m_\nu n_\nu}{3H_0^2} = 5.4 \times 10^{-4} \left( \frac{\sum m_\nu}{0.05 \text{ eV}} \right) h_{0.70}^{-2} . \quad (10.109)$$

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 non-relativistic, then the effective energy-weighted number  $g_{\rho,0}$  of relativistic species today is

$$g_{\rho,0} = 2 + \left( \frac{4}{11} \right)^{4/3} \frac{7}{8} 2 = 2.45 . \quad (10.110)$$

The density  $\Omega_r$  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 momentum point in the same direction is called right-handed, while a particle whose spin and momentum

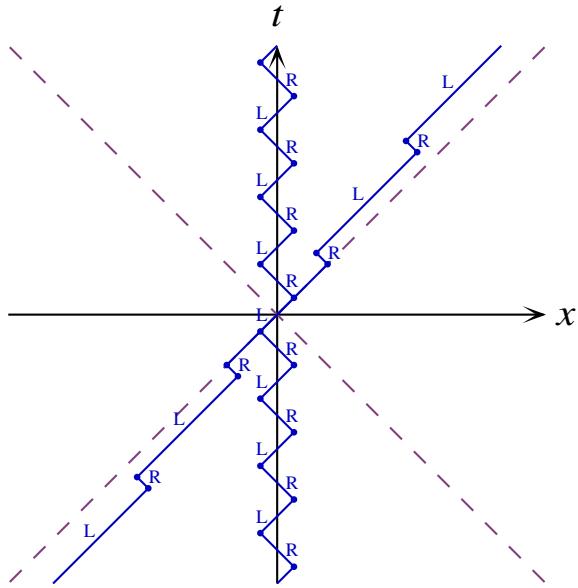


Figure 10.14 According to the Standard Model of physics, a massive Dirac fermion acquires its mass by interacting with the Higgs field. The interaction flips the fermion between left- and right-handed chiralities as it propagates through spacetime, as illustrated schematically in this spacetime diagram. 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.

point in opposite directions is called left-handed. The handedness of a particle is also called its chirality. Chirality is Lorentz-invariant: a 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 10.14.

## 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 a 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,

$$dN = f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3} , \quad (10.111)$$

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  $x^{\alpha}$  and their conjugate generalized momenta  $\pi_{\alpha}$ .

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

$$\frac{dt d^3r}{d\lambda} = E d^3r , \quad (10.112)$$

since  $dt/d\lambda = E$ . The left hand side of equation (10.112) 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  $\tau$  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

$$\delta(E^2 - p^2 c^2 - m^2 c^4) dE d^3p = \frac{d^3p}{2E} , \quad (10.113)$$

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 a FLRW universe

The homogeneity and isotropy of a FLRW universe imply that, for a comoving observer, the occupation number  $f$  is independent of position and direction,

$$f(t, \mathbf{x}, \mathbf{p}) = f(t, p) . \quad (10.114)$$

### 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  $n^k$ ,

$$n^k = \int p^k f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3 p}{E(2\pi\hbar)^3} . \quad (10.115)$$

In particular, the number density  $n^0$ , with units number of particles per unit proper volume, is the time component of the number current,

$$n^0 = \int f \frac{g d^3 p}{(2\pi\hbar)^3} . \quad (10.116)$$

In a 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,

$$n \equiv n^0 = \int f(t, p) \frac{4\pi p^2 dp}{(2\pi\hbar)^3} . \quad (10.117)$$

### 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

$$T^{kl} = \int p^k p^l f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3 p}{E(2\pi\hbar)^3} . \quad (10.118)$$

For a 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  $\rho$  of a particle species is

$$\rho = \int E f(t, p) \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} , \quad (10.119)$$

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)

$$p = \int \frac{p^2}{3E} f(t, p) \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} . \quad (10.120)$$

### 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

$$f = \begin{cases} \frac{1}{e^{(E-\mu)/T} + 1} & \text{fermion ,} \\ \frac{1}{e^{(E-\mu)/T} - 1} & \text{boson ,} \end{cases} \quad (10.121)$$

where  $\mu$  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

$$f = e^{(-E+\mu)/T} \quad \text{Boltzmann .} \quad (10.122)$$

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,

$$p + e \leftrightarrow H + \gamma , \quad (10.123)$$

separately preserves proton and electron number, hydrogen being composed of one proton and one electron. In thermodynamic equilibrium, the chemical potential  $\mu_H$  of hydrogen is the sum of the chemical potentials  $\mu_p$  and  $\mu_e$  of protons and electrons,

$$\mu_p + \mu_e = \mu_H . \quad (10.124)$$

Photon number is not conserved, so photons have zero chemical potential,

$$\mu_\gamma = 0 , \quad (10.125)$$

which is closely associated with the fact that photons are their own antiparticles. For photons, which are bosons, the thermodynamic distribution (10.121) becomes the Planck distribution,

$$f = \frac{1}{e^{E/T} - 1} \quad \text{Planck .} \quad (10.126)$$

**Exercise 10.13 Distribution of non-interacting particles initially in thermodynamic equilibrium.** The number  $dN$  of a particle species in an interval  $d^3rd^3p$  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

(non-relativistic, relativistic, or anything in between) in thermodynamic equilibrium at temperature  $T$  and chemical potential  $\mu$  is

$$dN = f \frac{g d^3 r d^3 p}{(2\pi\hbar)^3} , \quad (10.127)$$

where the occupation number  $f$  is (units  $k = 1$ , where  $k$  is the Boltzmann constant)

$$f = \frac{1}{e^{(E-\mu)/T} \pm 1} , \quad (10.128)$$

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  $\mu$  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  $d^3 r d^3 p$  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 a FLRW universe as long as the particles remain relativistic. How do the temperature  $T$  and chemical potential  $\mu$  of the relativistic distribution vary with cosmic scale factor  $a$ ?
3. **Non-relativistic particles.** Show similarly that a distribution of non-interacting non-relativistic particles initially in thermodynamic equilibrium remains thermal. How do the temperature  $T$  and chemical potential  $\mu - m$  of the non-relativistic distribution vary with cosmic scale factor  $a$ ?
4. **Transition from relativistic to non-relativistic.** What happens to a distribution of non-interacting particles that are relativistic in thermodynamic equilibrium, but redshift to being non-relativistic?

**Exercise 10.14 The first law of thermodynamics with non-conserved particle number.** As seen in §10.9.2, the first law of thermodynamics in the form

$$T d(a^3 s) = d(a^3 \rho) + p d(a^3) = 0 \quad (10.129)$$

is built into Friedmann's equations. But what happens when for example the temperature falls through the temperature  $T \approx 0.5$  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

$$T d(a^3 s) = d(a^3 \rho) + p d(a^3) - \sum_X \mu_X d(a^3 n_X) , \quad (10.130)$$

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  $\mu_X$  is associated with a conserved number, so the additional terms contribute zero change to the entropy,

$$\sum_X \mu_X d(a^3 n_X) = 0 , \quad (10.131)$$

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

$$\mu_e d(a^3 n_e) + \mu_{\bar{e}} d(a^3 n_{\bar{e}}) = \mu_e d(a^3 n_e - a^3 n_{\bar{e}}) = 0 , \quad (10.132)$$

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 (10.129) remains correct in a FLRW universe even when number changing processes are occurring.

**Exercise 10.15 Number, energy, pressure, and entropy of a relativistic ideal gas at zero chemical potential.** The number density  $n$ , energy density  $\rho$ , and pressure  $p$  of an ideal gas of a single species of free particles are given by equations (10.117), (10.119), and (10.120), with occupation number (10.128). 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  $\rho$ , and pressure  $p$  are (units  $k = 1$ ; number density  $n$  in units 1/volume, energy density  $\rho$  and pressure  $p$  in units energy/volume)

$$n = g \frac{\zeta(3)T^3}{\pi^2 c^3 \hbar^3} , \quad \rho = 3p = g \frac{\pi^2 T^4}{30 c^3 \hbar^3} , \quad (10.133)$$

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

$$s = \frac{\rho + p}{T} . \quad (10.134)$$

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)

$$s = g \frac{2\pi^2 T^3}{45 c^3 \hbar^3} . \quad (10.135)$$

**Exercise 10.16 A relation between thermodynamic integrals.** Prove that

$$\int_0^\infty \frac{x^{n-1} dx}{e^x + 1} = (1 - 2^{1-n}) \int_0^\infty \frac{x^{n-1} dx}{e^x - 1} . \quad (10.136)$$

[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

$$\frac{n_f}{n_b} = \frac{3}{4} , \quad \frac{\rho_f}{\rho_b} = \frac{s_f}{s_b} = \frac{7}{8} . \quad (10.137)$$

Conclude that if the number  $n$ , energy  $\rho$ , and entropy  $s$  of a mixture of bosonic and fermion species in thermodynamic equilibrium at the same temperature  $T$  are written in the form of equations (10.133) and (10.135), 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,

$$g_n = g_b + \frac{3}{4} g_f , \quad g_\rho = g_s = g_b + \frac{7}{8} g_f . \quad (10.138)$$

**Exercise 10.17 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

$$\mu_{\bar{X}} = -\mu_X . \quad (10.139)$$

If the particle-antiparticle asymmetry is denoted  $\eta$ , defined for relativistic particles by

$$n_X - n_{\bar{X}} = \eta n_X , \quad (10.140)$$

then  $\mu_X/T \sim \eta$ . More accurately, to linear order in  $\eta$ ,

$$\frac{\mu_X}{T} \approx \eta \frac{\pi^2}{3\zeta(3)} \times \begin{cases} 1 & (\text{bosons}) \\ \frac{2}{3} & (\text{fermions}) \end{cases} . \quad (10.141)$$

**Exercise 10.18 Entropy per particle.** The entropy of an ideal gas of free particles in thermodynamic equilibrium is

$$s = \frac{\rho + p - \mu n}{T} . \quad (10.142)$$

Argue that the entropy per particle  $s/n$  is a quantity of order unity, whether particles are relativistic or non-relativistic.

**Solution.** For relativistic bosons with zero chemical potential, equations (10.133) and (10.135) imply that the entropy per particle is

$$\frac{s}{n} = \frac{2\pi^4}{45\zeta(3)} \times \begin{cases} 1 & = 3.6 \quad (\text{bosons}) \\ \frac{7}{6} & = 4.2 \quad (\text{fermions}) \end{cases} . \quad (10.143)$$

For a non-relativistic species, the number density  $n$  is related to the temperature  $T$  and chemical potential  $\mu$  by

$$n = g \left( \frac{mT}{2\pi\hbar^2} \right)^{3/2} e^{(\mu-m)/T} . \quad (10.144)$$

Under cosmological conditions, the occupation number of non-relativistic 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 non-relativistic species is

$$\frac{s}{n} = \frac{5}{2} + \ln \left[ \frac{g}{n} \left( \frac{mT}{2\pi\hbar^2} \right)^{3/2} \right] , \quad (10.145)$$

which remains modest even if the argument of the logarithm is huge.

**Exercise 10.19 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 (10.102).

**Exercise 10.20 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 (10.106) and (10.107). The oscillation data do not constrain the offset from zero mass. A neutrino of mass  $\lesssim 10^{-4} \text{ eV}$  would remain relativistic at the present time, equation (10.108), and would produce a Cosmic Neutrino Background. Neutrinos that are non-relativistic 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

$$a^3 T^3 \left( g_\gamma + \frac{7}{8} g_e \right) = T_\gamma^3 g_\gamma , \quad (10.146a)$$

$$a^3 T^3 g_\nu = T_\nu^3 g_\nu , \quad (10.146b)$$

where the left hand sides refer to quantities before  $e\bar{e}$  annihilation, which happened at  $T \sim 0.5 \text{ MeV}$ , and the right hand sides to quantities after  $e\bar{e}$  annihilation (including today). Deduce the ratio of neutrino to photon temperatures today,

$$\frac{T_\nu}{T_\gamma} . \quad (10.147)$$

Does the temperature ratio (10.147) 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 (10.138). 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 non-relativistic?

**Solution.** The ratio of neutrino to photon temperatures after  $e\bar{e}$  annihilation is

$$\frac{T_\nu}{T_\gamma} = \left( \frac{g_\gamma}{g_\gamma + \frac{7}{8} g_e} \right)^{1/3} = \left( \frac{4}{11} \right)^{1/3} = 0.714 . \quad (10.148)$$

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 (10.148) implies that the CNB temperature is given by equation (10.103). 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

$$g_{n,0} = g_\gamma + \left( \frac{T_\nu}{T_\gamma} \right)^3 \frac{3}{4} g_\nu = 2 + \frac{4}{11} \frac{3}{4} 6 = \frac{40}{11} = 3.64 , \quad (10.149a)$$

$$g_{\rho,0} = g_\gamma + \left( \frac{T_\nu}{T_\gamma} \right)^4 \frac{7}{8} g_\nu = 2 + \left( \frac{4}{11} \right)^{4/3} \frac{7}{8} 6 = 3.36 , \quad (10.149b)$$

$$g_{s,0} = g_\gamma + \left( \frac{T_\nu}{T_\gamma} \right)^3 \frac{7}{8} g_\nu = 2 + \frac{4}{11} \frac{7}{8} 6 = \frac{43}{11} = 3.91 . \quad (10.149c)$$

Neutrinos today interact too weakly to annihilate, so their number and entropy today is that of relativistic species even if they are non-relativistic today. However, their energy density today is not that of relativistic particles.

## 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 10.21, 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

$$ds^2 = - \left( 1 - \frac{1}{3} \Lambda r_s^2 \right) dt_s^2 + \frac{dr_s^2}{1 - \frac{1}{3} \Lambda r_s^2} + r_s^2 d\Omega^2 , \quad (10.150)$$

where the time  $t_s$  and radius  $r_s$  are subscripted s for stationary to distinguish them from FLRW time  $t$  and radius  $r$ .

Spacetimes that are homogeneous, isotropic, and stationary, and are therefore described by the metric (10.150), 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  $r_H = \sqrt{3/\Lambda}$ . 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

$$R_{\kappa\lambda\mu\nu} = \frac{1}{3} \Lambda (g_{\kappa\mu} g_{\lambda\nu} - g_{\kappa\nu} g_{\lambda\mu}) , \quad R_{\kappa\mu} = \Lambda g_{\kappa\mu} , \quad R = 4\Lambda , \quad G_{\kappa\mu} = -\Lambda g_{\kappa\mu} . \quad (10.151)$$

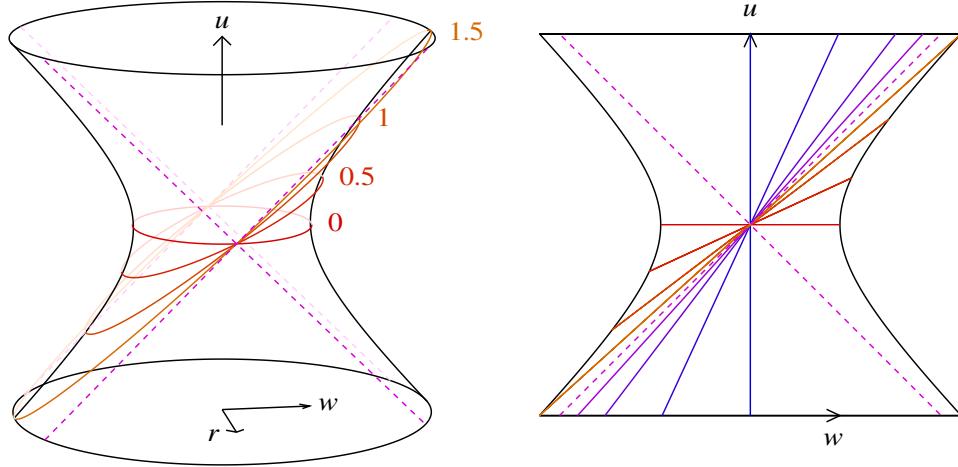


Figure 10.15 Embedding spacetime diagram of de Sitter space, shown on the left in 3D, on the right in a 2D projection on to the  $u$ - $w$  plane. Objects are confined to the surface of the embedded hyperboloid. The vertical direction is timelike, while the horizontal directions are spacelike. The position 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. Antipodal 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  $u$ - $w$  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).

### 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

$$-u^2 + x^2 + y^2 + z^2 + w^2 = r_H^2 = \text{constant} , \quad (10.152)$$

with  $r_H$  the horizon radius

$$r_H = \sqrt{\frac{3}{\Lambda}} . \quad (10.153)$$

The de Sitter hyperboloid is illustrated in Figure 10.15. Let  $r \equiv (x^2 + y^2 + z^2)^{1/2}$ , and introduce the boost

angle  $\psi$  and rotation angle  $\chi$  defined by

$$u = r_H \sinh \psi , \quad (10.154a)$$

$$r = r_H \cosh \psi \sin \chi , \quad (10.154b)$$

$$w = r_H \cosh \psi \cos \chi . \quad (10.154c)$$

The radius  $r$  defined by equation (10.154b) is the same as the radius  $r_s$  in the stationary metric (10.150). In terms of the angles  $\psi$  and  $\chi$ , the metric on the de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 [-d\psi^2 + \cosh^2 \psi (d\chi^2 + \sin^2 \chi do^2)] . \quad (10.155)$$

The metric (10.155) is of FLRW form (10.28) with  $t = r_H \psi$  and  $a(t) = \kappa^{1/2} r_H \cosh \psi$ , 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 10.15.

The spatial position at  $r = 0$  and  $w > 0$  defines a “north pole” of de Sitter space. Antipodal 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  $u-w$  plane. The time  $t_s$  in the stationary metric (10.150) is, modulo a factor of  $r_H$ , the boost angle of this Lorentz boost, which is

$$t_s = \begin{cases} r_H \operatorname{atanh}(u/w) & |w| > |u| , \\ r_H \operatorname{atanh}(w/u) & |w| < |u| . \end{cases} \quad (10.156)$$

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 (10.152) 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  $\psi$  and  $\chi$  by

$$u = r_H \sinh \psi \cosh \chi , \quad (10.157a)$$

$$r = r_H \sinh \psi \sinh \chi , \quad (10.157b)$$

$$w = r_H \cosh \psi . \quad (10.157c)$$

The  $r$  defined by equation (10.157b) is *not* the same as the  $r_s$  in the stationary metric (10.150); rather, it is  $w$  that equals  $r_s$ . In terms of the angles  $\psi$  and  $\chi$  defined by equations (10.157), the metric on the de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 [-d\psi^2 + \sinh^2 \psi (d\chi^2 + \sinh^2 \chi do^2)] . \quad (10.158)$$

The metric (10.158) is in FLRW form (10.28) with  $t = r_H \psi$  and  $a(t) = (-\kappa)^{1/2} r_H \sinh \psi$ , and an open spatial geometry. Whereas the coordinates  $\{\psi, \chi\}$ , equation (10.154), for de Sitter with closed spatial geometry cover the entire embedded hyperboloid shown in Figure 10.15, the coordinates  $\{\psi, \chi\}$ , equation (10.157), 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  $u$ - $r$  plane, bounded by the null cone at  $u = |r|$ .

In the open FLRW metric (10.158) for de Sitter space, the coordinates  $t_s$  and  $r_s$  of the stationary metric (10.150) are respectively spacelike and timelike. Lines of constant stationary time  $t_s$ , equation (10.156), coincide with geodesics of comoving observers, at constant  $\chi$ , while lines of constant stationary radius  $r_s = w$ , equation (10.157c), coincide with lines of constant FLRW time  $\psi$ ,

$$t_s/r_H = \chi , \quad (10.159a)$$

$$r_s/r_H = w/r_H = \cosh \psi . \quad (10.159b)$$

### 10.28.3 Penrose diagram of de Sitter space

Figure 10.16 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 10.15 become vertical lines on the Penrose diagram. These vertical lines are geodesics for comoving observers, lines of constant  $\chi$ , in the closed FLRW form (10.155) form of the de Sitter metric. The corresponding Penrose time coordinate  $t_P$  follows from solving for the radial null geodesics of the metric (10.155), whence  $t_P = \int d\psi / \cosh \psi$ . The resulting Penrose coordinates for de Sitter space are

$$t_P = \text{atan}(\sinh \psi) = \text{atan}(u/r_H) , \quad (10.160a)$$

$$r_P = \chi = \text{atan}(r/w) . \quad (10.160b)$$

The radial coordinate  $r$  in both the closed and open FLRW forms (10.155) and (10.158) 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 10.16 depicts both closed and open FLRW geometries, but the open geometry is shifted by  $90^\circ$  to the equator, so that it appears to interleave with the closed geometry instead of overlapping it. The thick (pink) null lines at  $45^\circ$  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

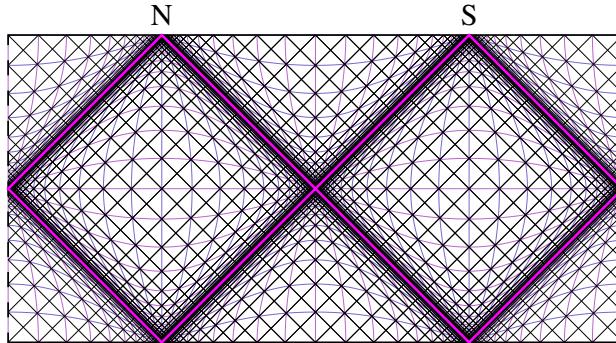


Figure 10.16 Penrose diagram of de Sitter space. The left and right edges are identified. The topology is that of a 3-sphere in the horizontal (spatial) direction times the real line in the vertical (time) direction. The thick (pink) null lines are past and future horizons for observers who follow (vertical) 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 (10.150) 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 (10.171). 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 (10.158) 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.

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 10.16 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

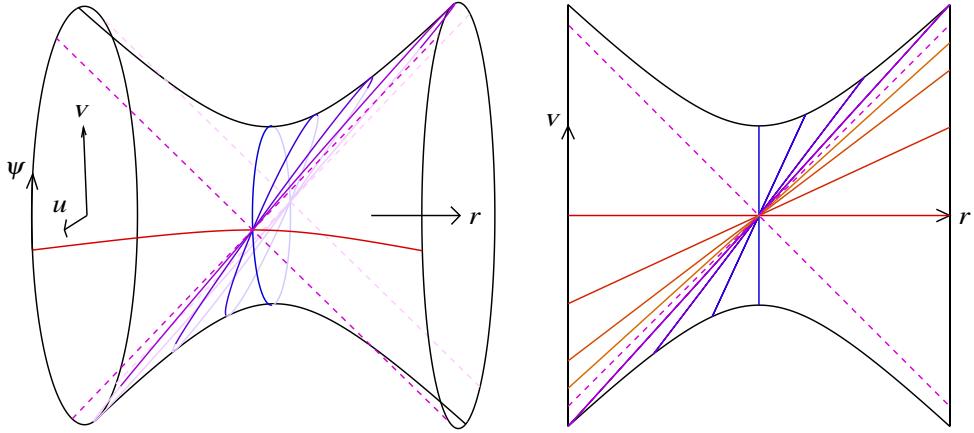


Figure 10.17 Embedding spacetime diagram of anti de Sitter space, shown on the left in 3D, on the right in a 2D projection on to the  $v$ - $r$  plane. The vertical direction winding around the hyperboloid is timelike, while the horizontal direction is spacelike. The position of a non-accelerating observer 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  $u$ - $v$  plane. The (bluish) lines at less than  $45^\circ$  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  $\pi r_H$ . The 2D diagram on the right shows in addition (reddish) lines of constant stationary time  $t_s$  for observers on the various geodesics.

coordinate  $r_s^*$

$$r_s^* \equiv \int \frac{dr_s}{1 - r_s^2/r_H^2} = r_H \operatorname{atanh}(r_s/r_H). \quad (10.161)$$

The contours in the Penrose diagram 10.16 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

$$t_P \pm r_P = \operatorname{atan} \left[ \sinh \left( \frac{t_s \pm r_s^*}{r_H} \right) \right]. \quad (10.162)$$

#### 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

$$-u^2 - v^2 + x^2 + y^2 + z^2 = -r_H^2 = \text{constant}, \quad (10.163)$$

with  $r_H \equiv \sqrt{-3/\Lambda}$ . The anti de Sitter hyperboloid is illustrated in Figure 10.17. Let  $r \equiv (x^2 + y^2 + z^2)^{1/2}$ , and introduce the boost angle  $\chi$  and rotation angle  $\psi$  defined by

$$u = r_H \cosh \chi \cos \psi , \quad (10.164a)$$

$$v = r_H \cosh \chi \sin \psi , \quad (10.164b)$$

$$r = r_H \sinh \chi . \quad (10.164c)$$

The time coordinate  $\psi$  defined by equations (10.164) appears to be periodic, with period  $2\pi$ , but this is an artefact of the embedding. In a causal spacetime, the time coordinate would not loop back on itself. Rather, the coordinate  $\psi$  can be taken to increase monotonically as it loops around the hyperboloid, extending from  $-\infty$  to  $\infty$ . In terms of the angles  $\psi$  and  $\chi$ , the metric on the anti de Sitter 4D hyperboloid is

$$ds^2 = r_H^2 (-\cosh^2 \chi d\psi^2 + d\chi^2 + \sinh^2 \chi do^2) . \quad (10.165)$$

The metric (10.165) is of stationary form (10.150) 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 10.17 for anti de Sitter space,

$$u = r_H \cos \psi , \quad (10.166a)$$

$$v = r_H \sin \psi \cosh \chi , \quad (10.166b)$$

$$r = r_H \sin \psi \sinh \chi . \quad (10.166c)$$

yields a metric in FLRW form with an open spatial geometry,

$$ds^2 = r_H^2 [-d\psi^2 + \sin^2 \psi (d\chi^2 + \sinh^2 \chi do^2)] . \quad (10.167)$$

Whereas the coordinates (10.164) cover all of the anti de Sitter hyperboloid 10.17, the open coordinates (10.166) 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 10.17. 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  $v$ - $x$  plane. In the open FLRW form (10.167) 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

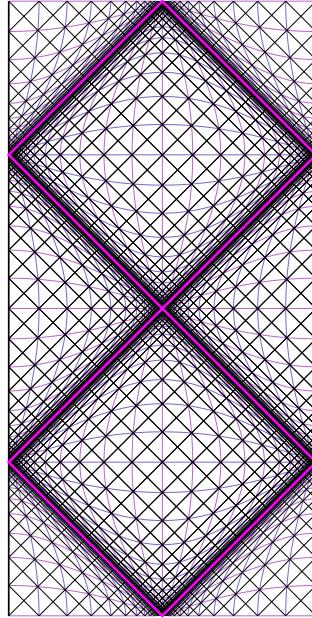


Figure 10.18 Penrose diagram of anti de Sitter space. The diagram repeats vertically indefinitely. The topology is that of Euclidean 3-space in the horizontal (spatial) direction times the real line in the vertical (time) direction. The thick (pink) null lines outline the causal diamonds of observers in the open FLRW form (10.167) 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 (10.169), at the left and right edges of the diagram. The approximately horizontal and vertical contours are lines of constant  $\psi$  and  $\chi$ , uniformly spaced by 0.4 in  $\chi$  and  $\psi^*$ , equation (10.172), both in the open FLRW diamonds and in the left and right Rindler wedges, equations (10.167) and (10.169). The coordinates  $\psi$  and  $\chi$  are respectively timelike (vertical) and spacelike (horizontal) in the open diamonds, and respectively spacelike and timelike in the Rindler wedges.

the  $v-x$  plane. Rindler time and space coordinates  $\{\chi, \psi\}$  are defined by

$$u = r_H \cosh \psi , \quad (10.168a)$$

$$v = r_H \sinh \psi \sinh \chi , \quad (10.168b)$$

$$x = r_H \sinh \psi \cosh \chi , \quad (10.168c)$$

yielding the AdS Rindler metric

$$ds^2 = r_H^2 (-\sinh^2 \psi d\chi^2 + d\psi^2) + dy^2 + dz^2 . \quad (10.169)$$

### 10.28.7 Penrose diagram of anti de Sitter space

Figure 10.18 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 10.17 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 (10.165) 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 (10.165), whence  $r_P = \int d\chi / \cosh \chi$ . The resulting Penrose coordinates for de Sitter space are

$$t_P = \psi = t_s/r_H , \quad (10.170a)$$

$$r_P = \text{atan}(\sinh \chi) = r_s^*/r_H , \quad (10.170b)$$

where  $r_s^*$  is the tortoise radial coordinate

$$r_s^* \equiv \int \frac{dr_s}{1 + r_s^2/r_H^2} = r_H \text{atan}(r_s/r_H) . \quad (10.171)$$

Thus, for anti de Sitter, lines of constant time  $t_s$  and radius  $r_s$  in the stationary metric (10.150) correspond also to lines of constant Penrose time and radius  $t_P$  and  $r_P$ .

The thick (pink) null lines in the Penrose diagram 10.18 outline the causal diamonds of comoving observers in the open FLRW (10.167) form of the anti de Sitter metric. The null lines also outline the causal wedges of Rindler observers in the Rindler (10.169) form of the anti de Sitter metric.

The approximately horizontal and vertical contours in the Penrose diagram 10.18 are lines of constant  $\psi$  and  $\chi$  in both the open FLRW (10.167) and Rindler (10.169) forms of the anti de Sitter metric. In the open FLRW causal diamonds, the horizontal lines are lines of constant cosmic time  $\psi$ , while the vertical contours are geodesics, lines of constant  $\chi$ . In the Rindler causal wedges, the horizontal contours are lines of constant boost angle  $\chi$ , while the vertical contours are worldlines of Rindler observers, lines of constant  $\psi$ .

Anti de Sitter space is symmetric under boosts in the  $v$ - $x$  plane, corresponding to translations of the coordinate  $\chi$  in either of the open FLRW (10.167) or Rindler (10.169) forms of the anti de Sitter metric. The contours in the Penrose diagram 10.18 are uniformly spaced in  $\chi$  by 0.4 so as to manifest this symmetry. A similarly symmetric appearance for the  $\psi$  coordinate is achieved by choosing contours to be uniformly spaced by 0.4 in the tortoise coordinate  $\psi^*$

$$\psi^* \equiv \begin{cases} \int \frac{d\psi}{\sin \psi} = \ln \tan(\psi/2) & \text{open ,} \\ \int \frac{d\psi}{\sinh \psi} = \ln \tanh(\psi/2) & \text{Rindler .} \end{cases} \quad (10.172)$$

#### Exercise 10.21 Maximally symmetric spaces.

- 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  $\rho$  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

$$a(t) = \begin{cases} H_0 t & \Omega_k = 1, \Omega_\Lambda = 0, \\ \exp(H_0 t) & \Omega_k = 0, \Omega_\Lambda = 1, \\ \sqrt{-\Omega_k/\Omega_\Lambda} \cosh(\sqrt{\Omega_\Lambda} H_0 t) & \Omega_k < 0, \Omega_\Lambda > 0, \\ \sqrt{\Omega_k/\Omega_\Lambda} \sinh(\sqrt{\Omega_\Lambda} H_0 t) & \Omega_k > 0, \Omega_\Lambda > 0, \\ \sqrt{-\Omega_k/\Omega_\Lambda} \sin(\sqrt{-\Omega_\Lambda} H_0 t) & \Omega_k > 0, \Omega_\Lambda < 0, \end{cases} \quad (10.173)$$

with

$$\Omega_\Lambda H_0^2 = \frac{1}{3}\Lambda, \quad \Omega_k H_0^2 = -\kappa. \quad (10.174)$$

As elsewhere in this chapter,  $H = H_0$  at  $a = 1$ , and the  $\Omega$ 's sum to unity,  $\Omega_k + \Omega_\Lambda = 1$ .

3. Show that the FLRW metric transforms into the explicitly stationary form (10.150) under a coordinate transformation to proper radius  $r_s = a(t)x$  and stationary time  $t_s$  given by

$$t_s = \begin{cases} \sqrt{1 - \kappa x^2} t & \Omega_k = 1, \Omega_\Lambda = 0, \\ t - \frac{1}{H_0} \ln \sqrt{1 - H_0^2 r_s^2} & \Omega_k = 0, \Omega_\Lambda = 1, \\ \frac{1}{\sqrt{\Omega_\Lambda} H_0} \operatorname{acoth} [\sqrt{1 - \kappa x^2} \coth (\sqrt{\Omega_\Lambda} H_0 t)] & \Omega_k < 0, \Omega_\Lambda > 0, \\ \frac{1}{\sqrt{\Omega_\Lambda} H_0} \operatorname{atanh} [\sqrt{1 - \kappa x^2} \tanh (\sqrt{\Omega_\Lambda} H_0 t)] & \Omega_k > 0, \Omega_\Lambda > 0, \\ \frac{1}{\sqrt{-\Omega_\Lambda} H_0} \operatorname{atan} [\sqrt{1 - \kappa x^2} \tan (\sqrt{-\Omega_\Lambda} H_0 t)] & \Omega_k > 0, \Omega_\Lambda < 0. \end{cases} \quad (10.175)$$

Note that in all cases  $t_s = t$  at the origin  $r_s = 0$ .

**Concept question 10.22 Milne Universe.** In Exercise 10.21 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 10.23 Stationary FLRW metrics with different curvature constants describe the same spacetime.** How can it be that stationary FLRW metrics with different curvature constants  $\kappa$  (but the same cosmological constant  $\Lambda$ ) 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  $\partial_m$  mean physically?
6. Is the directed derivative  $\partial_m$  coordinate gauge-invariant?
7. Is the tetrad metric  $\gamma_{mn}$  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  $\gamma_m$  change with time?  $\partial_0 \gamma_m = 0$ ?  $D_0 \gamma_m = 0$ ?
  - c. Does the tetrad metric  $\gamma_{mn}$  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  $\mathbf{p} = \boldsymbol{\gamma}_m p^m$  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  $p^m \equiv \boldsymbol{\gamma}^m \cdot \mathbf{p}$  where  $\boldsymbol{\gamma}^m$  are the contravariant locally inertial rest axes of the observer. Similarly the proper covariant components are  $p_m \equiv \boldsymbol{\gamma}_m \cdot \mathbf{p}$ .]
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  $\delta p^m$  in the two 4-vectors detected by the two eyes to infer the binocular distance to the object. Is the difference  $\delta p^m$  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  $\Gamma_{kmn}$  mean?
  15. What is the physical significance of the fact that  $\Gamma_{kmn}$  is antisymmetric in its first two indices (if the tetrad metric  $\gamma_{mn}$  is constant)?
  16. Are the tetrad connections  $\Gamma_{kmn}$  coordinate gauge-invariant?

---

## What's important?

This part of the notes 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 (inverse) 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  $\partial_m$  is the analogue of the coordinate partial derivative  $\partial/\partial x^{\mu}$  of the coordinate formalism. Directed derivatives  $\partial_m$  do not commute, whereas coordinate derivatives  $\partial/\partial x^{\mu}$  do commute.

# 11

---

## The tetrad formalism

### 11.1 Tetrad

A **tetrad** (greek foursome)  $\gamma_m(x)$  is a set of axes

$$\boxed{\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}} \quad (11.1)$$

attached to each point  $x^\mu$  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  $\eta_{mn}$

$$\gamma_m \cdot \gamma_n = \eta_{mn} . \quad (11.2)$$

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

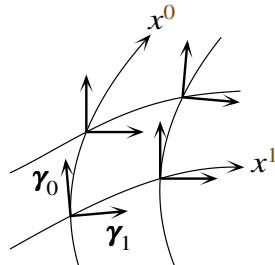


Figure 11.1 Tetrad vectors  $\gamma_m$  form a basis of vectors at each point. A common choice, depicted here, is for the basis vectors  $\gamma_m$  to form an orthonormal set, meaning that their dot products constitute the Minkowski metric,  $\gamma_m \cdot \gamma_n = \eta_{mn}$  at each point. The 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.

$e_\mu$  is a tetrad). In general, the tetrad metric is some symmetric matrix  $\gamma_{mn}$

$$\boxed{\boldsymbol{\gamma}_m \cdot \boldsymbol{\gamma}_n \equiv \gamma_{mn}} . \quad (11.3)$$

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  $\mu$ )

$$\boxed{\boldsymbol{\gamma}_m = e_m^\mu \boldsymbol{e}_\mu} . \quad (11.4)$$

The letter  $e$  stems from the German word einheit for unity. The vierbein is a  $4 \times 4$  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

$$e^m_\mu e_m^\nu = \delta_\mu^\nu , \quad e^m_\mu e_n^\mu = \delta_m^n . \quad (11.5)$$

Thus equation (11.4) inverts to

$$\boxed{\boldsymbol{e}_\mu = e^m_\mu \boldsymbol{\gamma}_m} . \quad (11.6)$$

## 11.3 The metric 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 metrics are commonly written encodes not only a metric but also an inverse vierbein, hence a tetrad. For example, the Schwarzschild metric

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \quad (11.9)$$

takes the form (11.7) with an orthonormal (Minkowski) tetrad metric  $\gamma_{mn} = \eta_{mn}$ , and an inverse 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 inverse vierbein of the Schwarzschild metric 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 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 (11.12) 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  $\gamma_k$  at each point by a Lorentz transformation  $L_k^m$ , while keeping the background coordinates  $x^{\mu}$  unchanged:

$$\boxed{\gamma_k \rightarrow \gamma'_k = L_k^m \gamma_m} . \quad (11.13)$$

In the case that the tetrad axes  $\gamma_k$  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  $\gamma_{mn}$  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  $\gamma_{kl}$  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  $\Delta L_k^m$ , 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  $\gamma_{kl}$ , equation (11.14). It follows that

$$\Delta L_{kl} + \Delta L_{lk} = 0 , \quad (11.17)$$

that is, the generator  $\Delta L_{kl}$  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  $\gamma_k$  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  $\gamma_{mn}$  and its inverse  $\gamma^{mn}$ , 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

$$L_k{}^m L^l{}_m = L_{km} L^{lm} = L_{mk} L^{ml} = L^m{}_k L_m{}^l = \delta_k^l . \quad (11.25)$$

### 11.5.4 Abstract vector

A 4-vector can be written in a coordinate- and tetrad- independent fashion as an abstract 4-vector  $\mathbf{A}$ ,

$$\mathbf{A} = \boldsymbol{\gamma}_m A^m = e_\mu A^\mu . \quad (11.26)$$

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 tetrad and coordinate components of the 4-vector  $\mathbf{A}$  are related by the vierbein,

$$A_m = e_m^{\mu} A_{\mu}, \quad A_{\mu} = e^m_{\mu} A_m. \quad (11.27)$$

### 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**  $A_{mn...}^{kl...}$  is an object that transforms under tetrad (Lorentz) transformations (11.13) as

$$A'_{mn...}^{kl...} = L^k_a L^l_b \dots L_m^c L_n^d \dots A_{cd...}^{ab...}. \quad (11.29)$$

## 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**  $A^{\mu}$ , 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  $x^{\mu}$  of the spacetime without actually changing the spacetime or whatever lies in it. A coordinate vector  $A^{\mu}$  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  $\gamma_m$  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  $\gamma_{mn}$  and its inverse  $\gamma^{mn}$ ;
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**  $\partial_m$  are defined to be the directional derivatives along the axes  $\gamma_m$

$$\boxed{\partial_m \equiv \gamma_m \cdot \partial = \gamma_m \cdot e^\mu \frac{\partial}{\partial x^\mu} = e_m{}^\mu \frac{\partial}{\partial x^\mu}} \quad \text{a tetrad 4-vector .} \quad (11.30)$$

The directed derivative  $\partial_m$  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  $\partial_m$  do not commute. Their commutator is

$$\begin{aligned} [\partial_m, \partial_n] &= \left[ e_m{}^\mu \frac{\partial}{\partial x^\mu}, e_n{}^\nu \frac{\partial}{\partial x^\nu} \right] \\ &= e_m{}^\mu \frac{\partial e_n{}^\nu}{\partial x^\mu} \frac{\partial}{\partial x^\nu} - e_n{}^\nu \frac{\partial e_m{}^\mu}{\partial x^\nu} \frac{\partial}{\partial x^\mu} \\ &= (-d_{nm}^k + d_{mn}^k) \partial_k \quad \text{not a tetrad tensor} \end{aligned} \quad (11.31)$$

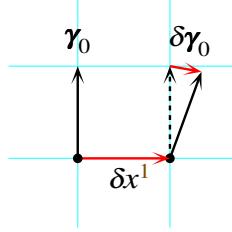


Figure 11.2 The change  $\delta\gamma_0$  in the tetrad vector  $\gamma_0$  over a small coordinate interval  $\delta x^1$  of spacetime is defined to be the difference between the tetrad vector  $\gamma_0(x^1 + \delta x^1)$  at the shifted position  $x^1 + \delta x^1$  and the tetrad vector  $\gamma_0(x^1)$  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.

where  $d_{lmn} \equiv \gamma_{lk} d_m^k$  is the vierbein derivative

$$d_{lmn} \equiv -\gamma_{lk} e^k_\mu e_n^\nu \frac{\partial e_m^\kappa}{\partial x^\nu} \quad \text{not a tetrad tensor .} \quad (11.32)$$

Since the vierbein and inverse vierbein are inverse to each other, an equivalent definition of  $d_{lmn}$  in terms of the inverse vierbein is

$$d_{lmn} \equiv \gamma_{lk} e_m^\mu e_n^\nu \frac{\partial e^k_\mu}{\partial x^\nu} \quad \text{not a tetrad tensor .} \quad (11.33)$$

## 11.9 Tetrad covariant derivative

The derivation of tetrad covariant derivatives  $D_m$  follows precisely the analogous derivation of coordinate covariant derivatives  $D_\mu$ . 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  $\partial_m$  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  $\Gamma_{lmn}$  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  $\Phi$  is a scalar, then  $\partial_m \Phi$  is a tetrad 4-vector. The tetrad covariant derivative of a scalar is just the directed derivative

$$D_m \Phi = \partial_m \Phi \quad \text{a tetrad 4-vector .} \quad (11.34)$$

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  $\gamma_m$  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  $\gamma_m$  at two different points  $x$  and  $x+\delta x$ . 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  $\delta x$  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**,  $\Gamma_{mn}^k$ , also known as Ricci rotation coefficients (or, in the context of Newman-Penrose tetrads, spin coefficients),

$$\boxed{\partial_n \gamma_m \equiv \Gamma_{mn}^k \gamma_k} \quad \text{not a tetrad tensor.} \quad (11.37)$$

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  $\Gamma$  term for each contravariant index, and a negative  $\Gamma$  term for each covariant index.

## 11.10 Relation between tetrad and coordinate connections

The relation between the tetrad connections  $\Gamma_{mn}^k$  and their coordinate counterparts  $\Gamma_{\mu\nu}^\kappa$  follows from

$$\begin{aligned}\frac{\partial e_\mu}{\partial x^\nu} &= \Gamma_{\mu\nu}^\kappa e_\kappa = \frac{\partial e^m}{\partial x^\nu} \gamma_m \quad \text{not a tetrad tensor} \\ &= \frac{\partial e^m}{\partial x^\nu} \gamma_m + e^m \frac{\partial \gamma_m}{\partial x^\nu} \\ &= e^m \frac{\partial e^n}{\partial x^\nu} (d_{mn}^k + \Gamma_{mn}^k) \gamma_k .\end{aligned}\tag{11.43}$$

Thus the relation is

$$d_{lmn} + \Gamma_{lmn} = e_l^\lambda e_m^\mu e_n^\nu \Gamma_{\lambda\mu\nu} \quad \text{not a tetrad tensor}\tag{11.44}$$

where

$$\Gamma_{lmn} \equiv \gamma_{lk} \Gamma_{mn}^k .\tag{11.45}$$

## 11.11 Antisymmetry of the tetrad connections

The directed derivative of the tetrad metric is

$$\begin{aligned}\partial_n \gamma_{lm} &= \partial_n (\gamma_l \cdot \gamma_m) \\ &= \gamma_l \cdot \partial_n \gamma_m + \gamma_m \cdot \partial_n \gamma_l \\ &= \Gamma_{lmn} + \Gamma_{mln} .\end{aligned}\tag{11.46}$$

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

$$\Gamma_{lmn} = -\Gamma_{mln} .\tag{11.47}$$

This antisymmetry reflects the fact that  $\Gamma_{lmn}$  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 (2.56), by the commutator of the covariant derivative acting on a scalar  $\Phi$

$$[D_k, D_l] \Phi = S_{kl}^m \partial_m \Phi \quad \text{a tetrad tensor} .\tag{11.48}$$

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 (2.57) 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

$$\begin{aligned} \partial_n \gamma_{lm} + \partial_m \gamma_{ln} - \partial_l \gamma_{mn} &= \Gamma_{lmn} + \Gamma_{mln} + \Gamma_{lnm} + \Gamma_{nlm} - \Gamma_{mnl} - \Gamma_{nml} \\ &= 2\Gamma_{lmn} + S_{lnm} + S_{mln} + S_{nlm} - d_{lnm} + d_{lmn} - d_{mln} + d_{mnl} - d_{nlm} + d_{nml} \end{aligned} \quad (11.52)$$

implies that the tetrad connections  $\Gamma_{lmn}$  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 (11.32)

$$\boxed{\Gamma_{lmn} = \frac{1}{2} (d_{lnm} - d_{lmn} + d_{mln} - d_{mnl} + d_{nlm} - d_{nml})} \quad \text{not a tetrad tensor .} \quad (11.54)$$

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  $\circ$  overscript. The torsion-full tetrad connection  $\Gamma_{lmn}$  is a sum of the torsion-free (Levi-Civita) connection  $\overset{\circ}{\Gamma}_{lmn}$  and the contortion tensor  $K_{lmn}$ ,

$$\Gamma_{lmn} = \overset{\circ}{\Gamma}_{lmn} + K_{lmn}, \quad (11.55)$$

where from equation (11.53) the contortion tensor  $K_{lmn}$  and the torsion tensor  $S_{lmn}$  are related by

$$K_{lmn} = \frac{1}{2}(S_{lmn} - S_{mln} + S_{nml}) = -S_{nlm} + \frac{3}{2}S_{[lmn]} \quad \text{a tetrad tensor ,} \quad (11.56a)$$

$$S_{lmn} = K_{lmn} - K_{lnm} = -K_{mnl} + 3K_{[lmn]} \quad \text{a tetrad tensor .} \quad (11.56b)$$

Like the tetrad connection  $\Gamma_{lmn}$ , 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  $\overset{\circ}{D}_n$  by the contortion,

$$D_n A^k \equiv \overset{\circ}{D}_n A^k + K_{mn}^k A^m \quad \text{a tetrad tensor .} \quad (11.57)$$

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 (2.107), 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

$$[D_k, D_l] A_m \equiv R_{klmn} A^n \quad \text{a tetrad tensor .} \quad (11.59)$$

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:

$$R_{klmn} = \partial_k \Gamma_{mln} - \partial_l \Gamma_{mln} + \Gamma_{ml}^p \Gamma_{pnk} - \Gamma_{mk}^p \Gamma_{pln} + (\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p) \Gamma_{mnp} \quad \text{a tetrad tensor .} \quad (11.60)$$

The formula has extra terms  $(\Gamma_{kl}^p - \Gamma_{lk}^p - S_{kl}^p) \Gamma_{mnp}$  compared to the formula (2.109) for the coordinate-frame Riemann tensor  $R_{\kappa\lambda\mu\nu}$ . If torsion vanishes, as general relativity assumes, then

$$R_{klmn} = \partial_k \Gamma_{mln} - \partial_l \Gamma_{mln} + \Gamma_{ml}^p \Gamma_{pnk} - \Gamma_{mk}^p \Gamma_{pln} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \Gamma_{mnp} \quad \text{a tetrad tensor .} \quad (11.61)$$

The symmetries of the tetrad-frame Riemann tensor are the same as those of the coordinate-frame Riemann tensor. For vanishing torsion, these are

$$R_{klmn} = R_{([kl][mn])}, \quad (11.62)$$

$$R_{k[lmn]} = 0. \quad (11.63)$$

**Exercise 11.3 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  $D_k \gamma^{np} = 0$ . To complete the derivation, show that

$$\partial_k (\Gamma_{mnl} + \Gamma_{nml}) - \partial_l (\Gamma_{mnk} + \Gamma_{nmk}) = [\partial_k, \partial_l] \gamma_{mn} = (\Gamma_{lk}^p - \Gamma_{kl}^p + S_{kl}^p)(\Gamma_{mnp} + \Gamma_{nmp}). \quad (11.66)$$

Equation (11.66) implies the antisymmetry of  $R_{klmn}$  in  $mn$ .]

**Exercise 11.4 Antisymmetry of the Riemann tensor.** Argue that the antisymmetry of  $R_{klmn}$  in  $mn$ , with or without torsion, can be deduced from

$$0 = [D_k, D_l] \gamma_{mn} = S_{kl}^p D_p \gamma_{mn} + R_{klmp} \delta_n^p + R_{klnp} \delta_m^p = R_{klmn} + R_{klnm}. \quad (11.67)$$

**Exercise 11.5 Cyclic symmetry of the Riemann tensor.** Show that the cyclic symmetry (11.63) 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  $\Phi$  is a scalar, then

$$\begin{aligned} 2D_{[k} D_{l]} D_{m]} \Phi &= [D_{[k}, D_{l}]] D_{m]} \Phi = (R_{[klm]}{}^n - S_{[kl}^p S_{m]p}^n) D_n \Phi + S_{[kl}^n D_{m]} D_n \Phi \\ &= D_{[k} [D_{l]}, D_{m}]] \Phi = (D_{[k} S_{lm]}^n) D_n \Phi + S_{[kl}^n D_{m]} D_n \Phi. \end{aligned} \quad (11.68)$$

Consequently

$$R_{[klm]}{}^n = D_{[k} S_{lm]}^n + S_{[kl}^p S_{m]p}^n. \quad (11.69)$$

An equivalent expression in terms of the torsion-free covariant derivative  $\dot{D}_k$  and the contortion  $K_{mnl}$  is

$$R_{[klm]}{}^n = \dot{D}_{[k} S_{lm]}^n + K_{p[k}^n S_{lm]}^p. \quad (11.70)$$

**Exercise 11.6 Symmetry of the Riemann tensor.** Show that the cyclic symmetry (11.63) implies the symmetry  $kl \leftrightarrow mn$ , given the antisymmetries  $k \leftrightarrow l$  and  $m \leftrightarrow n$ . Given Exercise 11.5, this shows that the symmetry  $kl \leftrightarrow mn$  is, like the cyclic symmetry, a consequence of vanishing torsion.

**Solution.** Show that

$$2(R_{klmn} - R_{mnlk}) = 3(R_{k[lmn]} - R_{l[kmn]} - R_{m[nkl]} + R_{n[mkl]}) , \quad (11.71)$$

or alternatively,

$$2(R_{klmn} - R_{mnlk}) = 3(R_{[klm]n} - R_{[kln]m} - R_{[mnk]l} + R_{[mnl]k}) . \quad (11.72)$$

**Exercise 11.7 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 11.8 Must connections vanish if Riemann vanishes?** Must the tetrad connections  $\Gamma_{lmn}$  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 (20.21)).

---

### 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  $D_\kappa$  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  $\Gamma_{\mu\nu\kappa}$ . 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 (2.109) 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

$$\boxed{S_{\kappa\lambda}^m = -\frac{\partial e^m_\lambda}{\partial x^\kappa} + \frac{\partial e^m_\kappa}{\partial x^\lambda} - \Gamma_{l\kappa}^m e^l_\lambda + \Gamma_{k\lambda}^m e^k_\kappa} \quad \text{a coordinate-tetrad tensor}. \quad (11.78)$$

Equations (11.77) and (11.78) constitute Cartan's equations of structure (Cartan, 1904) (see §16.14.1).

## 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:

$$\boxed{R_{km} \equiv \gamma^{ln} R_{klmn}}. \quad (11.79)$$

Ricci scalar:

$$\boxed{R \equiv \gamma^{km} R_{km}}. \quad (11.80)$$

Einstein tensor:

$$\boxed{G_{km} \equiv R_{km} - \frac{1}{2} \gamma_{km} R}. \quad (11.81)$$

Einstein's equations:

$$\boxed{G_{km} = 8\pi G T_{km}}. \quad (11.82)$$

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,

$$R_{km} = 8\pi G (T_{km} - \frac{1}{2} \gamma_{km} T). \quad (11.83)$$

Bianchi identities in the absence of torsion:

$$\boxed{D_k R_{lmnp} + D_l R_{mknp} + D_m R_{klnp} = 0}, \quad (11.84)$$

which most importantly imply covariant conservation of the Einstein tensor, hence conservation of energy-momentum

$$\boxed{D^k T_{km} = 0} . \quad (11.85)$$

## 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  $\gamma_0$  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  $\Gamma_{lmn}$ , 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

$$R_{km} = \mathring{R}_{km} - \mathring{D}_k K_m - \mathring{D}^n K_{mnk} + K_{mpn} K^{np}{}_k - K_{mpk} K^p , \quad (11.88)$$

and the Ricci scalar is

$$R = \mathring{R} - 2\mathring{D}_n K^n + K_{mpn} K^{np}{}^m - K_n K^n . \quad (11.89)$$

The antisymmetric part of the Einstein tensor is, from contracting equation (11.69),

$$G_{[km]} = \frac{3}{2} R_{[klm]}{}^l = \frac{3}{2} \left( D_{[k} S_{lm]}^l + S_{[kl}^p S_{m]p}^l \right) , \quad (11.90)$$

which vanishes for vanishing torsion.

The Jacobi identity (2.123) 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

$$-\frac{3}{2} \gamma^{kn} \gamma^{lp} \left( D_{[k} R_{lm]np} + S_{[kl}^q R_{m]qnp} \right) = D^k G_{mk} - \frac{1}{2} S_{kl}^q R_{mq}{}^{kl} - S_{km}^q R_q{}^k = 0 . \quad (11.92)$$

**Exercise 11.9 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  $\theta$  and  $\phi$  represent the transverse (angular) directions, the non-zero components of the tetrad-frame Riemann tensor are

$$\frac{1}{2}R_{trtr} = -R_{tot\theta} = -R_{t\phi t\phi} = R_{r\theta r\theta} = R_{r\phi r\phi} = -\frac{1}{2}R_{\theta\phi\theta\phi} = C \quad (11.93)$$

where

$$C = -M/r^3 \quad (11.94)$$

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 (11.95),

$$\frac{D^2\delta\xi_m}{D\tau^2} + R_{klmn}\delta\xi^k u^l u^n = 0 , \quad (11.95)$$

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  $\xi^m$  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 (11.93) 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 meter 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

$$\tau = \frac{2}{3}\sqrt{\frac{r^3}{2M}} . \quad (11.96)$$

How long, in seconds, does it take to fall to the singularity from the place where the tidal acceleration is 1 gee per meter? 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 meter? 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

$$\frac{dr}{d\tau} = -\sqrt{\frac{2M}{r}} . \quad (11.97)$$

Show that radial component ( $\delta\xi^r$ ) of the equation of geodesic deviation (11.95) for such a person solves to

$$\delta\xi^r = \frac{A}{\sqrt{r}} + Br^2, \quad (11.98)$$

where  $A$  and  $B$  are constants. If a person tears apart when the tidal acceleration is 1 gee per meter, 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 (11.98), can you interpret what it means?]

### Exercise 11.10 Totally antisymmetric tensor.

1. In an orthonormal tetrad  $\gamma_m$  where  $\gamma_0$  points to the future and  $\gamma_1, \gamma_2, \gamma_3$  are right-handed, the contravariant totally antisymmetric tensor  $\varepsilon^{klmn}$  is defined by (this is the opposite sign from the Misner, Thorne, and Wheeler (1973) notation)

$$\varepsilon^{klmn} \equiv [klmn], \quad (11.99)$$

and hence

$$\varepsilon_{klmn} = -[klmn], \quad (11.100)$$

where  $[klmn]$  is the totally antisymmetric symbol

$$[klmn] \equiv \begin{cases} +1 & \text{if } klmn \text{ is an even permutation of 0123 ,} \\ -1 & \text{if } klmn \text{ is an odd permutation of 0123 ,} \\ 0 & \text{if } klmn \text{ are not all different .} \end{cases} \quad (11.101)$$

Argue that in a general basis  $e_\mu$  the contravariant totally antisymmetric tensor  $\varepsilon^{\kappa\lambda\mu\nu}$  is

$$\varepsilon^{\kappa\lambda\mu\nu} = e_k{}^\kappa e_l{}^\lambda e_m{}^\mu e_n{}^\nu \varepsilon^{klmn} = e[\kappa\lambda\mu\nu], \quad (11.102)$$

while its covariant counterpart is

$$\varepsilon_{\kappa\lambda\mu\nu} = -(1/e)[\kappa\lambda\mu\nu], \quad (11.103)$$

where  $e \equiv |e_m{}^\mu|$  is the determinant of the vierbein.

2. Show that in 4 dimensions

$$\varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -4! e^{[k}{}_\kappa e^{l}{}_\lambda e^{m}{}_\mu e^{n]}{}_\nu. \quad (11.104)$$

Conclude that

$$e_k{}^\kappa \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -6 e^{[l}{}_\lambda e^{m}{}_\mu e^{n]}{}_\nu, \quad (11.105a)$$

$$e_k{}^\kappa e_l{}^\lambda \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -4 e^{[m}{}_\mu e^{n]}{}_\nu, \quad (11.105b)$$

$$e_k{}^\kappa e_l{}^\lambda e_m{}^\mu \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -6 e^n{}_\nu, \quad (11.105c)$$

$$e_k{}^\kappa e_l{}^\lambda e_m{}^\mu e_n{}^\nu \varepsilon^{klmn} \varepsilon_{\kappa\lambda\mu\nu} = -24. \quad (11.105d)$$

The coefficient of the  $p$ 'th contraction is  $-p!(4-p)!$ .



# 12

---

## Spin and Newman-Penrose tetrads

THIS CHAPTER NEEDS REWRITING.

This chapter discusses spin tetrads (§???) and Newman-Penrose tetrads (§36.1). 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, tauons, 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 (35.3), 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,  $\pm 1$ , while gravitational waves have two possible spins, or helicities,  $\pm 2$ . 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  $\gamma_z$ , but there is no loss of generality in making this choice.

Start with an orthonormal tetrad  $\{\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y, \boldsymbol{\gamma}_z\}$ . If the preferred tetrad axis is the  $z$ -axis  $\boldsymbol{\gamma}_z$ , then the spin tetrad axes  $\{\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  are defined to be complex combinations of the transverse axes  $\{\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y\}$ ,

$$\boxed{\boldsymbol{\gamma}_+ \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_x + i\boldsymbol{\gamma}_y)} , \quad (12.1a)$$

$$\boxed{\boldsymbol{\gamma}_- \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_x - i\boldsymbol{\gamma}_y)} . \quad (12.1b)$$

The tetrad metric of the spin tetrad  $\{\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_z, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  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  $\{\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  are themselves null,  $\boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_+ = \boldsymbol{\gamma}_- \cdot \boldsymbol{\gamma}_- = 0$ , whereas their scalar product with each other is non-zero  $\boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_- = 1$ . 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, §36.1, 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  $\chi$  about the preferred axis  $\boldsymbol{\gamma}_z$ , the transverse axes  $\boldsymbol{\gamma}_x$  and  $\boldsymbol{\gamma}_y$  transform as

$$\begin{aligned} \boldsymbol{\gamma}_x &\rightarrow \cos \chi \boldsymbol{\gamma}_x + \sin \chi \boldsymbol{\gamma}_y , \\ \boldsymbol{\gamma}_y &\rightarrow \sin \chi \boldsymbol{\gamma}_x - \cos \chi \boldsymbol{\gamma}_y . \end{aligned} \quad (12.3)$$

It follows that the spin axes  $\boldsymbol{\gamma}_+$  and  $\boldsymbol{\gamma}_-$  transform under a right-handed rotation by angle  $\chi$  about  $\boldsymbol{\gamma}_z$  as

$$\boldsymbol{\gamma}_\pm \rightarrow e^{\mp i\chi} \boldsymbol{\gamma}_\pm . \quad (12.4)$$

The transformation (35.5) identifies the spin axes  $\boldsymbol{\gamma}_+$  and  $\boldsymbol{\gamma}_-$  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  $\chi$  about the preferred axis  $\boldsymbol{\gamma}_z$ . 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  $\boldsymbol{\gamma}_z$  axis, unchanged by a rotation of any angle about the axis. The  $\boldsymbol{\gamma}_z$  axis itself is spin-0, as is the time axis  $\boldsymbol{\gamma}_t$ .

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:

$$\boldsymbol{\gamma}_+ \leftrightarrow \boldsymbol{\gamma}_- , \quad (12.7)$$

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  $\chi$ . 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- $\pm 1$  components, and 2 spin- $\pm 2$  components:

$$\begin{aligned} -2: & G_{--}, \\ -1: & G_{t-}, G_{z-}, \\ 0: & G_{tt}, G_{tz}, G_{zz}, G_{+-}, \\ +1: & G_{t+}, G_{z+}, \\ +2: & G_{++}. \end{aligned} \tag{12.11}$$

The 4 spin-0 components are all real; in particular  $G_{+-}$  is real since  $G_{+-}^* = G_{-+} = G_{+-}$ . The 4 spin- $\pm 1$  and 2 spin- $\pm 2$  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- $\pm 1$ ), and tensor (spin- $\pm 2$ ).

## 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  $\gamma_v$  (outgoing) and  $\gamma_u$  (ingoing) along the direction of propagation are chosen to be lightlike, while the two axes  $\gamma_+$  and  $\gamma_-$  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  $\Gamma_{kmn}$  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  $\psi_{2-s}$  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  $\{\boldsymbol{\gamma}_v, \boldsymbol{\gamma}_u, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  is defined in terms of an orthonormal tetrad  $\{\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y, \boldsymbol{\gamma}_z\}$  by

$$\boxed{\boldsymbol{\gamma}_v \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_t + \boldsymbol{\gamma}_z)}, \quad (12.13a)$$

$$\boxed{\boldsymbol{\gamma}_u \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_z)}, \quad (12.13b)$$

$$\boxed{\boldsymbol{\gamma}_+ \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_x + i\boldsymbol{\gamma}_y)}, \quad (12.13c)$$

$$\boxed{\boldsymbol{\gamma}_- \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_x - i\boldsymbol{\gamma}_y)}, \quad (12.13d)$$

or in matrix form

$$\begin{pmatrix} \boldsymbol{\gamma}_v \\ \boldsymbol{\gamma}_u \\ \boldsymbol{\gamma}_+ \\ \boldsymbol{\gamma}_- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & i & 0 \\ 0 & 1 & -i & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_t \\ \boldsymbol{\gamma}_x \\ \boldsymbol{\gamma}_y \\ \boldsymbol{\gamma}_z \end{pmatrix}. \quad (12.14)$$

All four tetrad axes are null

$$\boldsymbol{\gamma}_v \cdot \boldsymbol{\gamma}_v = \boldsymbol{\gamma}_u \cdot \boldsymbol{\gamma}_u = \boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_+ = \boldsymbol{\gamma}_- \cdot \boldsymbol{\gamma}_- = 0. \quad (12.15)$$

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  $\{\boldsymbol{\gamma}_v, \boldsymbol{\gamma}_u, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  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  $\theta$  along the  $\boldsymbol{\gamma}_z$  axis multiplies the outgoing and ingoing axes  $\boldsymbol{\gamma}_v$  and  $\boldsymbol{\gamma}_u$  by a blueshift factor  $e^\theta$  and its reciprocal,

$$\begin{aligned} \boldsymbol{\gamma}_v &\rightarrow e^\theta \boldsymbol{\gamma}_v, \\ \boldsymbol{\gamma}_u &\rightarrow e^{-\theta} \boldsymbol{\gamma}_u. \end{aligned} \quad (12.17)$$

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  $\theta$  along the preferred direction  $\gamma_z$ . Thus  $\gamma_v$  has boost weight +1, and  $\gamma_u$  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:

$$\boxed{\text{boost weight } n = \text{number of } v \text{ minus } u \text{ covariant indices}} \quad . \quad (12.20)$$

### 12.2.3 Lorentz transformations

Under a Lorentz transformation consisting of a combination of a Lorentz boost by rapidity  $\xi$  about  $t-x$  and a rotation by angle  $\zeta$  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  $6 \times 6$  symmetric bivector matrix, organized as a  $2 \times 2$  matrix of  $3 \times 3$  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  $3 \times 3$  blocks  $C_{EE}$  and  $C_{BB}$  are symmetric, while  $C_{BE} = C_{EB}^\top$ . The cyclic symmetry (11.63) of the Riemann, hence Weyl, tensor implies that the off-diagonal  $3 \times 3$  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  $3 \times 3$  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  $\{\boldsymbol{\gamma}_v, \boldsymbol{\gamma}_u, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$ , equation (36.1), the Weyl tensor  $C_{klmn}$  has 5 distinct complex components, here denoted  $\psi_s$ , of spins respectively  $s = -2, -1, 0, +1$ , and  $+2$ :

$$\begin{aligned} -2 : \quad & \psi_{-2} && \equiv C_{u-u-}, \\ -1 : \quad & \psi_{-1} && \equiv C_{uvu-} = C_{+-u-}, \\ 0 : \quad & \psi_0 \equiv \frac{1}{2}(C_{uvuv} + C_{uv+-}) = \frac{1}{2}(C_{-+--} + C_{uv+-}) = C_{v+-u}, \\ +1 : \quad & \psi_1 && \equiv C_{vuv+} = C_{-+v+}, \\ +2 : \quad & \psi_2 && \equiv C_{v+v+}. \end{aligned} \quad (12.27)$$

The complex conjugates  $\psi_s^*$  of the 5 NP components of the Weyl tensor are:

$$\begin{aligned} \psi_{-2}^* &= C_{u+u+}, \\ \psi_{-1}^* &= C_{uvu+} = C_{-+u+}, \\ \psi_0^* = \frac{1}{2}(C_{uvuv} + C_{uv+-}) &= \frac{1}{2}(C_{-+--} + C_{uv+-}) = C_{v+-u}, \\ \psi_1^* &= C_{vuv-} = C_{-+v-}, \\ \psi_2^* &= C_{v-v-}. \end{aligned} \quad (12.28)$$

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:

$$C_{v+v-} = C_{u+u-} = C_{v+u+} = C_{v-u-} = 0. \quad (12.29)$$

The above convention that the index  $s$  on the NP component  $\psi_s$  labels its spin differs from the standard convention, where the spin  $s$  component of the Weyl tensor is impenetrably denoted  $\psi_{2-s}$  (e.g. Chandrasekhar (1983)):

$$\begin{aligned} -2 : \quad & \psi_4, \\ -1 : \quad & \psi_3, \\ 0 : \quad & \psi_2, \quad (\text{standard convention, not followed here}) \\ +1 : \quad & \psi_1, \\ +2 : \quad & \psi_0. \end{aligned} \quad (12.30)$$

### 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

$$\begin{aligned} \tilde{C}_{u-u-} &= \psi_{-2}, \\ \tilde{C}_{uvu-} &= \tilde{C}_{+-u-} = \psi_{-1}, \\ \tilde{C}_{uvuv} = \tilde{C}_{-+--} &= \tilde{C}_{uv+-} = \tilde{C}_{v+-u} = \psi_0, \\ \tilde{C}_{vuv+} &= \tilde{C}_{-+v+} = \psi_1, \\ \tilde{C}_{v+v+} &= \psi_2. \end{aligned} \quad (12.31)$$

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  $\psi_s$ .

With respect to a triple of bivector indices ordered as  $\{u-, uv, +v\}$ , the NP components of the complexified Weyl tensor constitute the  $3 \times 3$  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  $3 \times 3$  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  $\psi_s$  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  $\psi_{-2}$  (the one conventionally called  $\psi_4$ ) 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  $\psi_2$  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  $3 \times 3$  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  $\boldsymbol{\gamma}_{tz}$  unchanged, which is to say, transformations generated by the Lorentz rotor  $\exp(\zeta\boldsymbol{\gamma}_{tz})$  where  $\zeta$  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.

# 13

---

## 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 super geometric algebra and the super spacetime algebra, covered in Chapters 35 and 36. 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) and Grassmann (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  $\gamma$ -matrices, which explains the adoption of the symbol  $\gamma_m$  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  $\gamma_m$  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  $a$  as  $a \rightarrow Ra\bar{R}$  and a spinor  $\varphi$  as  $\varphi \rightarrow R\varphi$ . This, along with the standard definition (13.18) for the pseudoscalar, has the consequence that a right-handed rotation corresponds to  $R = e^{-i\theta/2}$  with  $\theta$  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- $\frac{1}{2}$

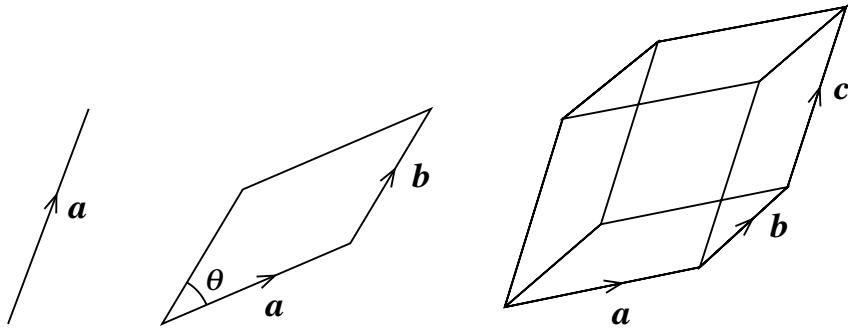


Figure 13.1 Multivectors of grade 1, 2, and 3: a vector  $\mathbf{a}$  (left), a bivector  $\mathbf{a} \wedge \mathbf{b}$  (middle), and a trivector  $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$  (right).

object with a transformation law (13.58) different from that (13.54) of multivectors. Earlier latin indices  $a, b, \dots$  run over spatial indices 1, 2, ... only, while mid latin indices  $m, n, \dots$  run over both time and space indices 0, 1, 2, ....

### 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  $\theta$  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

$$\mathbf{a} = \sum_{\text{distinct } \{a,b,\dots,d\} \subseteq \{1,2,\dots,N\}} a^{ab\dots d} \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b \wedge \dots \wedge \boldsymbol{\gamma}_d \quad (13.4)$$

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  $\boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b \wedge \dots \wedge \boldsymbol{\gamma}_d$ .

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  $\boldsymbol{\gamma}_a$  of the orthonormal basis with itself is a scalar, and with any other element of the basis is a bivector:

$$\boldsymbol{\gamma}_a \boldsymbol{\gamma}_b = \begin{cases} 1 & (a = b) \\ \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b & (a \neq b) \end{cases} . \quad (13.6)$$

Conversely, the rules (13.6), plus distributivity, imply the multiplication rule (13.5). A generalization of the rule (13.6) completes the definition of the geometric product:

$$\boldsymbol{\gamma}_a \boldsymbol{\gamma}_b \dots \boldsymbol{\gamma}_d = \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b \wedge \dots \wedge \boldsymbol{\gamma}_d \quad (a, b, \dots, d \text{ all distinct}) . \quad (13.7)$$

The rules (13.6) and (13.7), along with the usual requirements of associativity and distributivity, combined with commutativity of scalars and anticommutativity of pairs of  $\boldsymbol{\gamma}_a$ , uniquely define multiplication over the space of multivectors. For example, the product of the bivector  $\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$  with the vector  $\boldsymbol{\gamma}_1$  is

$$(\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2) \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_1 \boldsymbol{\gamma}_2 \boldsymbol{\gamma}_1 = -\boldsymbol{\gamma}_2 \boldsymbol{\gamma}_1 \boldsymbol{\gamma}_1 = -\boldsymbol{\gamma}_2 . \quad (13.8)$$

Sometimes it is convenient to denote the outer product (13.7) of distinct basis elements by the abbreviated symbol  $\boldsymbol{\gamma}_{ab\dots d}$  (with  $\boldsymbol{\gamma}$  in boldface)

$$\boldsymbol{\gamma}_{ab\dots d} \equiv \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b \wedge \dots \wedge \boldsymbol{\gamma}_d \quad (a, b, \dots, d \text{ all distinct}) . \quad (13.9)$$

By construction,  $\boldsymbol{\gamma}_{ab\dots d}$  is antisymmetric in its indices. The product of two general multivectors  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  and  $\mathbf{b} = b^B \boldsymbol{\gamma}_B$  is

$$\mathbf{ab} = a^A b^B \boldsymbol{\gamma}_A \boldsymbol{\gamma}_B , \quad (13.10)$$

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,

$$(1 + \boldsymbol{\gamma}_1)(1 - \boldsymbol{\gamma}_1) = 0 \quad (13.12)$$

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

$$\overline{\gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_m} \equiv \gamma_m \wedge \dots \wedge \gamma_b \wedge \gamma_a . \quad (13.13)$$

The reverse  $\bar{a}$  of any multivector  $a$  is the multivector obtained by reversing each of its components. Reversion leaves unchanged all multivectors whose grade is 0 or 1, modulo 4, and changes the sign of all multivectors whose grade is 2 or 3, modulo 4. Thus the reverse of a multivector  $a$  of pure grade  $p$  is

$$\bar{a} = (-)^{[p/2]} a , \quad (13.14)$$

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

$$\overline{a + b} = \bar{a} + \bar{b} , \quad (13.15)$$

$$\overline{ab} = \bar{b}\bar{a} . \quad (13.16)$$

Among other things, it follows that the reverse of any product of multivectors is the reversed product, as you would hope:

$$\overline{ab \dots c} = \bar{c} \dots \bar{b}\bar{a} . \quad (13.17)$$

### 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.

Define the **pseudoscalar**  $I_N$  in  $N$  dimensions to be

$$I_N \equiv \gamma_1 \wedge \gamma_2 \wedge \dots \wedge \gamma_N \quad (13.18)$$

with reverse

$$\bar{I}_N = (-)^{[N/2]} I_N , \quad (13.19)$$

where  $[N/2]$  signifies the largest integer less than or equal to  $N/2$ . The square of the pseudoscalar is

$$I_N^2 = (-)^{[N/2]} = \begin{cases} 1 & \text{if } N = (0 \text{ or } 1) \text{ modulo 4} \\ -1 & \text{if } N = (2 \text{ or } 3) \text{ modulo 4} \end{cases} . \quad (13.20)$$

The pseudoscalar anticommutes (commutes) with vectors  $\mathbf{a}$ , that is, with multivectors of grade 1, if  $N$  is even (odd):

$$\begin{aligned} I_N \mathbf{a} &= -\mathbf{a} I_N && \text{if } N \text{ is even} \\ I_N \mathbf{a} &= \mathbf{a} I_N && \text{if } N \text{ is odd}. \end{aligned} \quad (13.21)$$

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).

**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  $\gamma_a \wedge \gamma_b \wedge \dots \wedge \gamma_m$ . Since multiplication by a basis element permutes the basis elements amongst each other (and multiplies each by  $\pm 1$ ), 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 (13.21), commute with all other basis elements. Thus  $\mathbf{a}$  must be some linear combination of 1 and, if  $N$  is odd, the pseudoscalar  $I_N$ .

The **Hodge dual**  ${}^* \mathbf{a}$  of a multivector  $\mathbf{a}$  in  $N$  dimensions is defined by pre-multiplication by the pseudoscalar  $I_N$ ,

$${}^* \mathbf{a} \equiv I_N \mathbf{a}. \quad (13.22)$$

In 3 dimensions, the Hodge duals of the basis vectors  $\gamma_a$  are the bivectors

$$I_3 \gamma_1 = \gamma_2 \wedge \gamma_3, \quad I_3 \gamma_2 = \gamma_3 \wedge \gamma_1, \quad I_3 \gamma_3 = \gamma_1 \wedge \gamma_2. \quad (13.23)$$

Thus in 3 dimensions the bivector  $\mathbf{a} \wedge \mathbf{b}$  is seen to be the pseudovector Hodge dual to the familiar vector product  $\mathbf{a} \times \mathbf{b}$ :

$$\mathbf{a} \wedge \mathbf{b} = I_3 \mathbf{a} \times \mathbf{b}. \quad (13.24)$$

### 13.5 General wedge and dot 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

$$\langle \mathbf{a} \rangle_p, \quad (13.25)$$

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 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

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{pq-n^2} \langle \mathbf{ba} \rangle_{p+q-2n} \quad (13.26)$$

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  $\mathbf{b} = b^A \boldsymbol{\gamma}_B$  is

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{[n/2]} a^{AC} b_C^B \boldsymbol{\gamma}_A \wedge \boldsymbol{\gamma}_B , \quad (13.27)$$

implicitly summed over distinct sequences  $A$ ,  $B$ , and  $C$  of respectively  $p-n$ ,  $q-n$ , and  $n$  indices. Only components with the  $p+q+n$  indices of  $ABC$  all distinct contribute. The sign comes from  $\boldsymbol{\gamma}_C \boldsymbol{\gamma}_C = (-)^{[n/2]}$  for each index sequence  $C$ . Equation (13.27) can also be written

$$\langle \mathbf{ab} \rangle_{p+q-2n} = (-)^{[n/2]} \frac{(p+q-2n)!}{(p-n)!(q-n)!} a^{[AC]} b_C^B \boldsymbol{\gamma}_{AB} , \quad (13.28)$$

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.

The wedge product of multivectors of arbitrary grade is defined, consistent with the convention of differential forms, §15.7, to be the highest possible grade component of the 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

$$\boxed{\mathbf{a} \wedge \mathbf{b} \equiv \langle \mathbf{ab} \rangle_{p+q}} . \quad (13.29)$$

The definition (13.29) 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,

$$\mathbf{a} \wedge \mathbf{b} = (-)^{pq} \mathbf{b} \wedge \mathbf{a} , \quad (13.30)$$

which is a special case of equation (13.26). The wedge product is associative,

$$(\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c} = \mathbf{a} \wedge (\mathbf{b} \wedge \mathbf{c}) . \quad (13.31)$$

In accordance with the definition (13.29), 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,

$$\mathbf{a} \wedge \mathbf{b} = \mathbf{ab} \quad \text{if } a \text{ is a scalar} , \quad (13.32)$$

again consistent with the convention of differential forms.

The dot product of multivectors of arbitrary grade is defined to be the lowest grade component of their geometric product,

$$\boxed{\mathbf{a} \cdot \mathbf{b} \equiv \langle \mathbf{ab} \rangle_{|p-q|}} , \quad (13.33)$$

except that the dot product of a scalar, a zero grade multivector, with any multivector is defined to be zero,

$$a \cdot \mathbf{b} = 0 \quad \text{if } a \text{ is a scalar} \quad (13.34)$$

(this convention is adopted to ensure that, if  $\mathbf{b}$  is a vector, then  $\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b}$  for any multivector  $\mathbf{a}$ , including the case where  $\mathbf{a}$  is a scalar). The dot product is symmetric or antisymmetric,

$$\mathbf{a} \cdot \mathbf{b} = (-)^{(p-q)q} \mathbf{b} \cdot \mathbf{a} \quad \text{for } p \geq q . \quad (13.35)$$

The dot product is *not* associative.

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

$$\mathbf{a} \cdot \mathbf{b} = a^A \gamma_A \cdot b^B \gamma_B = (-)^{[p/2]} a^A b_A , \quad (13.36)$$

implicitly summed over distinct sequences  $A$  of  $p$  indices. Equation (13.36) is a special case of equation (13.27).

### 13.5.1 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

$$\langle \mathbf{abc} \rangle_0 = \langle \langle \mathbf{ab} \rangle_r \mathbf{c} \rangle_0 = \langle \mathbf{a} \langle \mathbf{bc} \rangle_p \rangle_0 . \quad (13.37)$$

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

$$\langle \mathbf{abc} \rangle_s = \sum_{n=|r-s|}^{r+s} \langle \langle \mathbf{ab} \rangle_n \mathbf{c} \rangle_s = \sum_{n=|p-s|}^{p+s} \langle \mathbf{a} \langle \mathbf{bc} \rangle_n \rangle_s . \quad (13.38)$$

Often some terms vanish, simplifying the relation. As an example of the triple-product relation (13.38), 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

$$I_N(\mathbf{a} \wedge \mathbf{b}) = (I_N \mathbf{a}) \cdot \mathbf{b} , \quad (\mathbf{a} \wedge \mathbf{b}) I_N = \mathbf{a} \cdot (b I_N) , \quad (13.39)$$

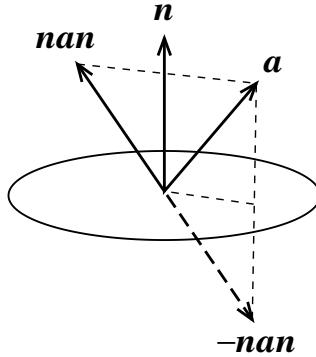
where  $I_N$  is the pseudoscalar (13.18).

## 13.6 Reflection

Multiplying a vector (a multivector of grade 1) by a vector shifts the grade (dimension) of the vector by  $\pm 1$ . 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

$$\mathbf{n} : \mathbf{a} \rightarrow \mathbf{n} \mathbf{a} \mathbf{n} , \quad (13.40)$$

Figure 13.2 Reflection of a vector  $\mathbf{a}$  through axis  $\mathbf{n}$ .

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 (13.40) is

$$\mathbf{n} : \mathbf{a}_{\parallel} + \mathbf{a}_{\perp} \rightarrow \mathbf{a}_{\parallel} - \mathbf{a}_{\perp}, \quad (13.41)$$

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:

$$\mathbf{n} : \mathbf{a} \rightarrow \mathbf{n}\mathbf{a}\mathbf{n}. \quad (13.42)$$

For example, the product  $\mathbf{ab}$  of two vectors transforms as

$$\mathbf{n} : \mathbf{ab} \rightarrow \mathbf{n}(\mathbf{ab})\mathbf{n} = (\mathbf{n}\mathbf{a}\mathbf{n})(\mathbf{n}\mathbf{b}\mathbf{n}) \quad (13.43)$$

which works because  $\mathbf{n}^2 = 1$ .

A reflection leaves any scalar  $\lambda$  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  $\mathbf{n}$ :

$$\mathbf{nm} : \mathbf{a} \rightarrow \mathbf{n}\mathbf{mamn}. \quad (13.44)$$

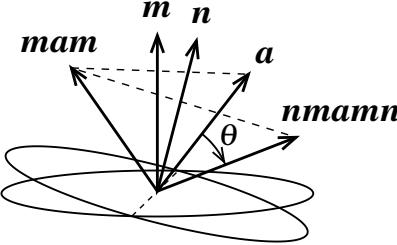


Figure 13.3 Two successive reflections of a vector  $\mathbf{a}$ , first through  $\mathbf{m}$ , then through  $\mathbf{n}$ , yield a rotation of a vector  $\mathbf{a}$  by the bivector  $\mathbf{mn}$ . Baffled? Hey, draw your own picture.

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 (13.44), since each reflection flips the sign of  $\mathbf{a}_\perp$ :

$$\mathbf{nm} : \mathbf{a}_\perp \rightarrow \mathbf{nma}_\perp \mathbf{mn} = -\mathbf{n}\mathbf{a}_\perp \mathbf{n} = \mathbf{a}_\perp . \quad (13.45)$$

Rotations inherit from reflections the property of preserving the lengths of, and angles between, all vectors. Thus the transformation (13.44) 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  $\theta$  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  $\pi$ , and the transformation (13.44) rotates the vector  $\mathbf{a}_\parallel$  by  $\theta = 0$  or  $2\pi$ , 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 (13.44) rotates  $\mathbf{a}_\parallel$  by  $\theta = \pm\pi$ , that is, it maps  $\mathbf{a}_\parallel$  to  $-\mathbf{a}_\parallel$ .

The rotation (13.44) can be abbreviated

$$R : \mathbf{a} \rightarrow R\mathbf{a}\bar{R} \quad (13.46)$$

where  $R = \mathbf{nm}$  is called a **rotor**, and  $\bar{R} = \mathbf{mn}$  is its reverse. Rotors are **unimodular**, satisfying  $R\bar{R} = \bar{R}R = 1$ . According to the discussion above, the transformation (13.46) corresponds to a rotation by angle  $\theta$  in the  $\mathbf{m}$ - $\mathbf{n}$  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  $\mathbf{m}$ - $\mathbf{n}$  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  $\theta$  in the  $\boldsymbol{\gamma}_1$ - $\boldsymbol{\gamma}_2$  plane is given by

$$R = \cos \frac{\theta}{2} - (\boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2) \sin \frac{\theta}{2} . \quad (13.47)$$

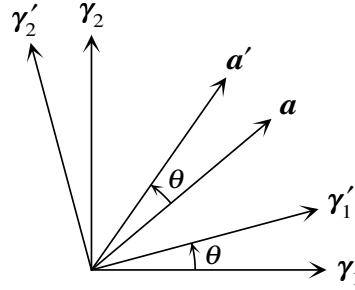


Figure 13.4 Right-handed rotation of a vector  $\mathbf{a}$  by angle  $\theta$  in the  $\gamma_1-\gamma_2$  plane. A rotation in the geometric algebra is an active rotation, which rotates the axes  $\gamma_a \rightarrow \gamma'_a$  while leaving the components  $a^a$  of a multivector unchanged, equation (13.52). 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.

The rotor (13.47) can also be written as an exponential of the bivector  $\theta = \theta \gamma_1 \wedge \gamma_2$ ,

$$R = e^{-\theta/2}. \quad (13.48)$$

It is straightforward to check that the action of the rotor (13.47) on the basis vectors  $\gamma_a$  is

$$R : \gamma_1 \rightarrow R\gamma_1\bar{R} = \gamma_1 \cos \theta + \gamma_2 \sin \theta, \quad (13.49a)$$

$$R : \gamma_2 \rightarrow R\gamma_2\bar{R} = \gamma_2 \cos \theta - \gamma_1 \sin \theta, \quad (13.49b)$$

$$R : \gamma_a \rightarrow R\gamma_a\bar{R} = \gamma_a \quad (a \neq 1, 2), \quad (13.49c)$$

which corresponds to a right-handed rotation of the basis vectors  $\gamma_a$  by angle  $\theta$  in the  $\gamma_1-\gamma_2$  plane. The inverse rotation is

$$\bar{R} : \mathbf{a} \rightarrow \bar{R}\mathbf{a}R \quad (13.50)$$

with

$$\bar{R} = \cos \frac{\theta}{2} + (\gamma_1 \wedge \gamma_2) \sin \frac{\theta}{2}. \quad (13.51)$$

A rotation of the form (13.47), a rotation in a single plane, is called a **simple rotation**.

In the geometric algebra, a rotation is considered to rotate the axes  $\gamma_a \rightarrow \gamma'_a$  while leaving the components  $a^a$  of a multivector unchanged. Thus a rotation transforms a vector  $\mathbf{a}$  as

$$R : \mathbf{a} = a^a \gamma_a \rightarrow \mathbf{a}' = a^a \gamma'_a. \quad (13.52)$$

Figure 13.4 illustrates a right-handed rotation by angle  $\theta$  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

$$SR : \mathbf{a} \rightarrow SR\mathbf{a}\bar{R}\bar{S} = SR\mathbf{a}\bar{S}\bar{R}. \quad (13.53)$$

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 (13.47), but with complex angle  $\theta$  and two orthogonal planes of rotation combined into a complex pair of planes.

Simple rotors are both even and unimodular, and composition preserves those properties. A rotor  $R$  is defined in general to be any even, unimodular ( $R\bar{R} = 1$ ) element of the geometric algebra. The set of rotors defines a group the **rotor group**, also referred to here as the **rotation group**. A rotor  $R$  rotates not only vectors, but multivectors  $a$  in general:

$$R : a \rightarrow Ra\bar{R}. \quad (13.54)$$

For example, the product  $ab$  of two vectors transforms as

$$R : ab \rightarrow R(ab)\bar{R} = (Ra\bar{R})(Rb\bar{R}) \quad (13.55)$$

which works because  $\bar{R}R = 1$ .

**Concept question 13.2 How fast do bivectors rotate?** If vectors rotate twice as fast as rotors, do bivectors rotate twice as fast as vectors? What happens to a bivector when you rotate it by  $\pi$  radians? Construct a mental picture of a rotating bivector.

To summarize, the characterization of rotations by rotors has considerable advantages. Firstly, the transformation (13.54) applies to multivectors  $a$  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  $\theta$  in equation (13.47).

### 13.8 A multivector rotation is an active rotation

So far in this book, indices have indicated how an object transforms, so that the notation

$$a^m \gamma_m \quad (13.56)$$

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  $\gamma_m$ . However, the transformation (13.54) of a multivector is to be understood as an **active** transformation that rotates the basis vectors  $\gamma_A$  while keeping the coefficients  $a^A$  fixed, as opposed to a **passive** transformation that rotates the tetrad while keeping the thing itself unchanged. Thus a multivector  $a \equiv a^A \gamma_A$  (implicit summation

over distinct antisymmetrized  $A \subseteq \{1, \dots, N\}$ ) is not a scalar under the transformation (13.54), but rather transforms to the multivector  $a' \equiv a^A \gamma'_A$  given by

$$R : a^A \gamma_A \rightarrow a^A \bar{R} \gamma_A R = a^A \gamma'_A . \quad (13.57)$$

Figure 13.4 illustrates the example of a right-handed rotation by angle  $\theta$  in the  $\gamma_1 - \gamma_2$  plane, equations (13.49).

### 13.9 A rotor is a spin- $\frac{1}{2}$ object

A rotor was defined in §13.7 as an even, unimodular element of the geometric algebra. As a multivector, a rotor  $R$  would transform under a rotation by the rotor  $S$  as  $R \rightarrow SRS$ . As a rotor, however, the rotor  $R$  transforms under a rotation by the rotor  $S$  as

$$S : R \rightarrow SR , \quad (13.58)$$

according to the transformation law (13.53). That is, composition in the rotor group is defined by the transformation (13.58):  $R$  rotated by  $S$  is  $SR$ .

The expression (13.47) for a simple rotation in the  $\gamma_1 - \gamma_2$  plane shows that the rotor corresponding to a rotation by  $2\pi$  is  $-1$ . Thus under a rotation (13.58) by  $2\pi$ , a rotor  $R$  changes sign:

$$2\pi : R \rightarrow -R . \quad (13.59)$$

A rotation by  $4\pi$  is necessary to bring the rotor  $R$  back to its original value:

$$4\pi : R \rightarrow R . \quad (13.60)$$

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  $a$  to  $Ra\bar{R}$ . In the primed frame rotated by rotor  $S$  from the unprimed frame,  $a' = Sa\bar{S}$ , the transformed rotor is  $SRS$ , since

$$a' = Sa\bar{S} \rightarrow SRa\bar{R}S = SRSa'SRS = SRSa'\bar{SRS} . \quad (13.61)$$

By contrast, the transformation (13.58) of a rotor as a rotor answers the question, what is the rotor corresponding to a rotation  $R$  followed by a rotation  $S$ ?

### 13.10 Generator of a rotation

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  $\theta$  in the  $\gamma_a - \gamma_b$  plane can be thought of as being built up from an

infinite number of infinitesimal rotations  $\exp(-\frac{1}{2}\delta\theta \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b)$  by angles  $\delta\theta$ . To linear order, an infinitesimal rotation by angle  $\delta\theta$  in the  $\boldsymbol{\gamma}_a$ - $\boldsymbol{\gamma}_b$  plane is

$$\exp(-\frac{1}{2}\delta\theta \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b) = 1 - \frac{1}{2}\delta\theta \boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b . \quad (13.62)$$

The bivector  $-\boldsymbol{\gamma}_a \wedge \boldsymbol{\gamma}_b$  is said to be the **generator** of a right-handed rotation in the  $\boldsymbol{\gamma}_a$ - $\boldsymbol{\gamma}_b$  plane.

The Baker-Campbell-Hausdorff formula states that the product of exponentials of not-necessarily-commuting elements  $\boldsymbol{\theta}$  and  $\boldsymbol{\phi}$  is

$$\exp(\boldsymbol{\theta}) \exp(\boldsymbol{\phi}) = \exp(\boldsymbol{\theta} + \boldsymbol{\phi} + \frac{1}{2}[\boldsymbol{\theta}, \boldsymbol{\phi}] + \frac{1}{12}[[\boldsymbol{\theta}, \boldsymbol{\phi}], \boldsymbol{\phi}] - \frac{1}{12}[[\boldsymbol{\theta}, \boldsymbol{\phi}], \boldsymbol{\theta}] + \dots) , \quad (13.63)$$

where  $[\boldsymbol{\theta}, \boldsymbol{\phi}] \equiv \boldsymbol{\theta}\boldsymbol{\phi} - \boldsymbol{\phi}\boldsymbol{\theta}$  is the commutator of  $\boldsymbol{\theta}$  and  $\boldsymbol{\phi}$ . 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 algebra of bivectors constitute the Lie algebra of the Lie group of rotations.

### 13.11 2D rotations and complex numbers

Section 13.7 identified the rotation group in  $N$  dimensions with the geometric subalgebra of even, unimodular multivectors. In two dimensions, the even grade multivectors are linear combinations of the basis set

$$\begin{array}{ll} 1, & I_2, \\ \text{1 scalar} & \text{1 bivector (pseudoscalar)} \end{array} \quad (13.64)$$

forming a linear space of dimension 2. The sole bivector is the pseudoscalar  $I_2 \equiv \boldsymbol{\gamma}_1 \wedge \boldsymbol{\gamma}_2$ , equation (13.18), the highest grade element in 2 dimensions. The rotor  $R$  that produces a right-handed rotation by angle  $\theta$  is, according to equation (13.47),

$$R = e^{-\boldsymbol{\theta}/2} = e^{-I_2 \theta/2} = \cos \frac{\theta}{2} - I_2 \sin \frac{\theta}{2} , \quad (13.65)$$

where  $\boldsymbol{\theta} = I_2 \theta$  is the bivector whose magnitude is  $(\bar{\boldsymbol{\theta}}\boldsymbol{\theta})^{1/2} = \theta$ .

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

$$I_2 \leftrightarrow i \quad (13.66)$$

defines an isomorphism between the algebra of even grade multivectors in 2 dimensions and the field of complex numbers

$$a + I_2 b \leftrightarrow a + i b . \quad (13.67)$$

With the isomorphism (13.67), the rotor  $R$  that produces a right-handed rotation by angle  $\theta$  is equivalent to the complex number

$$\boxed{R = e^{-i\theta/2}} . \quad (13.68)$$

The associated reverse rotor  $\bar{R}$  is

$$\bar{R} = e^{i\theta/2}, \quad (13.69)$$

the complex conjugate of  $R$ . The group of 2D rotors is isomorphic to the group of complex numbers of unit magnitude, the unitary group  $U(1)$ ,

$$2\text{D rotors} \leftrightarrow U(1). \quad (13.70)$$

Let  $\mathbf{z}$  denote an even multivector, equivalent to some complex number by the isomorphism (13.67). According to the transformation formula (13.54), under the rotation  $R = e^{-i\theta/2}$ , the even multivector, or complex number,  $\mathbf{z}$  transforms as

$$R : \mathbf{z} \rightarrow e^{-i\theta/2} \mathbf{z} e^{i\theta/2} = e^{-i\theta/2} e^{i\theta/2} \mathbf{z} = \mathbf{z}, \quad (13.71)$$

which is true because even multivectors in 2 dimensions commute, as complex numbers should. Equation (13.71) 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  $\theta$  in the Argand plane? The answer is that the rotation  $R : \mathbf{a} \rightarrow R\mathbf{a}\bar{R}$  rotates vectors  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_2$  (Exercise 13.3), as already seen in the transformation (13.49). 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 13.3 Rotation of a vector.** Confirm that a right-handed rotation by angle  $\theta$  rotates the axes  $\boldsymbol{\gamma}_a$  by

$$R : \boldsymbol{\gamma}_1 \rightarrow e^{-i\theta/2} \boldsymbol{\gamma}_1 e^{i\theta/2} = \boldsymbol{\gamma}_1 \cos \theta + \boldsymbol{\gamma}_2 \sin \theta, \quad (13.72a)$$

$$R : \boldsymbol{\gamma}_2 \rightarrow e^{-i\theta/2} \boldsymbol{\gamma}_2 e^{i\theta/2} = \boldsymbol{\gamma}_2 \cos \theta - \boldsymbol{\gamma}_1 \sin \theta, \quad (13.72b)$$

in agreement with (13.49). 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

$$q = a + ib_1 + jb_2 + kb_3, \quad (13.73)$$

where  $a$  and  $b_a$  ( $a = 1, 2, 3$ ) are real numbers, and the three imaginary numbers  $\iota, j, k$ , are defined to satisfy<sup>1</sup>

$$\iota^2 = j^2 = k^2 = -\iota j k = -1 . \quad (13.74)$$

Remark the dotless  $\iota$  (and  $j$ ), to distinguish these quaternionic imaginaries from other possible imaginaries. A consequence of equations (13.74) is that each pair of imaginary numbers anticommutes:

$$\iota j = -j\iota = -k , \quad jk = -kj = -\iota , \quad k\iota = -\iota k = -j . \quad (13.75)$$

It is convenient to abbreviate the three imaginaries by  $\iota_a$  with  $a = 1, 2, 3$ ,

$$\{\iota, j, k\} \equiv \{\iota_1, \iota_2, \iota_3\} . \quad (13.76)$$

The quaternion (13.73) can then be expressed compactly as a sum of its scalar  $a$  and vector (actually pseudovector, as will become apparent below from the isomorphism (13.90))  $\mathbf{b} = \iota_a b_a$  parts

$$q = a + \mathbf{b} = a + \iota_a b_a , \quad (13.77)$$

implicitly summed over  $a = 1, 2, 3$ . A fundamentally useful formula, which follows from the defining equations (13.74), is

$$\mathbf{a}\mathbf{b} = (\iota_a a_a)(\iota_b b_b) = -\mathbf{a} \cdot \mathbf{b} - \mathbf{a} \times \mathbf{b} = -a_a b_a - \iota_a \varepsilon_{abc} a_b b_c , \quad (13.78)$$

where  $\mathbf{a} \cdot \mathbf{b}$  and  $\mathbf{a} \times \mathbf{b}$  denote the usual 3D scalar and vector products, and  $\varepsilon_{abc}$  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

$$\begin{aligned} pq &= (a + \mathbf{b})(c + \mathbf{d}) = (a + \iota_a b_a)(c + \iota_b d_b) \\ &= ac - \mathbf{b} \cdot \mathbf{d} + ad + cb - \mathbf{b} \times \mathbf{d} = ac - b_a d_a + \iota_a(ad_a + cb_a - \varepsilon_{abc} b_b d_c) . \end{aligned} \quad (13.79)$$

The **quaternionic conjugate**  $\bar{q}$  of a quaternion  $q \equiv a + \mathbf{b}$  is (the overbar symbol  $\bar{\phantom{x}}$  for quaternionic conjugation distinguishes it from the asterisk symbol  $*$  for complex conjugation)

$$\bar{q} = a - \mathbf{b} = a - \iota_a b_a . \quad (13.80)$$

The quaternionic conjugate of a product is the reversed product of quaternionic conjugates

$$\bar{p}\bar{q} = \bar{q}\bar{p} \quad (13.81)$$

just like reversion in the geometric algebra, equation (13.16). 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

$$|q| = (\bar{q}q)^{1/2} = (q\bar{q})^{1/2} = (a^2 + \mathbf{b} \cdot \mathbf{b})^{1/2} = (a^2 + b_a b_a)^{1/2} . \quad (13.82)$$

<sup>1</sup> The choice  $\iota j k = 1$  in the definition (13.74) is the *opposite* of the conventional definition  $ijk = -1$  famously carved by W. R. 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  $\iota = -i$ ,  $j = -j$ ,  $k = -k$ , or alternatively  $\iota = i$ ,  $j = -j$ ,  $k = k$ , or  $\iota = k$ ,  $j = j$ ,  $k = i$ . The adopted choice  $\iota j k = 1$  has the merit that it avoids a treacherous minus sign in the isomorphism (13.90) between 3-dimensional pseudovectors and quaternions. The present choice also conforms to the convention used by OpenGL and other computer graphics programs.

The **inverse**  $q^{-1}$  of the quaternion, satisfying  $qq^{-1} = q^{-1}q = 1$ , is

$$q^{-1} = \bar{q}/(\bar{q}q) = (a - \mathbf{b})/(a^2 + \mathbf{b} \cdot \mathbf{b}) = (a - \iota_a b_a)/(a^2 + b_b b_b) . \quad (13.83)$$

### 13.13 3D rotations and quaternions

As before, the rotation group is the group of even, unimodular multivectors of the geometric algebra. In three dimensions, the even grade multivectors are linear combinations of the basis set

1 ,	$I_3\gamma_1$ ,	$I_3\gamma_2$ ,	$I_3\gamma_3$ ,	
1 scalar	3 bivectors (pseudovectors)			

(13.84)

forming a linear space of dimension 4. The three bivectors are pseudovectors, equation (13.23). The squares of the pseudovector basis elements are all minus one,

$$(I_3\gamma_1)^2 = (I_3\gamma_2)^2 = (I_3\gamma_3)^2 = -1 , \quad (13.85)$$

and they anticommute with each other,

$$\begin{aligned} (I_3\gamma_1)(I_3\gamma_2) &= -(I_3\gamma_2)(I_3\gamma_1) = -I_3\gamma_3 , \\ (I_3\gamma_2)(I_3\gamma_3) &= -(I_3\gamma_3)(I_3\gamma_2) = -I_3\gamma_1 , \\ (I_3\gamma_3)(I_3\gamma_1) &= -(I_3\gamma_1)(I_3\gamma_3) = -I_3\gamma_2 . \end{aligned} \quad (13.86)$$

The rotor  $R$  that produces a rotation by angle  $\theta$  right-handedly about unit direction  $n_a = \{n_1, n_2, n_3\}$ , satisfying  $n_a n_a = 1$ , is, according to equation (13.47),

$$R = e^{-\theta/2} = e^{-\mathbf{n}\theta/2} = \cos \frac{\theta}{2} - \mathbf{n} \sin \frac{\theta}{2} .$$

(13.87)

where  $\theta$  is the bivector

$$\theta \equiv \mathbf{n}\theta = I_3\gamma_a n_a \theta . \quad (13.88)$$

of magnitude  $(\bar{\theta}\theta)^{1//2} = \theta$  and unit direction  $\mathbf{n} \equiv I_3\gamma_a n_a$  (satisfying  $\bar{n}\mathbf{n} = 1$ ). The pseudovector  $I_3$  is a commuting imaginary, commuting with all members of the 3D geometric algebra, both odd and even, and satisfying

$$I_3^2 = -1 . \quad (13.89)$$

Comparison of equations (13.85) and (13.86) to equations (13.74) and (13.75), shows that the mapping

$$I_3\gamma_a \leftrightarrow \iota_a \quad (a = 1, 2, 3) \quad (13.90)$$

defines an isomorphism between the space of even multivectors in 3 dimensions and the non-commutative division algebra of quaternions

$$a + I_3\gamma_a b_a \leftrightarrow a + \iota_a b_a . \quad (13.91)$$

With the equivalence (13.91), the rotor  $R$  given by equation (13.87) can be interpreted as a quaternion, with  $\theta$  the quaternion

$$\theta \equiv \mathbf{n} \theta = \iota_a n_a \theta . \quad (13.92)$$

The associated reverse rotor  $\bar{R}$  is

$$\bar{R} = e^{\theta/2} = e^{\mathbf{n} \theta/2} = \cos \frac{\theta}{2} + \mathbf{n} \sin \frac{\theta}{2} , \quad (13.93)$$

the quaternionic conjugate of  $R$ .

The group of rotors is isomorphic to the group of unit quaternions, quaternions  $q = a + \iota_1 b_1 + \iota_2 b_2 + \iota_3 b_3$  satisfying  $q\bar{q} = a^2 + b_1^2 + b_2^2 + b_3^2 = 1$ . Unit 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 13.4 3D rotation matrices.** This exercise is a precursor to Exercise 14.8. The principle 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 (13.78) to show that the vector  $\mathbf{b}$  transforms under a right-handed rotation by angle  $\theta$  about unit direction  $\mathbf{n} = \boldsymbol{\gamma}_a n_a$  as

$$R : \mathbf{b} \rightarrow R \mathbf{b} \bar{R} = \mathbf{b} + 2 \sin \frac{\theta}{2} \mathbf{n} \times \left( \cos \frac{\theta}{2} \mathbf{b} + \sin \frac{\theta}{2} \mathbf{n} \times \mathbf{b} \right) . \quad (13.94)$$

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 (13.24). 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 (13.94) is (note that the  $3 \times 3$  rotation matrix is written to the left of the vector, in accordance with the physics convention that rotations accumulate to the left):

$$R : \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} \rightarrow \begin{pmatrix} w^2 + x^2 - y^2 - z^2 & 2(xy - wz) & 2(zx + wy) \\ 2(xy + wz) & w^2 - x^2 + y^2 - z^2 & 2(yz - wx) \\ 2(zx - wy) & 2(yz + wx) & w^2 - x^2 - y^2 + z^2 \end{pmatrix} \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} . \quad (13.95)$$

Confirm that the  $3 \times 3$  rotation matrix on the right hand side of the transformation (13.95) is an orthogonal matrix (its inverse is its transpose) provided that the rotor is unimodular,  $R\bar{R} = 1$ , so that  $w^2 + x^2 + y^2 + z^2 = 1$ . As a simple example, show that the transformation (13.95) in the case of a right-handed rotation by angle  $\theta$  about the 3-axis (the 1–2 plane), where  $w = \cos(\theta/2)$  and  $z = -\sin(\theta/2)$ , is

$$R : \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} \rightarrow \begin{pmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \boldsymbol{\gamma}_1 \\ b_2 \boldsymbol{\gamma}_2 \\ b_3 \boldsymbol{\gamma}_3 \end{pmatrix} . \quad (13.96)$$

### 13.14 Pauli matrices

The multiplication rules of the basis vectors  $\gamma_a$  of the 3D geometric algebra are identical to those of the Pauli matrices  $\sigma_a$  used in the theory of non-relativistic spin- $\frac{1}{2}$  particles.

The **Pauli matrices** form a vector of  $2 \times 2$  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 ( $\sigma_a^{-1} = \sigma^\dagger$ ):

$$\sigma_1 \equiv \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 \equiv \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (13.97)$$

The Pauli matrices anticommute with each other

$$\sigma_1\sigma_2 = -\sigma_2\sigma_1 = i\sigma_3, \quad \sigma_2\sigma_3 = -\sigma_3\sigma_2 = i\sigma_1, \quad \sigma_3\sigma_1 = -\sigma_1\sigma_3 = i\sigma_2. \quad (13.98)$$

The particular choice (13.97) of Pauli matrices is conventional but not unique: any three traceless, Hermitian, unitary, anticommuting  $2 \times 2$  complex matrices will do. The product of the 3 Pauli matrices is  $i$  times the unit matrix,

$$\sigma_1\sigma_2\sigma_3 = i \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (13.99)$$

If the scalar 1 in the geometric algebra is identified with the unit  $2 \times 2$  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

$$1 \leftrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \gamma_a \leftrightarrow \sigma_a, \quad I_3 \leftrightarrow i \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (13.100)$$

The 3D pseudoscalar  $I_3$  commutes with all elements of the 3D geometric algebra.

**Concept question 13.5 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?

The rotation group is the group of even, unimodular multivectors of the geometric algebra. The isomorphism (13.100) establishes that the rotation group is isomorphic to the group of complex  $2 \times 2$  matrices of the form

$$a + i\sigma_a b_a, \quad (13.101)$$

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 13.6) that the group of such matrices constitutes the group of unitary complex  $2 \times 2$  matrices of unit determinant, the special unitary group  $SU(2)$ . The isomorphisms

$$a + I_3 \gamma_a b_a \leftrightarrow a + \iota_a b_a \leftrightarrow a + i\sigma_a b_a \quad (13.102)$$

have thus established isomorphisms between the group of 3D rotors, the group of unit quaternions, and the special unitary group of complex  $2 \times 2$  matrices

$$\text{3D rotors} \leftrightarrow \text{unit quaternions} \leftrightarrow SU(2) . \quad (13.103)$$

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  $2 \times 2$  complex matrices may be termed the **Pauli representation**. The Pauli representation is the lowest dimensional representation of the 3D rotation group.

**Exercise 13.6 Translate a rotor into an element of  $SU(2)$ .** Show that the rotor  $R = e^{-i_a n_a \theta/2}$ , equation (13.87), corresponding to a right-handed rotation by angle  $\theta$  about unit axis  $n_a$  is equivalent to the special unitary  $2 \times 2$  matrix

$$R \leftrightarrow \begin{pmatrix} \cos \frac{\theta}{2} - i n_3 \sin \frac{\theta}{2} & -(n_2 + i n_1) \sin \frac{\theta}{2} \\ (n_2 - i n_1) \sin \frac{\theta}{2} & \cos \frac{\theta}{2} + i n_3 \sin \frac{\theta}{2} \end{pmatrix} . \quad (13.104)$$

Show that the reverse rotor  $\bar{R}$  is equivalent to the Hermitian conjugate  $R^\dagger$  of the corresponding  $2 \times 2$  matrix. Show that the determinant of the matrix equals  $\bar{R}R$ , which is 1.

### 13.15 Pauli spinors as quaternions, or scaled rotors

Any Pauli spinor  $\varphi$  can be expressed uniquely in the form of a  $2 \times 2$  matrix  $\mathbf{q}$ , the Pauli representation of a quaternion  $q$ , acting on the spin-up basis element  $\gamma_\uparrow$  (the precise translation between Pauli spinors and quaternions is left as Exercises 13.7 and 13.8):

$$\varphi = \mathbf{q} \gamma_\uparrow . \quad (13.105)$$

In this section (and in the Exercises) the  $2 \times 2$  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  $\lambda$  and a rotor, or unit quaternion,  $R$ . The real scalar  $\lambda$  can be taken without loss of generality to be positive, since any minus sign can be absorbed into a rotation by  $2\pi$  of the rotor  $R$ . Thus a Pauli spinor  $\varphi$  can also be expressed as a scaled rotor  $\lambda R$  acting on the spin-up basis element  $\gamma_\uparrow$ ,

$$\varphi = \lambda R \gamma_\uparrow . \quad (13.106)$$

One is used to thinking of a Pauli spinor as an intrinsically quantum-mechanical object. The mapping (13.105) or (13.106) between Pauli spinors and quaternions or scaled rotors shows that Pauli spinors also have a classical interpretation: they encode a real amplitude  $\lambda$ , 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.

**Exercise 13.7 Translate a Pauli spinor into a quaternion.** Given any Pauli spinor

$$\varphi \equiv \varphi^\uparrow \gamma_\uparrow + \varphi^\downarrow \gamma_\downarrow = \begin{pmatrix} \varphi^\uparrow \\ \varphi^\downarrow \end{pmatrix}, \quad (13.107)$$

show that the corresponding real quaternion  $q$ , and the equivalent  $2 \times 2$  complex matrix  $\mathbf{q}$  in the Pauli representation (13.97), such that  $\varphi = \mathbf{q} \gamma_\uparrow$ , are

$$q = \{\operatorname{Re} \varphi^\uparrow, \operatorname{Im} \varphi^\downarrow, -\operatorname{Re} \varphi^\downarrow, \operatorname{Im} \varphi^\uparrow\} \leftrightarrow \mathbf{q} = \begin{pmatrix} \varphi^\uparrow & -\varphi^{\downarrow*} \\ \varphi^\downarrow & \varphi^{\uparrow*} \end{pmatrix}. \quad (13.108)$$

Show that the reverse quaternion  $\bar{q}$  and the equivalent  $2 \times 2$  matrix  $\bar{\mathbf{q}}$  in the Pauli representation are

$$\bar{q} = \{\operatorname{Re} \varphi^\uparrow, -\operatorname{Im} \varphi^\downarrow, \operatorname{Re} \varphi^\downarrow, -\operatorname{Im} \varphi^\uparrow\} \leftrightarrow \bar{\mathbf{q}} = \begin{pmatrix} \varphi^{\uparrow*} & \varphi^{\downarrow*} \\ -\varphi^\downarrow & \varphi^\uparrow \end{pmatrix}. \quad (13.109)$$

Conclude that the reverse matrix  $\bar{\mathbf{q}}$  equals its Hermitian conjugate,  $\bar{\mathbf{q}} = \mathbf{q}^\dagger$ , and that the reverse Pauli spinor defined by  $\bar{\varphi} \equiv \gamma_\uparrow^\top \bar{\mathbf{q}}$  is

$$\bar{\varphi} \equiv \gamma_\uparrow^\top \bar{\mathbf{q}} = \gamma_\uparrow^\top \mathbf{q}^\dagger = (\varphi^{\uparrow*} \quad \varphi^{\downarrow*}) = \varphi^\dagger. \quad (13.110)$$

**Exercise 13.8 Translate a quaternion into a Pauli spinor.** Show that the quaternion  $q \equiv w + ix + jy + kz$  is equivalent in the Pauli representation (13.97) to the  $2 \times 2$  matrix  $\mathbf{q}$

$$q = \{w, x, y, z\} \leftrightarrow \mathbf{q} = \begin{pmatrix} w + iz & ix + y \\ ix - y & w - iz \end{pmatrix}. \quad (13.111)$$

Conclude that the Pauli spinor  $\varphi = \mathbf{q} \gamma_\uparrow$  corresponding to the quaternion  $q$  is

$$\varphi \equiv \mathbf{q} \gamma_\uparrow = \begin{pmatrix} w + iz \\ ix - y \end{pmatrix}, \quad (13.112)$$

and that the reverse spinor  $\bar{\varphi}$  is

$$\bar{\varphi} = \varphi^\dagger = \gamma_\uparrow^\top \mathbf{q}^\dagger = (w - iz \quad -ix - y). \quad (13.113)$$

Hence conclude that  $\bar{\varphi} \varphi$  is the real positive scalar magnitude squared  $\lambda^2 = \bar{q} q$  of the quaternion  $q$ ,

$$\bar{\varphi} \varphi = \varphi^\dagger \varphi = \bar{q} q = \lambda^2, \quad (13.114)$$

with

$$\lambda^2 = w^2 + x^2 + y^2 + z^2. \quad (13.115)$$

### 13.16 Spin axis

In the Pauli representation, the spinor basis elements  $\gamma_a$  are eigenvectors of the Pauli operator  $\sigma_3$  with eigenvalues  $\pm 1$ ,

$$\sigma_3 \gamma_{\uparrow} = +\gamma_{\uparrow}, \quad \sigma_3 \gamma_{\downarrow} = -\gamma_{\downarrow}. \quad (13.116)$$

The **spin axis** of a Pauli spinor  $\varphi$  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  $\gamma_{\uparrow}$  is the positive 3-axis, while the spin axis of the spin-down basis spinor  $\gamma_{\downarrow}$  is the negative 3-axis. The spin axis of a Pauli spinor  $\varphi = \lambda R \gamma_{\uparrow}$  is the unit direction  $\{n_1, n_2, n_3\}$  of the rotated  $z$ -axis, given by

$$\sigma_a n_a = R \sigma_3 \bar{R}. \quad (13.117)$$

Equation (13.117) is confirmed by the fact that  $\sigma_a n_a$  has eigenvalue  $+1$  acting on  $\varphi$ :

$$\sigma_a n_a \varphi = (R \sigma_3 \bar{R}) (\lambda R \gamma_{\uparrow}) = \lambda R \sigma_3 \gamma_{\uparrow} = \lambda R \gamma_{\uparrow} = \varphi. \quad (13.118)$$

**Exercise 13.9 Orthonormal eigenvectors of the spin operator.** Show that, in the Pauli representation, the orthonormal eigenvectors  $\gamma_{\uparrow n}$  and  $\gamma_{\downarrow n}$  of the spin operator  $\sigma_a n_a$  projected along the unit direction  $\{n_1, n_2, n_3\}$  are

$$\gamma_{\uparrow n} = \frac{1}{\sqrt{2(1+n_3)}} \begin{pmatrix} 1+n_3 \\ n_1+in_2 \end{pmatrix}, \quad \gamma_{\downarrow n} = \frac{1}{\sqrt{2(1-n_3)}} \begin{pmatrix} -1+n_3 \\ n_1+in_2 \end{pmatrix}. \quad (13.119)$$

# 14

---

## 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 one of the dimensions is timelike and the others are spacelike. Happily, the elegant formalism of the geometric algebra carries through to the spacetime algebra.

### 14.1 Spacetime algebra

Let  $\gamma_m$  ( $m = 0, 1, 2, 3$ ) denote an orthonormal basis of spacetime, with  $\gamma_0$  representing the time axis, and  $\gamma_a$  ( $a = 1, 2, 3$ ) the spatial axes. Geometric multiplication in the spacetime algebra is defined by

$$\gamma_m \gamma_n = \gamma_m \cdot \gamma_n + \gamma_m \wedge \gamma_n \quad (14.1)$$

just as in the geometric algebra, equation (13.5). The key difference between the spacetime basis  $\gamma_m$  and Euclidean bases is that scalar products of the basis vectors  $\gamma_m$  form the Minkowski metric  $\eta_{mn}$ ,

$$\gamma_m \cdot \gamma_n = \eta_{mn} \quad (14.2)$$

whereas scalar products of Euclidean basis elements  $\gamma_a$  formed the unit matrix,  $\gamma_a \cdot \gamma_b = \delta_{ab}$ , equation (13.6). 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  $\gamma_m$  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  $\gamma$ -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  $\gamma$ -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

$$\mathbf{a} = a^m \boldsymbol{\gamma}_m = a^0 \boldsymbol{\gamma}_0 + a^1 \boldsymbol{\gamma}_1 + a^2 \boldsymbol{\gamma}_2 + a^3 \boldsymbol{\gamma}_3 . \quad (14.6)$$

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

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b} = a^m b^n \boldsymbol{\gamma}_m \cdot \boldsymbol{\gamma}_n + a^m b^n \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n = a^m b^n \eta_{mn} + \frac{1}{2} a^m b^n [\boldsymbol{\gamma}_m, \boldsymbol{\gamma}_n] . \quad (14.7)$$

It is convenient to denote three of the six bivectors of the spacetime algebra by  $\boldsymbol{\sigma}_a$ ,

$$\boldsymbol{\sigma}_a \equiv \boldsymbol{\gamma}_0 \boldsymbol{\gamma}_a \quad (a = 1, 2, 3) . \quad (14.8)$$

The symbol  $\boldsymbol{\sigma}_a$  is used because the algebra of bivectors  $\boldsymbol{\sigma}_a$  is isomorphic to the algebra of Pauli matrices  $\sigma_a$ . The pseudoscalar, the highest grade basis element of the spacetime algebra, is denoted  $I$

$$\boldsymbol{\gamma}_0 \boldsymbol{\gamma}_1 \boldsymbol{\gamma}_2 \boldsymbol{\gamma}_3 = \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_2 \boldsymbol{\sigma}_3 = I . \quad (14.9)$$

The pseudoscalar  $I$  satisfies

$$I^2 = -1 , \quad I \boldsymbol{\gamma}_m = -\boldsymbol{\gamma}_m I , \quad I \boldsymbol{\sigma}_a = \boldsymbol{\sigma}_a I . \quad (14.10)$$

The basis elements of the 4-dimensional spacetime algebra are then

1 ,	$\boldsymbol{\gamma}_m$ ,	$\boldsymbol{\sigma}_a$ ,	$I \boldsymbol{\sigma}_a$ ,	$I \boldsymbol{\gamma}_m$ ,	$I$ ,
1 scalar	4 vectors	6 bivectors	4 pseudovectors	1 pseudoscalar	

(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 0 or 1, modulo 4, and changing the sign of multivectors of grade 2 or 3, modulo 4:

$$\bar{I} = 1 , \quad \bar{\boldsymbol{\gamma}}_m = \boldsymbol{\gamma}_m , \quad \bar{\boldsymbol{\sigma}}_a = -\boldsymbol{\sigma}_a , \quad \bar{I} \bar{\boldsymbol{\sigma}}_a = -I \boldsymbol{\sigma}_a , \quad \bar{I} \bar{\boldsymbol{\gamma}}_m = -I \boldsymbol{\gamma}_m , \quad \bar{I} = I . \quad (14.12)$$

The mapping

$$\boldsymbol{\gamma}_a^{(3)} \leftrightarrow \boldsymbol{\sigma}_a \quad (a = 1, 2, 3) \quad (14.13)$$

(the superscript <sup>(3)</sup> 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 (14.13) implies the equivalence of the 3D spatial pseudoscalar  $I_3$  and the 4D spacetime pseudoscalar  $I$ ,

$$I_3 \leftrightarrow I , \quad (14.14)$$

since  $I_3 = \boldsymbol{\gamma}_1^{(3)} \boldsymbol{\gamma}_2^{(3)} \boldsymbol{\gamma}_3^{(3)}$  and  $I = \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_2 \boldsymbol{\sigma}_3$ .

## 14.2 Complex quaternions

A **complex quaternion** (also called a **biquaternion** by W. R. Hamilton) is a quaternion

$$q = a + \mathbf{b} = a + \imath_a b_a , \quad (14.15)$$

in which the four coefficients  $a, b_a$  ( $a = 1, 2, 3$ ) are each complex numbers

$$a = a_R + Ia_I , \quad b_a = b_{a,R} + Ib_{a,I} . \quad (14.16)$$

The imaginary  $I$  is taken to commute with each of the quaternionic imaginaries  $\imath_a$ . The choice of symbol  $I$  is deliberate: in the isomorphism (14.31) 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 (13.78)–(13.83). The quaternionic conjugate  $\bar{q}$  of a complex quaternion  $q \equiv a + \mathbf{b}$  is conjugated with respect to the quaternionic imaginaries  $\imath_a$ , but the complex coefficients  $a$  and  $b_a$  are *not* conjugated with respect to the complex imaginary  $I$ ,

$$\bar{q} = a + \bar{\mathbf{b}} = a - \mathbf{b} = a - \imath_a b_a . \quad (14.17)$$

The magnitude  $|q|$  of a complex quaternion  $q \equiv a + \mathbf{b}$ ,

$$|q| = (\bar{q}q)^{1/2} = (q\bar{q})^{1/2} = (a^2 + \mathbf{b} \cdot \mathbf{b})^{1/2} = (a^2 + b_a b_a)^{1/2} , \quad (14.18)$$

is a complex number, not a real number. 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  $i$ )

$$q^* = a^* + \mathbf{b}^* = a^* + \imath_a b_a^* , \quad (14.19)$$

in which the complex coefficients  $a$  and  $b_a$  are conjugated with respect to the imaginary  $I$ , but the quaternionic imaginaries  $\imath_a$  are *not* conjugated.

A non-zero complex quaternion can have zero magnitude (unlike a real quaternion), in which case it is **null**. The null condition

$$\bar{q}q = a^2 + b_a b_a = 0 \quad (14.20)$$

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

$$q = p(1 + In) = p(1 + I\imath_a n_a) , \quad (14.21)$$

where  $p$  is a real quaternion, and  $\mathbf{n} = \iota_a n_a$  is a real unit vector quaternion, with real components  $\{n_1, n_2, n_3\}$  satisfying  $n_a n_a = 1$ . Equivalently,

$$q = (1 + I\mathbf{n}')p = (1 + I\iota_a n'_a)p , \quad (14.22)$$

where  $\mathbf{n}'$  is the real unit vector quaternion

$$\mathbf{n}' = \frac{p\mathbf{n}\bar{p}}{|p|^2} , \quad (14.23)$$

with components  $\{n'_1, n'_2, n'_3\}$  satisfying  $n'_a n'_a = 1$ .

**Solution.** Write the null quaternion  $q$  as

$$q = p + Ir \quad (14.24)$$

where  $p$  and  $r$  are real quaternions, both of which must be non-zero if  $q$  is non-trivial. Then equation (14.21) is true with

$$\mathbf{n} = \iota_a n_a = \frac{\bar{p}r}{|p|^2} . \quad (14.25)$$

The null condition is  $\bar{q}q = 0$ . The vanishing of the real part,  $\text{Re}(\bar{q}q) = \bar{p}p - \bar{r}r = 0$ , shows that  $|p|^2 = |r|^2$ . The vanishing of the imaginary ( $I$ ) part,  $\text{Im}(\bar{q}q) = \bar{p}r + \bar{r}p = \bar{p}r + \bar{\bar{p}r} = 0$  shows that the  $\bar{p}r$  must be a pure quaternionic imaginary, since the quaternionic conjugate of  $\bar{p}r$  is minus itself, so  $\bar{p}r/|p|^2$  must be of the form  $\mathbf{n} = \iota_a n_a$ . Its squared magnitude  $\mathbf{n}\bar{\mathbf{n}} = n_a n_a = \bar{p}r \bar{\bar{p}r}/|p|^4 = 1$  is unity, so  $\mathbf{n}$  is a unit 3-vector quaternion. It follows immediately from the manner of construction that the expression (14.21) 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

$$\mathbf{q} = \iota_a q_a \quad \text{with} \quad \mathbf{q} \cdot \mathbf{q} = q_a q_a = 0 \quad (14.26)$$

is nilpotent,

$$\mathbf{q}^2 = 0 . \quad (14.27)$$

Prove that a nilpotent complex quaternion must take the form (14.26). 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. Such rotations correspond, in the usual way, to even, unimodular elements of the geometric algebra of spacetime. The basis elements of the even spacetime algebra

are

$$\begin{array}{ccc} 1, & \boldsymbol{\sigma}_a, & I\boldsymbol{\sigma}_a, \\ 1 \text{ scalar} & 6 \text{ bivectors} & 1 \text{ pseudoscalar} \end{array} \quad (14.28)$$

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  $8 \div 2 = 4$  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

$$\begin{array}{ccc} 1, & I\boldsymbol{\sigma}_a, \\ 1 \text{ scalar} & 3 \text{ bivectors} \end{array} \quad (14.29)$$

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 (14.29) is that the basis  $\{1, I\boldsymbol{\sigma}_a\}$  is equivalent to the basis (13.84) of the even algebra of 3-dimensional Euclidean space through the isomorphism (14.13) and (14.14). This basis in turn is equivalent to the quaternionic basis  $\{1, \iota_a\}$  through the isomorphism (13.90):

$$I\boldsymbol{\sigma}_a \leftrightarrow I_3\boldsymbol{\gamma}_a^{(3)} \leftrightarrow \iota_a \quad (a = 1, 2, 3). \quad (14.30)$$

In other words, the even spacetime algebra is isomorphic to the algebra of quaternions with complex coefficients:

$$a + I\boldsymbol{\sigma}_a b_a \leftrightarrow a + \iota_a b_a \quad (14.31)$$

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 (14.31) between even elements of the spacetime algebra and complex quaternions implies that the group of Lorentz rotors, which are unimodular elements of the even spacetime algebra, is isomorphic to the group of unimodular complex quaternions

$$\text{spacetime rotors} \leftrightarrow \text{unimodular complex quaternions}. \quad (14.32)$$

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  $\theta$  about the unit complex direction  $n_a$  is, according to equation (13.47),

$$R = e^{-\theta/2} = e^{-\mathbf{n}\theta/2} = \cos \frac{\theta}{2} - \mathbf{n} \sin \frac{\theta}{2},$$

(14.33)

generalizing the 3D rotor (13.87). Here  $\theta$  is a bivector

$$\theta = \mathbf{n}\theta = I\boldsymbol{\sigma}_a n_a \theta, \quad (14.34)$$

whose magnitude is the complex angle  $(\bar{\theta}\theta)^{1/2} = \theta \equiv \theta_R + I\theta_I$ , and whose direction is the unit complex bivector  $\mathbf{n} = \mathbf{n}_R + I\mathbf{n}_I$ . The unit condition  $\bar{\mathbf{n}}\mathbf{n} = 1$  on  $\mathbf{n}$  is equivalent to the complex condition  $n_a n_a = 1$

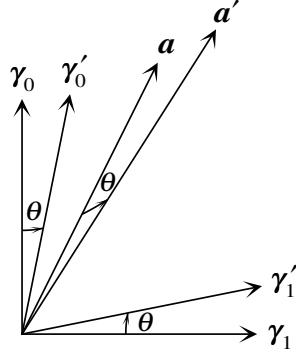


Figure 14.1 Lorentz boost of a vector  $\mathbf{a}$  by rapidity  $\theta$  in the  $\gamma_0$ - $\gamma_1$  plane. See Exercise 14.3.

on the components  $n_a \equiv \{n_1, n_2, n_3\}$ . The real and imaginary parts of the unit condition imply the two conditions

$$n_{a,R} n_{a,R} - n_{a,I} n_{a,I} = 1 , \quad 2 n_{R,a} n_{I,a} = 0 . \quad (14.35)$$

The complex angle  $\theta$  has 2 degrees of freedom, while the complex unit 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 (14.30), the Lorentz rotor  $R$  given by equation (14.33) can be reinterpreted as a complex quaternion, with  $\boldsymbol{\theta}$  the complex quaternion

$$\boldsymbol{\theta} = \mathbf{n} \theta = \iota_a n_a \theta , \quad (14.36)$$

whose complex magnitude is  $\theta = |\boldsymbol{\theta}| \equiv (\bar{\boldsymbol{\theta}} \boldsymbol{\theta})^{1/2}$  and whose complex unit direction is  $\mathbf{n} \equiv \iota_a n_a$ . The associated reverse rotor  $\bar{R}$  is

$$\bar{R} = e^{\boldsymbol{\theta}/2} = e^{\mathbf{n} \theta/2} = \cos \frac{\theta}{2} + \mathbf{n} \sin \frac{\theta}{2} \quad (14.37)$$

the quaternionic conjugate of  $R$ . Note that  $\theta$  and  $\mathbf{n}$  in equation (14.37) are *not* conjugated with respect to the imaginary  $I$ . The sine and cosine of the complex angle  $\theta$  appearing in equations (14.33) and (14.37) are related to its real and imaginary parts in the usual way,

$$\cos \frac{\theta}{2} = \cos \frac{\theta_R}{2} \cosh \frac{\theta_I}{2} - I \sin \frac{\theta_R}{2} \sinh \frac{\theta_I}{2} , \quad \sin \frac{\theta}{2} = \sin \frac{\theta_R}{2} \cosh \frac{\theta_I}{2} + I \cos \frac{\theta_R}{2} \sinh \frac{\theta_I}{2} . \quad (14.38)$$

For the case of a pure spatial rotation, the angle  $\theta = \theta_R$  and axis  $\mathbf{n} = \mathbf{n}_R$  in the rotor (14.33) are both real. The rotor corresponding to a pure spatial rotation by angle  $\theta_R$  right-handedly about unit real axis  $\mathbf{n}_R \equiv I \boldsymbol{\sigma}_a n_{a,R} = \iota_a n_{a,R}$  is the real quaternion

$$R = e^{-\mathbf{n}_R \theta_R/2} = \cos \frac{\theta_R}{2} - \mathbf{n}_R \sin \frac{\theta_R}{2} . \quad (14.39)$$

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  $\gamma_1$ -axis (the  $x$ -axis) is a rotation of spacetime in the  $\gamma_0-\gamma_1$  plane (the  $t-x$  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  $\theta_I$ , 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

$$R = e^{-I\mathbf{n}_R \theta_I/2} = \cosh \frac{\theta_I}{2} - I\mathbf{n}_R \sinh \frac{\theta_I}{2}. \quad (14.40)$$

**Exercise 14.3 Lorentz boost.** A Lorentz boost by rapidity  $\theta = \operatorname{atanh} v$  along the  $\gamma_1$ -axis ( $x$ -axis) (that is, a rotation in the  $\gamma_0-\gamma_1$  plane) is given by the Lorentz rotor

$$R = e^{-I\mathbf{n}_1 \theta/2} = \cosh \frac{\theta}{2} + \gamma_0 \wedge \gamma_1 \sinh \frac{\theta}{2}. \quad (14.41)$$

Confirm that the Lorentz boost transforms the axes  $\gamma_m$  as

$$R : \gamma_0 \rightarrow R\gamma_0 \bar{R} = \gamma_0 \cosh \theta + \gamma_1 \sinh \theta, \quad (14.42a)$$

$$R : \gamma_1 \rightarrow R\gamma_1 \bar{R} = \gamma_1 \cosh \theta + \gamma_0 \sinh \theta, \quad (14.42b)$$

$$R : \gamma_a \rightarrow R\gamma_a \bar{R} = \gamma_a \quad (a \neq 0, 1). \quad (14.42c)$$

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  $R = e^{-\iota_a n_a \theta/2}$  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

$$R = p + Iq \quad (14.43)$$

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  $L$

$$R = LU \quad (14.44)$$

in which

$$U = \frac{p}{|p|}, \quad L = |p| + I \frac{q\bar{p}}{|p|} \quad (14.45)$$

where  $|p| = (\bar{p}p)^{1/2}$  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,  $U\bar{U} = 1$ , and it follows that the boost rotor  $L = R\bar{U}$  is also unimodular, since  $R$  is unimodular. The spatial rotor  $U$  is a real quaternion, and therefore satisfies the form (14.39) of a pure spatial rotation. The real part  $|p|$  of the boost rotor  $L$  is pure real, while the imaginary part  $q\bar{p}/|p|$  is a pure quaternionic

imaginary, since unimodularity  $R\bar{R} = 1$  implies that  $\text{Im}(R\bar{R}) = q\bar{p} + p\bar{q} = q\bar{p} + \bar{q}\bar{p} = 0$ , i.e. the quaternionic conjugate of  $q\bar{p}$  is minus itself. Thus  $L$  satisfies the form (14.40) 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 (14.44), is

$$R = UL' \quad (14.46)$$

where  $U$  is the same spatial rotor as before, but the boost rotor  $L'$  is

$$L' = |p| + I \frac{\bar{p}q}{|p|} = \bar{U}LU \quad (14.47)$$

whose real part  $|p|$  is the same as for  $L$ , but whose imaginary part  $\bar{p}q/|p|$  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 all spatial directions while keeping the time direction unchanged. Spatial inversion may be accomplished by reflecting the spatial vector basis elements  $\gamma_a \rightarrow -\gamma_a$ , while keeping the time vector basis element  $\gamma_0$  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  $\iota_a$  are unchanged. Thus, if multivectors in the geometric spacetime algebra are written as linear combinations of products of  $\gamma_0$ ,  $\iota_a$ , and  $I$ , then spatial inversion  $P$  corresponds to the transformation

$$P : \gamma_0 \rightarrow \gamma_0, \quad \iota_a \rightarrow \iota_a, \quad I \rightarrow -I. \quad (14.48)$$

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 may be accomplished by reflecting the time vector basis element  $\gamma_0 \rightarrow -\gamma_0$ , while keeping the spatial vector basis elements  $\gamma_a$  unchanged. As with spatial inversion, this results in  $\sigma_a \rightarrow -\sigma_a$  and  $I \rightarrow -I$ , which keeps  $I\sigma_a$  hence  $\iota_a$  unchanged. If multivectors in the geometric spacetime algebra are written as linear combinations of products of  $\gamma_0$ ,  $\iota_a$ , and  $I$ , then time inversion  $T$  corresponds to the transformation

$$T : \gamma_0 \rightarrow -\gamma_0, \quad \iota \rightarrow \iota_a, \quad I \rightarrow -I. \quad (14.49)$$

For any multivector, time inversion corresponds to the instruction to flip  $\gamma_0$  and take the complex conjugate (with respect to  $I$ ).

The combined operation  $PT$  of inverting both space and time directions corresponds to

$$PT : \gamma_0 \rightarrow -\gamma_0, \quad \iota \rightarrow \iota_a, \quad I \rightarrow I. \quad (14.50)$$

For any multivector, spacetime inversion corresponds to the instruction to flip  $\gamma_0$ , while keeping  $\iota_a$  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 14.7). 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

$$q = w + \iota_1 x + \iota_2 y + \iota_3 z \quad (14.51)$$

with complex coefficients  $w, x, y, z$  (so  $w = w_R + Iw_I$ , etc.) can be stored as the 8-component object

$$q = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{Bmatrix}, \quad (14.52)$$

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  $\bar{q}$  of the quaternion (14.52) is

$$\bar{q} = \begin{Bmatrix} w_R & -x_R & -y_R & -z_R \\ w_I & -x_I & -y_I & -z_I \end{Bmatrix}, \quad (14.53)$$

while its complex conjugate  $q^*$  (with respect to  $I$ ) is

$$q^* = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ -w_I & -x_I & -y_I & -z_I \end{Bmatrix}. \quad (14.54)$$

A Lorentz rotor  $R$  corresponds to a complex quaternion of unit modulus. The unimodular condition  $R\bar{R} = 1$ , 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  $\theta$  right-handedly about the  $x$ -axis (the  $x$ -axis) is the real Lorentz rotor

$$R = e^{-\iota_1 \theta/2} = \cos(\theta/2) - \iota_1 \sin(\theta/2), \quad (14.55)$$

or, stored as a complex quaternion,

$$R = \begin{Bmatrix} \cos(\theta/2) & -\sin(\theta/2) & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Bmatrix}. \quad (14.56)$$

Note that this is the physics convention, where a right-handed rotation corresponds to  $R = e^{-i_a n_a \theta/2}$  and rotations accumulate to the left. The convention in OpenGL and other graphics software is that  $R = e^{i_a n_a \theta/2}$  and rotations accumulate to the right. To change to OpenGL convention, omit the minus sign in equation (14.56). 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

$$R = e^{-I i_1 \theta/2} = \cosh(\theta/2) - I i_1 \sinh(\theta/2), \quad (14.57)$$

or, stored as a complex quaternion,

$$R = \begin{Bmatrix} \cosh(\theta/2) & 0 & 0 & 0 \\ 0 & -\sinh(\theta/2) & 0 & 0 \end{Bmatrix}. \quad (14.58)$$

Again, this is the physics convention. To change to OpenGL convention, omit the minus sign in equation (14.58). 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  $\bar{R}$ .

Any even multivector  $\mathbf{q}$  is equivalent to a complex quaternion by the isomorphism (14.31). According to the usual transformation law (13.54) for multivectors, the rule for Lorentz transforming an even multivector  $\mathbf{q}$  is

$$\boxed{R : \mathbf{q} \rightarrow R\mathbf{q}\bar{R}} \quad (\text{even multivector}). \quad (14.59)$$

The transformation (14.59) instructs to multiply three complex quaternions  $R$ ,  $\mathbf{q}$ , and  $\bar{R}$ , a one-line expression in a c++ program. In OpenGL convention, the transformation rule is  $\mathbf{q} \rightarrow \bar{R}\mathbf{q}R$ .

As an example of an even multivector, the electromagnetic field  $\mathbf{F}$  is a bivector in the spacetime algebra,

$$\mathbf{F} = \frac{1}{2} F^{mn} \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n, \quad (14.60)$$

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  $\mathbf{E} = E_a i_a$  and  $\mathbf{B} = B_a i_a$

$$\mathbf{F} = -I(\mathbf{E} + I\mathbf{B}). \quad (14.61)$$

Under the parity transformation  $P$  (14.48), the electric field  $\mathbf{E}$  changes sign, whereas the magnetic field  $\mathbf{B}$  does not, which is as it should be:

$$\boxed{P : \mathbf{E} \rightarrow -\mathbf{E}, \quad \mathbf{B} \rightarrow \mathbf{B}}. \quad (14.62)$$

In view of the isomorphism (14.31), the electromagnetic field bivector  $\mathbf{F}$  can be written as the complex quaternion

$$\mathbf{F} = \begin{Bmatrix} 0 & B_1 & B_2 & B_3 \\ 0 & -E_1 & -E_2 & -E_3 \end{Bmatrix}. \quad (14.63)$$

According to the rule (14.59), the electromagnetic field bivector  $\mathbf{F}$  Lorentz transforms as  $\mathbf{F} \rightarrow R\mathbf{F}\bar{R}$ , which is a powerful and elegant way to Lorentz transform the electromagnetic field.

A 4-vector  $\mathbf{a} \equiv \boldsymbol{\gamma}_m a^m$  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  $I\mathbf{b} = I\boldsymbol{\gamma}_m b^m$ . The odd multivector can be written as the product of the time basis vector  $\boldsymbol{\gamma}_0$  and an even multivector  $\mathbf{q}$

$$\mathbf{a} + I\mathbf{b} = \boldsymbol{\gamma}_0 \mathbf{q} = \boldsymbol{\gamma}_0 (a^0 + I\iota_a a^a - Ib^0 + \iota_a b^a). \quad (14.64)$$

By the isomorphism (14.31), the even multivector  $\mathbf{q}$  is equivalent to the complex quaternion

$$\mathbf{q} = \begin{Bmatrix} a^0 & b^1 & b^2 & b^3 \\ -b^0 & a^1 & a^2 & a^3 \end{Bmatrix}. \quad (14.65)$$

According to the usual transformation law (13.54) for multivectors, the rule for Lorentz transforming the odd multivector  $\boldsymbol{\gamma}_0 \mathbf{q}$  is

$$R : \boldsymbol{\gamma}_0 \mathbf{q} \rightarrow R\boldsymbol{\gamma}_0 \mathbf{q} \bar{R} = \boldsymbol{\gamma}_0 R^* \mathbf{q} \bar{R}. \quad (14.66)$$

In the last expression of (14.66), the factor  $\boldsymbol{\gamma}_0$  has been brought to the left, to be consistent with the convention (14.64) that an odd multivector is  $\boldsymbol{\gamma}_0$  on the left times an even multivector on the right. Notice that commuting  $\boldsymbol{\gamma}_0$  through  $R$  converts the latter to its complex conjugate (with respect to  $I$ )  $R^*$ , which is true because  $\boldsymbol{\gamma}_0$  commutes with the quaternionic imaginaries  $\iota_a$ , but anticommutes with the pseudoscalar  $I$ . Thus if the components of an odd multivector are stored as a complex quaternion (14.65), then that complex quaternion  $\mathbf{q}$  Lorentz transforms as

$$R : \mathbf{q} \rightarrow R^* \mathbf{q} \bar{R} \quad (\text{odd multivector}). \quad (14.67)$$

In OpenGL convention,  $\mathbf{q} \rightarrow \bar{R}^* \mathbf{q} R$ . The rule (14.67) again instructs to multiply three complex quaternions  $R^*$ ,  $\mathbf{q}$ , and  $\bar{R}$ , a one-line expression in a C++ program. The transformation rule (14.67) for an odd multivector encoded as a complex quaternion differs from that (14.59) 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 (14.65), and Lorentz transformed simultaneously, enabling two transformations to be done for the price of one.

Finally, a Dirac spinor  $\psi$  is equivalent to a complex quaternion  $q$  (§14.9). It Lorentz transforms as

$$R : q \rightarrow Rq \quad (\text{spinor}). \quad (14.68)$$

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  $R(x)$  that corresponds to uniform rotation and acceleration between initial and final Lorentz rotors  $R_0$  and  $R_1$  as the parameter  $x$  varies uniformly from 0 to 1 is

$$R(x) = R_0 \exp [x \ln(R_1/R_0)] . \quad (14.69)$$

What are the exponential and logarithm of a complex quaternion in terms of its components? Address the issue of the multi-valued character of the logarithm.

**Exercise 14.7 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

$$f(x) = f_0 + f'_0 x + [3(f_1 - f_0) - 2f'_0 - f'_1] x^2 + [2(f_0 - f_1) + f'_0 + f'_1] x^3 . \quad (14.70)$$

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 14.8 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 (14.67) is a much better idea than Lorentz transforming by multiplying by an explicit  $4 \times 4$  matrix. Suppose that the Lorentz rotor  $R$  is the complex quaternion

$$R = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{Bmatrix} . \quad (14.71)$$

Show that the Lorentz transformation (14.67) 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  $4 \times 4$  rotation matrix is written to the left of the 4-vector in accordance with the physics convention that rotations accumulate to the left):

$$R : \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix} \rightarrow \begin{pmatrix} |w|^2 + |x|^2 + |y|^2 + |z|^2 & 2(-w_R x_I + w_I x_R + y_R z_I - y_I z_R) \\ 2(-w_R x_I + w_I x_R - y_R z_I + y_I z_R) & |w|^2 + |x|^2 - |y|^2 - |z|^2 \\ 2(-w_R y_I + w_I y_R - z_R x_I + z_I x_R) & 2(x_R y_R + x_I y_I + w_R z_R + w_I z_I) \\ 2(-w_R z_I + w_I z_R - x_R y_I + x_I y_R) & 2(z_R x_R + z_I x_I - w_R y_R - w_I y_I) \\ 2(-w_R y_I + w_I y_R + z_R x_I - z_I x_R) & 2(-w_R z_I + w_I z_R + x_R y_I - x_I y_R) \\ 2(x_R y_R + x_I y_I - w_R z_R - w_I z_I) & 2(z_R x_R + z_I x_I + w_R y_R + w_I y_I) \\ |w|^2 - |x|^2 + |y|^2 - |z|^2 & 2(y_R z_R + y_I z_I - w_R x_R - w_I x_I) \\ 2(y_R z_R + y_I z_I + w_R x_R + w_I x_I) & |w|^2 - |x|^2 - |y|^2 + |z|^2 \end{pmatrix} \begin{pmatrix} a^0 \gamma_0 \\ a^1 \gamma_1 \\ a^2 \gamma_2 \\ a^3 \gamma_3 \end{pmatrix} , \quad (14.72)$$

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 (14.72) in the case of a Lorentz boost by velocity  $v$  along the 1-axis, where the rotor  $R$  takes the form (14.41), is

$$R : \begin{pmatrix} a^0 \boldsymbol{\gamma}_0 \\ a^1 \boldsymbol{\gamma}_1 \\ a^2 \boldsymbol{\gamma}_2 \\ a^3 \boldsymbol{\gamma}_3 \end{pmatrix} \rightarrow \begin{pmatrix} \gamma & \gamma v & 0 & 0 \\ \gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a^0 \boldsymbol{\gamma}_0 \\ a^1 \boldsymbol{\gamma}_1 \\ a^2 \boldsymbol{\gamma}_2 \\ a^3 \boldsymbol{\gamma}_3 \end{pmatrix}, \quad (14.73)$$

with  $\gamma$  the familiar Lorentz gamma factor

$$\gamma = \cosh \theta = \frac{1}{\sqrt{1-v^2}}, \quad \gamma v = \sinh \theta = \frac{v}{\sqrt{1-v^2}}. \quad (14.74)$$


---

## 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  $\xi^\mu$ , §7.32, which shifts the coordinates by an infinitesimal amount,

$$x^\mu \rightarrow x^\mu + \epsilon \xi^\mu, \quad (14.75)$$

with  $\epsilon$  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 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

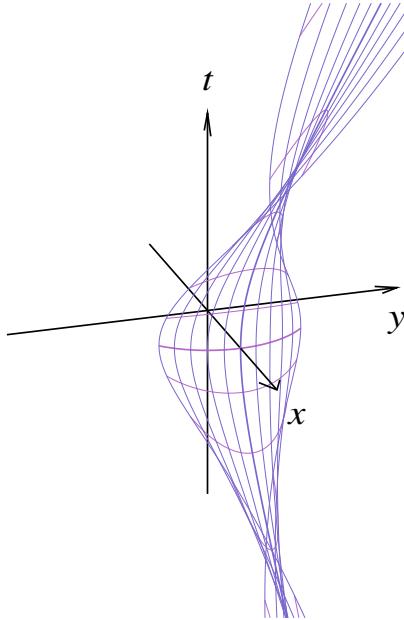


Figure 14.2 3D spacetime diagram of a sample of (blue) timelike Killing trajectories in Minkowski space. The two outermost of the trajectories shown lie on the light cylinder, and are lightlike. Motion in the  $z$  spatial direction is suppressed. The trajectories accelerate with uniform proper linear acceleration in the  $x$ -direction, and with uniform rotation about the  $y$ - $z$  plane. The trajectories shown are for the case of a Killing vector with equal linear and rotational components,  $\mu = 1$ , equation (14.83). The crossing (purple) lines are spacelike lines of constant argument  $\lambda$  of the transformation.

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 (14.61). The complex magnitude squared  $|\boldsymbol{\theta}|^2$  of the bivector,

$$|\boldsymbol{\theta}|^2 \equiv \bar{\boldsymbol{\theta}}\boldsymbol{\theta} = -\boldsymbol{\theta}^2 = \boldsymbol{\theta}_I^2 - \boldsymbol{\theta}_R^2 - 2I\boldsymbol{\theta}_R \cdot \boldsymbol{\theta}_I , \quad (14.76)$$

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 magnitude is fixed, that is,  $\boldsymbol{\theta}_I^2 - \boldsymbol{\theta}_R^2$  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 unit spatial direction,  $\mathbf{n}$ , say. So transformed, the bivector  $\boldsymbol{\theta}$  is the complex quaternion

$$\boldsymbol{\theta} = (\theta_R + I\theta_I) \mathbf{i} \cdot \mathbf{n} . \quad (14.77)$$

The bivector (14.77) generates a uniform proper spatial rotation about the  $\mathbf{n}$  axis, coupled with a uniform proper acceleration along the  $\mathbf{n}$  axis. A Killing trajectory  $\mathbf{x}(\lambda) \equiv x^m(\lambda) \boldsymbol{\gamma}_m$ , parametrized by real argument  $\lambda$  along the trajectory, is obtained by Lorentz transforming an initial 4-vector  $\mathbf{x}_0 \equiv \mathbf{x}(0)$  by a rotor  $R \equiv e^{-\lambda\theta/2}$ , equation (14.33),

$$\mathbf{x} = R\mathbf{x}_0\bar{R}. \quad (14.78)$$

Define  $\alpha$  and  $\beta$  by

$$\alpha \equiv \lambda\theta_R, \quad \beta \equiv \lambda\theta_I. \quad (14.79)$$

If the unit direction  $\mathbf{n}$  is taken to be the  $x$ -direction, then Minkowski coordinates  $x^m \equiv \{t, x, y, z\}$  along a Killing trajectory (14.78) starting at  $x_0^m \equiv \{0, x_0, r_0 \cos \phi_0, r_0 \sin \phi_0\}$  are

$$t = x_0 \sinh \beta, \quad (14.80a)$$

$$x = x_0 \cosh \beta, \quad (14.80b)$$

$$y = r_0 \cos(\phi_0 + \alpha), \quad (14.80c)$$

$$z = r_0 \sin(\phi_0 + \alpha). \quad (14.80d)$$

The metric in comoving Killing coordinates is

$$ds^2 = -x_0^2 d\beta^2 + dx_0^2 + dr_0^2 + r_0^2(d\phi_0 + d\alpha)^2. \quad (14.81)$$

The proper time along a Killing trajectory  $dx_0 = dr_0 = d\phi_0 = 0$  is

$$d\tau = \sqrt{x_0^2 d\beta^2 - r_0^2 d\alpha^2} = d\beta \sqrt{x_0^2 - \mu^2 r_0^2}, \quad (14.82)$$

where  $\mu$  is the real constant ratio

$$\mu \equiv \frac{d\alpha}{d\beta} = \frac{\theta_R}{\theta_I}. \quad (14.83)$$

A trajectory is timelike provided that

$$|x_0| > |\mu r_0|. \quad (14.84)$$

Null trajectories  $|x_0| = |\mu r_0|$  define the **light cylinder**. Trajectories outside the light cylinder are spacelike. The proper acceleration  $\kappa^m$  along a timelike Killing trajectory has a linear component  $\kappa^x$  along the  $x$ -direction, and an inward-pointing centrifugal component  $\kappa^\perp$  perpendicular to  $x$ -direction,

$$\{\kappa^x, \kappa^\perp\} \equiv \left\{ x_0 \left( \frac{d\beta}{d\tau} \right)^2, -r_0 \left( \frac{d\alpha}{d\tau} \right)^2 \right\} = \left\{ \frac{x_0}{x_0^2 - \mu^2 r_0^2}, -\frac{\mu^2 r_0}{x_0^2 - \mu^2 r_0^2} \right\}. \quad (14.85)$$

The ratio  $\kappa^\perp/\kappa^x$  of centrifugal to linear proper accelerations is

$$\frac{\kappa^\perp}{\kappa^x} = -\frac{\mu^2 r_0}{x_0}. \quad (14.86)$$

Figure 14.2 illustrates a sample of Killing trajectories for the case  $\mu = 1$ .

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,  $\bar{\boldsymbol{\theta}}\boldsymbol{\theta} = 0$ . 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,  $|\boldsymbol{\theta}_R| = |\boldsymbol{\theta}_I|$ , equation (14.76). A null bivector is also nilpotent,  $\boldsymbol{\theta}^2 = 0$ , so the rotor  $R$  obtained by exponentiating  $\boldsymbol{\theta}$  is linear in  $\boldsymbol{\theta}$ ,

$$R \equiv e^{-\lambda\boldsymbol{\theta}/2} = 1 - \lambda\boldsymbol{\theta}/2. \quad (14.87)$$

A Killing trajectory  $\mathbf{x}$  starting from an initial 4-vector  $\mathbf{x}_0$  is

$$\mathbf{x} \equiv R\mathbf{x}_0\bar{R} = \mathbf{x}_0 - \frac{\lambda}{2} [\boldsymbol{\theta}, \mathbf{x}_0], \quad (14.88)$$

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 if this is a useful result.

**Exercise 14.9 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.** As long as the electromagnetic field  $\mathbf{F} = \mathbf{B} - I\mathbf{E}$ , equation (14.61), 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 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  $\mathbf{n}$ .

## 14.7 Dirac matrices

The multiplication rules (14.1) for the basis vectors  $\gamma_m$  of the spacetime algebra are identical to the rules (14.5) governing the Clifford algebra of the **Dirac  $\gamma$ -matrices** used in the Dirac theory of relativistic spin- $\frac{1}{2}$  particles.

The Dirac  $\gamma$ -matrices are conventionally represented by  $4 \times 4$  complex (with respect to the quantum-mechanical imaginary  $i$ ) 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.

The requirement that the Dirac number current density be positive imposes the condition, equation (36.107), that taking the Hermitian conjugate of any of the basis vectors  $\gamma_m$  be equivalent to raising its index,

$$\gamma_m^\dagger = \gamma^m. \quad (14.89)$$

Condition (14.89) is the same as requiring that the basis vectors be unitary matrices,  $\gamma_m^{-1} = \gamma_m^\dagger$ .

The precise choice of matrices for  $\gamma_m$  is not fundamental: any set of 4 complex  $4 \times 4$  matrices satisfying conditions (14.5) and (14.89) will do. The  $\gamma$ -matrices are necessarily traceless. The traditional choice is the **Dirac representation**. The high-energy physics community commonly adopts the  $+---$  metric signature, which is opposite to the convention adopted here. With the high-energy  $+---$  signature, the standard Dirac representation of the Dirac  $\gamma$ -matrices is

$$\gamma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma_a = \begin{pmatrix} 0 & -\sigma_a \\ \sigma_a & 0 \end{pmatrix}, \quad (14.90)$$

where 1 denotes the unit  $2 \times 2$  matrix, and  $\sigma_a$  denote the three  $2 \times 2$  Pauli matrices (13.97). The choice of  $\gamma_0$  as a diagonal matrix is motivated by Dirac's discovery that eigenvectors of the time basis vector  $\gamma_0$  with eigenvalues of opposite sign define particles and antiparticles in their rest frames (see §14.8). To convert to the  $-+++$  metric signature adopted here while retaining the conventional set of eigenvectors, an additional factor of  $i$  must be inserted into the  $\gamma$ -matrices:

$$\gamma_0 = i \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma_a = i \begin{pmatrix} 0 & -\sigma_a \\ \sigma_a & 0 \end{pmatrix}. \quad (14.91)$$

In the representation (14.91), the bivectors  $\sigma_a$  and  $I\sigma_a$  and the pseudoscalar  $I$  of the spacetime algebra are

$$\gamma_0\gamma_a = \sigma_a = \begin{pmatrix} 0 & \sigma_a \\ \sigma_a & 0 \end{pmatrix}, \quad \frac{1}{2}\varepsilon_{abc}\gamma_b\gamma_c = I\sigma_a = i \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_a \end{pmatrix}, \quad I = i \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (14.92)$$

The Hermitian conjugates of the bivector and pseudoscalar basis elements are

$$\sigma_a^\dagger = \sigma_a, \quad (I\sigma_a)^\dagger = -I\sigma_a, \quad I^\dagger = -I. \quad (14.93)$$

The conventional chiral matrix  $\gamma_5$  of Dirac theory is defined to be  $-i$  times the pseudoscalar,

$$\gamma_5 \equiv -i\gamma_0\gamma_1\gamma_2\gamma_3 = -iI = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad (14.94)$$

whose representation is the same for either signature  $-+++$  or  $+---$ . The chiral matrix  $\gamma_5$  is Hermitian ( $\gamma_5^\dagger = \gamma_5$ ) and unitary ( $\gamma_5^{-1} = \gamma_5$ ), so its square is the unit matrix,

$$\gamma_5^\dagger = \gamma_5, \quad \gamma_5^2 = 1. \quad (14.95)$$

## 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  $\psi$  is a complex (with respect to the quantum-mechanical imaginary  $i$ ) linear combination of 4 basis spinors  $\gamma_a$  with indices  $a$  running over  $\{\uparrow\uparrow, \uparrow\downarrow, \downarrow\uparrow, \downarrow\downarrow\}$ , a total of 8 degrees of freedom in all,

$$\psi = \psi^a \gamma_a. \quad (14.96)$$

The basis spinors  $\gamma_a$  are basis elements of a super spacetime algebra, Chapter 36. In the Dirac representation (14.91), the four basis spinors are the column spinors

$$\gamma_{\uparrow\uparrow} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_{\uparrow\downarrow} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_{\downarrow\uparrow} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \gamma_{\downarrow\downarrow} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \quad (14.97)$$

The Dirac  $\gamma$ -matrices operate by pre-multiplication on Dirac spinors  $\psi$ , yielding other Dirac spinors. The basis spinors are eigenvectors of the time basis vector  $\gamma_0$  and of the bivector  $I\sigma_3$ , with  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  denoting eigenvectors of  $\gamma_0$ , and  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  eigenvectors of  $I\sigma_3$ ,

$$\gamma_0 \gamma_{\uparrow} = i \gamma_{\uparrow}, \quad \gamma_0 \gamma_{\downarrow} = -i \gamma_{\downarrow}, \quad I\sigma_3 \gamma_{\uparrow} = i \gamma_{\uparrow}, \quad I\sigma_3 \gamma_{\downarrow} = -i \gamma_{\downarrow}. \quad (14.98)$$

The bivector  $I\sigma_3$  is the generator of a spatial rotation about the 3-axis ( $z$ -axis), equation (14.30). Simultaneous eigenvectors of  $\gamma_0$  and  $I\sigma_3$  exist because  $\gamma_0$  and  $I\sigma_3$  commute.

A pure spin-up Dirac spinor  $\gamma_{\uparrow}$  can be rotated into a pure spin-down spinor  $\gamma_{\downarrow}$ , or vice versa, by a spatial rotation about the 1-axis or 2-axis. By contrast, a pure time-up spinor  $\gamma_{\uparrow}$  cannot be rotated into a pure time-down spinor  $\gamma_{\downarrow}$ , or vice versa, by any Lorentz transformation. Consider for example trying to rotate the pure time-up spin-up  $\gamma_{\uparrow\uparrow}$  spinor into any combination of pure time-down  $\gamma_{\downarrow}$  spinors. According to the expression (14.109), the Dirac spinor  $\psi$  obtained by Lorentz transforming the  $\gamma_{\uparrow\uparrow}$  spinor is pure  $\gamma_{\downarrow}$  only if the corresponding complex quaternion  $q$  is pure imaginary (with respect to  $I$ ). But a pure imaginary quaternion has negative squared magnitude  $\bar{q}q$ , so cannot be equivalent to any rotor of unit magnitude.

Thus the pure time-up and pure time-down spinors  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  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  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  do mix under Lorentz transformations. The manner in which Dirac spinors transform is described in §14.9.

The choice of time-axis  $\gamma_0$  and spin-axis  $\gamma_3$  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  $\gamma_0$  correspond to particles and antiparticles that are at rest in that frame. The eigenvectors associated with a particular spin-axis  $\gamma_3$  correspond to particles or antiparticles that are 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 (14.109). In the relativistic theory, the corresponding spin- $\frac{1}{2}$  object, a Dirac spinor  $\psi$ , is isomorphic (14.102) to a complex quaternion. The 4 complex degrees of freedom of the Dirac spinor  $\psi$  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, with zero magnitude, 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 (14.116), but if the Dirac spinor is null, then it is not simply a scaled rotor. The present section establishes an isomorphism (14.102) 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  $\gamma$ -matrices, then under rotation by Lorentz rotor  $R$ , the multivector  $\mathbf{a}$  operating on the Dirac spinor  $\psi$  transforms as

$$R : \mathbf{a}\psi \rightarrow (Ra\bar{R})(R\psi) = Ra\psi . \quad (14.99)$$

This shows that a Dirac spinor  $\psi$  Lorentz transforms, by construction, as

$$R : \psi \rightarrow R\psi . \quad (14.100)$$

The rule (14.100) is precisely the transformation rule (13.58) 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 (14.100) holds for any linear combination of Dirac spinors  $\psi$ . The isomorphism (14.32) between spacetime rotors and unit quaternions, coupled with linearity, shows that the algebra of Dirac spinors is isomorphic to the algebra of complex quaternions. Specifically, any Dirac spinor  $\psi$  can be expressed uniquely in the form of a  $4 \times 4$  matrix  $\mathbf{q}$ , the Dirac representation of a complex quaternion  $q$ , acting on the time-up spin-up eigenvector  $\gamma_{\uparrow\uparrow}$  (the precise translation between Dirac spinors and complex quaternions is left as Exercises 14.10 and 14.11):

$$\psi = \mathbf{q} \gamma_{\uparrow\uparrow} . \quad (14.101)$$

In this section (including the Exercises) the  $4 \times 4$  matrix  $\mathbf{q}$  is written in boldface to distinguish it from the quaternion  $q$  that it represents; but the distinction is not fundamental. The equivalence (14.101) establishes that Dirac spinors are isomorphic to complex quaternions

$$\psi \leftrightarrow q . \quad (14.102)$$

The isomorphism means that there is a one-to-one correspondence between Dirac spinors  $\psi$  and complex quaternions  $q$ , and that they transform in the same way under Lorentz transformations. The isomorphism does not mean that Dirac spinors and complex quaternions are identical. As will be discovered in Chapter 36, Dirac (and Majorana) spinors form a super spacetime algebra that is distinct from the algebra of quaternions.

**Exercise 14.10 Translate a Dirac spinor into a complex quaternion.** Given any Dirac spinor in the Dirac representation (14.91),

$$\psi = \psi^a \gamma_a = \begin{pmatrix} \psi^{\uparrow\uparrow} \\ \psi^{\uparrow\downarrow} \\ \psi^{\downarrow\uparrow} \\ \psi^{\downarrow\downarrow} \end{pmatrix} , \quad (14.103)$$

show that the corresponding complex quaternion  $q$ , and the equivalent  $4 \times 4$  matrix  $\mathbf{q}$  such that  $\psi = \mathbf{q} \gamma_{\uparrow\uparrow}$ , are (the complex conjugates  $\psi^{a*}$  of the components  $\psi^a$  of the spinor are with respect to the quantum-mechanical imaginary  $i$ )

$$q = \begin{Bmatrix} \operatorname{Re} \psi^{\uparrow\uparrow} & \operatorname{Im} \psi^{\uparrow\downarrow} & -\operatorname{Re} \psi^{\uparrow\downarrow} & \operatorname{Im} \psi^{\uparrow\uparrow} \\ -\operatorname{Im} \psi^{\uparrow\uparrow} & \operatorname{Re} \psi^{\downarrow\downarrow} & \operatorname{Im} \psi^{\downarrow\downarrow} & \operatorname{Re} \psi^{\downarrow\uparrow} \end{Bmatrix} \leftrightarrow \mathbf{q} = \begin{pmatrix} \psi^{\uparrow\uparrow} & -\psi^{\uparrow\downarrow*} & \psi^{\downarrow\uparrow} & \psi^{\downarrow\downarrow*} \\ \psi^{\uparrow\downarrow} & \psi^{\uparrow\uparrow*} & \psi^{\downarrow\downarrow} & -\psi^{\uparrow\downarrow*} \\ \psi^{\downarrow\uparrow} & \psi^{\downarrow\downarrow*} & \psi^{\uparrow\uparrow} & -\psi^{\uparrow\downarrow*} \\ \psi^{\downarrow\downarrow} & -\psi^{\downarrow\uparrow*} & \psi^{\uparrow\downarrow} & \psi^{\uparrow\uparrow*} \end{pmatrix}. \quad (14.104)$$

Show that the reverse complex quaternion  $\bar{q}$  and the equivalent  $4 \times 4$  matrix  $\bar{\mathbf{q}}$  in the Dirac representation (14.91), are

$$\bar{q} = \begin{Bmatrix} \operatorname{Re} \psi^{\uparrow\uparrow} & -\operatorname{Im} \psi^{\uparrow\downarrow} & \operatorname{Re} \psi^{\uparrow\downarrow} & -\operatorname{Im} \psi^{\uparrow\uparrow} \\ -\operatorname{Im} \psi^{\uparrow\uparrow} & -\operatorname{Re} \psi^{\downarrow\downarrow} & -\operatorname{Im} \psi^{\downarrow\downarrow} & -\operatorname{Re} \psi^{\uparrow\downarrow} \end{Bmatrix} \leftrightarrow \bar{\mathbf{q}} = \begin{pmatrix} \psi^{\uparrow\uparrow*} & \psi^{\uparrow\downarrow*} & -\psi^{\downarrow\uparrow*} & -\psi^{\downarrow\downarrow*} \\ -\psi^{\uparrow\downarrow} & \psi^{\uparrow\uparrow} & -\psi^{\downarrow\downarrow} & \psi^{\downarrow\uparrow} \\ -\psi^{\downarrow\uparrow*} & -\psi^{\downarrow\downarrow*} & \psi^{\uparrow\uparrow*} & \psi^{\uparrow\downarrow*} \\ -\psi^{\downarrow\downarrow} & \psi^{\downarrow\uparrow} & -\psi^{\uparrow\downarrow} & \psi^{\uparrow\uparrow} \end{pmatrix}. \quad (14.105)$$

Conclude that the reverse spinor defined by  $\bar{\psi} \equiv \gamma_{\uparrow\uparrow}^\top \bar{\mathbf{q}}$  is

$$\bar{\psi} \equiv (\gamma_{\uparrow\uparrow})^\dagger \bar{\mathbf{q}} = (\psi^{\uparrow\uparrow*} \ \psi^{\uparrow\downarrow*} \ -\psi^{\downarrow\uparrow*} \ -\psi^{\downarrow\downarrow*}). \quad (14.106)$$

**Exercise 14.11 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 (14.91) to the  $4 \times 4$  matrix  $\mathbf{q}$

$$q = \begin{Bmatrix} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{Bmatrix} \leftrightarrow \mathbf{q} = \begin{pmatrix} w_R + iz_R & ix_R + y_R & -iw_I + z_I & x_I - iy_I \\ ix_R - y_R & w_R - iz_R & x_I + iy_I & -iw_I - z_I \\ -iw_I + z_I & x_I - iy_I & w_R + iz_R & ix_R + y_R \\ x_I + iy_I & -iw_I - z_I & ix_R - y_R & w_R - iz_R \end{pmatrix}. \quad (14.107)$$

Show that the reverse quaternion  $\bar{q}$ , the complex conjugate (with respect to  $I$ ) quaternion  $q^*$ , and the reverse complex conjugate (with respect to  $I$ ) quaternion  $\bar{q}^*$  are respectively equivalent to the  $4 \times 4$  matrices

$$\bar{q} \leftrightarrow \bar{\mathbf{q}} \equiv -\boldsymbol{\gamma}_0 \mathbf{q}^\dagger \boldsymbol{\gamma}_0, \quad (14.108a)$$

$$q^* \leftrightarrow \bar{\mathbf{q}}^\dagger = -\boldsymbol{\gamma}_0 \mathbf{q} \boldsymbol{\gamma}_0, \quad (14.108b)$$

$$\bar{q}^* \leftrightarrow \mathbf{q}^\dagger = -\boldsymbol{\gamma}_0 \bar{\mathbf{q}} \boldsymbol{\gamma}_0. \quad (14.108c)$$

Conclude that the Dirac spinor  $\psi \equiv \mathbf{q} \gamma_{\uparrow\uparrow}$  corresponding to the complex quaternion  $q$  is

$$\psi \equiv \mathbf{q} \gamma_{\uparrow\uparrow} = \begin{pmatrix} w_R + iz_R \\ ix_R - y_R \\ -iw_I + z_I \\ x_I + iy_I \end{pmatrix}, \quad (14.109)$$

that the reverse spinor  $\bar{\psi} \equiv (\gamma_{\uparrow\uparrow})^\dagger \bar{\mathbf{q}}$  is

$$\bar{\psi} \equiv (\gamma_{\uparrow\uparrow})^\dagger \bar{\mathbf{q}} = (w_R - iz_R \ -ix_R - y_R \ -iw_I - z_I \ -x_I + iy_I), \quad (14.110)$$

and that the Hermitian conjugate spinor  $\psi^\dagger \equiv (\gamma_{\uparrow\uparrow})^\dagger \mathbf{q}^\dagger$  is

$$\psi^\dagger \equiv (\gamma_{\uparrow\uparrow})^\dagger \mathbf{q}^\dagger = \begin{pmatrix} w_R - iz_R & -ix_R - y_R & iw_I + z_I & x_I - iy_I \end{pmatrix}. \quad (14.111)$$

Hence conclude that  $\bar{\psi}\psi$  and  $\psi^\dagger\psi$  are respectively the real part of, and the absolute value of, the complex magnitude squared  $\bar{q}q \equiv \lambda^2$  of the complex quaternion  $q$ ,

$$\bar{\psi}\psi = \lambda_R^2 - \lambda_I^2, \quad (14.112a)$$

$$\psi^\dagger\psi = \lambda_R^2 + \lambda_I^2, \quad (14.112b)$$

with

$$\lambda_R^2 = w_R^2 + x_R^2 + y_R^2 + z_R^2, \quad (14.113a)$$

$$\lambda_I^2 = w_I^2 + x_I^2 + y_I^2 + z_I^2. \quad (14.113b)$$

**Exercise 14.12 Relation between  $\bar{\psi}$  and  $\psi^\dagger$ .** Show that  $\bar{\psi}$  and  $\psi^\dagger$  are related by

$$\bar{\psi} = -i\psi^\dagger \boldsymbol{\gamma}_0, \quad \psi^\dagger = -i\bar{\psi} \boldsymbol{\gamma}_0, \quad (14.114)$$

by showing from equations (14.110) and (14.111) that

$$\psi^\dagger = -i(\gamma_{\uparrow\uparrow})^\dagger \bar{\mathbf{q}} \boldsymbol{\gamma}_0. \quad (14.115)$$

The same result follows from equation (14.108b). The Hermitian conjugate matrix is  $\mathbf{q}^\dagger = -\boldsymbol{\gamma}_0 \bar{\mathbf{q}} \boldsymbol{\gamma}_0$ , and  $(\gamma_{\uparrow\uparrow})^\dagger \boldsymbol{\gamma}_0 = i(\gamma_{\uparrow\uparrow})^\dagger$ , so  $\psi^\dagger = (\gamma_{\uparrow\uparrow})^\dagger \mathbf{q}^\dagger = -i(\gamma_{\uparrow\uparrow})^\dagger \bar{\mathbf{q}} \boldsymbol{\gamma}_0$ .

---

## 14.10 Non-null Dirac spinor

A non-null, or massive, Dirac spinor  $\psi$ , one for which  $\bar{\psi}\psi \neq 0$ , is isomorphic (14.102) to a non-null complex quaternion  $q$ , which can be factored as a non-zero complex scalar times a unit quaternion, a rotor. Thus a non-null Dirac spinor can be expressed as the product of a complex scalar  $\lambda = \lambda_R + I\lambda_I$  and a Lorentz rotor  $R$ , acting on the rest-frame eigenvector  $\gamma_{\uparrow\uparrow}$ ,

$$\psi = q \gamma_{\uparrow\uparrow}, \quad q = \lambda R. \quad (14.116)$$

The complex scalar  $\lambda$  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  $2\pi$  of the rotor  $R$ . There is no further ambiguity in the decomposition (14.116) into scalar and rotor, because the squared magnitude  $\bar{\lambda}R\lambda R = \lambda^2$  of the scaled rotor  $\lambda R$  is the same for any decomposition.

The fact that a non-null Dirac spinor  $\psi$  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  $\gamma_{\uparrow}$  represents a particle in its own rest frame, while a pure time-down eigenvector  $\gamma_{\downarrow}$  represents an antiparticle in its own rest frame. The time-up spin-up eigenvector

$\gamma_{\uparrow\uparrow}$  is by definition (14.116) equivalent to the unit scaled rotor,  $\lambda R = 1$ , so in this case the scalar  $\lambda$  is pure real. Lorentz transforming the eigenvector multiplies it by a rotor, but leaves the scalar  $\lambda$  unchanged, therefore pure real. Conversely, if the time-up spin-up eigenvector  $\gamma_{\uparrow\uparrow}$  is multiplied by the imaginary  $I$ , then according to the expression (14.109) the resulting spinor can be Lorentz transformed into a pure  $\gamma_{\downarrow\downarrow}$  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 product  $\psi^\dagger\psi$  of the Dirac spinor with its Hermitian conjugate, equation (14.112b),

$$\psi^\dagger\psi = |\lambda|^2 = \lambda_R^2 + \lambda_I^2 , \quad (14.117)$$

is the sum of the probabilities  $\lambda_R^2$  of particles and  $\lambda_I^2$  of antiparticles. The magnitude  $\bar{\psi}\psi$  of the Dirac spinor, equation (14.112a),

$$\bar{\psi}\psi = \lambda_R^2 - \lambda_I^2 , \quad (14.118)$$

is the difference between the probabilities  $\lambda_R^2$  of particles and  $\lambda_I^2$  of antiparticles.

## 14.11 Null Dirac Spinor

A **null Dirac spinor** is a Dirac spinor  $\psi$  whose magnitude is zero,

$$\bar{\psi}\psi = 0 . \quad (14.119)$$

Such a spinor is equivalent to a null complex quaternion  $q$  acting on the rest-frame eigenvector  $\gamma_{\uparrow\uparrow}$ ,

$$\psi = q\gamma_{\uparrow\uparrow} , \quad \bar{q}q = 0 . \quad (14.120)$$

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  $\psi = \lambda R\gamma_{\uparrow\uparrow}$  would be unity,  $R = 1$ , and the spinor, being non-trivial,  $\lambda \neq 0$ , would not be null. The null condition (14.119) 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.

Any non-trivial null complex quaternion  $q$  can be written uniquely as the product of a real quaternion  $\lambda U$  and a null factor  $(1 - In)$  (Exercise 14.1):

$$q = \lambda U(1 - In) . \quad (14.121)$$

Here  $\lambda$  is a positive real scalar,  $U$  is a purely spatial (i.e. real, with no  $I$  part) rotor, and  $n = \iota_a n_a$ ,  $a = 1, 2, 3$ , is a real unit vector quaternion, satisfying  $n_a n_a = 1$ . Physically, equation (14.121) contains the instruction to boost to light speed in the direction  $n$ , then scale by the real scalar  $\lambda$  and rotate spatially by  $U$ . The minus sign in front of  $In$  comes from the fact that a boost in direction  $n$  is described by a rotor  $R = \cosh(\theta/2) - In \sinh(\theta/2)$ , equation (14.40), which becomes proportional to  $1 - In$  as the boost tends to infinity,  $\theta \rightarrow \infty$ . The  $1 + 3 + 2 = 6$  degrees of freedom from the real scalar  $\lambda$ , the spatial rotor  $U$ , and the real unit vector  $n$  in the expression (14.121) 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 transformed (pre-multiplied) by any complex quaternion  $p + Ir$ , then the result

$$(p + Ir)(1 - I\mathbf{n}) = (p + r\mathbf{n})(1 - I\mathbf{n}) \quad (14.122)$$

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 (14.122) is true because  $\mathbf{n}^2 = -1$ . The null Dirac spinor  $\psi$  corresponding to the null complex quaternion  $q$ , equation (14.121), is

$$\psi \equiv q \gamma_{\uparrow\uparrow} = \lambda U(1 - I\mathbf{n}) \gamma_{\uparrow\uparrow}. \quad (14.123)$$

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 14.13 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 (14.122) shows that a Lorentz transformation of the null Dirac spinor  $\psi$  (14.123) is equivalent to a scaling and a spatial rotation of that spinor,

$$(p + Ir)\psi = (p + Ir)\lambda U(1 - I\mathbf{n}) \gamma_{\uparrow\uparrow} = (p + r\mathbf{n}')\psi, \quad (14.124)$$

where

$$\mathbf{n}' = U\mathbf{n}\bar{U}. \quad (14.125)$$

The real quaternion  $p + r\mathbf{n}'$  on the right hand side of equation (14.124) 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  $\theta$  along the boost axis  $\mathbf{n}$ , equation (14.40), multiplies the null spinor  $(1 - I\mathbf{n})\gamma_{\uparrow\uparrow}$  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 (14.123) is

$$\bar{\psi} \equiv (\gamma_{\uparrow\uparrow})^\dagger \bar{q} = (\gamma_{\uparrow\uparrow})^\dagger (1 + I\mathbf{n})\lambda \bar{U}. \quad (14.126)$$

The spinor is null,  $\bar{\psi}\psi = 0$ , 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 (14.123) 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 (14.91), then for a right-handed spinor, the boost direction is  $\mathbf{n} = \iota_3$ , while for a left-handed spinor, the boost direction is  $\mathbf{n} = -\iota_3$ .

If the quaternionic components of the spatial rotor in equation (14.123) are  $\lambda U = \{w, x, y, z\}$ , then the complex quaternionic components of the right- or left-handed Weyl spinor are

$$\psi_{\text{R}}^{\text{L}} = \left\{ \begin{array}{cccc} w & x & y & z \\ \mp z & \mp y & \pm x & \pm w \end{array} \right\}. \quad (14.127)$$

**Concept question 14.14 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.** Weyl spinors form a complete set of eigenfunctions with respect to the commuting transformations of spatial rotations about any particular axis, such as the 3-axis ( $z$ -axis), and Lorentz boosts along that axis.

# 15

---

## 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, §15.8. 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 (15.59). A closely related derivative is the covariant spacetime derivative  $D$  defined by

$$D \equiv e^\nu D_\nu = \gamma^n D_n , \quad (15.2)$$

where  $D_\nu$  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  $\dot{D} \wedge$  (see equation (15.75) for a more precise statement of the isomorphism),

$$d \leftrightarrow \dot{D} \wedge . \quad (15.3)$$

The first part of this chapter shows how to take the covariant derivative of a multivector, and defines the covariant spacetime derivative  $D$ . 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 (2.67). In sections 15.1–15.4, the covariant derivative  $D_n$  and the covariant spacetime derivative  $D$  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  $D$  is the torsion-free derivative, 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 (13.46) by rotor  $R$  transforms a multivector  $\mathbf{a}$  by  $\mathbf{a} \rightarrow R\mathbf{a}\bar{R}$ . 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 (13.30). In the second place, the anti-commutator selects all grade components of the geometric product that anticommute, per equation (13.26). 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  $\Gamma_{kmn}$ , 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  $\partial_n$  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

$$\frac{1}{2}[\boldsymbol{\Gamma}_n, \boldsymbol{\gamma}_m] = \Gamma_{mn}^k \boldsymbol{\gamma}_k . \quad (15.8)$$

The  $N$ -tuple (not vector) of bivectors  $\boldsymbol{\Gamma}_n$  satisfying equation (15.8) is

$$\boxed{\boldsymbol{\Gamma}_n = \frac{1}{2}\Gamma_{klm} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l} \quad (15.9)$$

(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  $\boldsymbol{\Gamma}_n$  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

$$D_n \mathbf{a} = \boldsymbol{\gamma}^m \partial_n a_m + \frac{1}{2}[\boldsymbol{\Gamma}_n, \boldsymbol{\gamma}^m] a_m , \quad (15.11)$$

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

$$\frac{1}{2}[\boldsymbol{\Gamma}_n, \boldsymbol{\gamma}^m] = -\Gamma_{kn}^m \boldsymbol{\gamma}^k . \quad (15.13)$$

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:

$$\hat{\Gamma}_n \mathbf{ab} = \frac{1}{2}[\boldsymbol{\Gamma}_n, \mathbf{ab}] = \frac{1}{2}[\boldsymbol{\Gamma}_n, \mathbf{a}] \mathbf{b} + \frac{1}{2} \mathbf{a} [\boldsymbol{\Gamma}_n, \mathbf{b}] = (\hat{\Gamma}_n \mathbf{a}) \mathbf{b} + \mathbf{a} (\hat{\Gamma}_n \mathbf{b}) . \quad (15.14)$$

To summarize, the covariant derivative of a multivector  $\mathbf{a}$  can be written

$$\boxed{D_n \mathbf{a} = \partial_n \mathbf{a} + \frac{1}{2}[\boldsymbol{\Gamma}_n, \mathbf{a}] ,} \quad (15.15)$$

with the  $N$ -tuple of bivectors  $\boldsymbol{\Gamma}_n$  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  $\partial_n$  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

$$D_n \mathbf{a}_m = \partial_n \mathbf{a}_m - \Gamma_{mn}^k \mathbf{a}_k + \frac{1}{2}[\boldsymbol{\Gamma}_n, \mathbf{a}_m] . \quad (15.16)$$

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

$$D_n(\boldsymbol{\gamma}_m a^m) = \boldsymbol{\gamma}_m D_n a^m , \quad (15.18)$$

which is true by construction.

The covariant derivative of a multivector  $\mathbf{a}$  can also be expressed as a coordinate derivative

$$D_{\nu} \mathbf{a} = \frac{\partial \mathbf{a}}{\partial x^{\nu}} + \frac{1}{2} [\boldsymbol{\Gamma}_{\nu}, \mathbf{a}] , \quad (15.19)$$

where the coordinate and directed derivatives are related as usual by  $\partial/\partial x^{\nu} = e^n_{\nu} \partial_n$ , and where the coordinate connection vector  $\boldsymbol{\Gamma}_{\nu}$  is related to the tetrad connection  $\boldsymbol{\Gamma}_n$  defined by equation (15.9) by

$$\boldsymbol{\Gamma}_{\nu} \equiv e^n_{\nu} \boldsymbol{\Gamma}_n . \quad (15.20)$$

The components  $\Gamma_{kl\nu} \equiv e^n_{\nu} \Gamma_{kln}$  of  $\boldsymbol{\Gamma}_{\nu} \equiv \frac{1}{2} \Gamma_{kl\nu} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l$  constitute a coordinate-frame vector, but not a tetrad-frame tensor. The connection  $\boldsymbol{\Gamma}_{\nu}$  is given by equation (15.20), *not* by a direct relation to the coordinate-frame connections  $\Gamma_{\mu\nu\kappa}$ , that is,  $\boldsymbol{\Gamma}_{\nu} \neq \frac{1}{2} \Gamma_{\kappa\lambda\nu} \mathbf{e}^{\kappa} \wedge \mathbf{e}^{\lambda}$ .

## 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 (2.111). Define the Riemann antisymmetric tensor of bivectors  $\mathbf{R}_{kl}$  by

$$\boxed{\mathbf{R}_{kl} = \frac{1}{2} R_{klmn} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n} \quad (15.22)$$

(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  $\boldsymbol{\Gamma}_k$ , equation (15.9), by

$$\mathbf{R}_{kl} = \partial_k \boldsymbol{\Gamma}_l - \partial_l \boldsymbol{\Gamma}_k + \frac{1}{2} [\boldsymbol{\Gamma}_k, \boldsymbol{\Gamma}_l] + (\Gamma_{kl}^m - \Gamma_{lk}^m - S_{kl}^m) \boldsymbol{\Gamma}_m , \quad (15.24)$$

where, in conformity with the convention of equation (15.15), directed derivatives  $\partial_k \boldsymbol{\Gamma}_l$  are to be interpreted as acting only on the components  $\Gamma_{ml}$  of  $\boldsymbol{\Gamma}_l \equiv \frac{1}{2} \Gamma_{mln} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n$ , not on the tetrad axes  $\boldsymbol{\gamma}_m$ . 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 §16.14.1. The components of  $\mathbf{R}_{\kappa\lambda} = \frac{1}{2} R_{\kappa\lambda mn} \gamma^m \wedge \gamma^n$  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))

$$\mathbf{S}_{\kappa\lambda} \equiv -S_{m\kappa\lambda} \gamma^m . \quad (15.27)$$

In components, the torsion tensor of vectors  $\mathbf{S}_{\kappa\lambda}$  is, from equation (11.49),

$$\mathbf{S}_{\kappa\lambda} = \left( \frac{\partial e^m{}_\lambda}{\partial x^\kappa} - \frac{\partial e^m{}_\kappa}{\partial x^\lambda} + \Gamma^m{}_{\kappa\lambda} - \Gamma^m{}_{\lambda\kappa} \right) \gamma_m , \quad (15.28)$$

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  $\mathbf{e}_\kappa \equiv e^k{}_\kappa \gamma_k$  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  $e^k{}_\kappa$  of  $\mathbf{e}_\kappa$ , not on the tetrad axes  $\gamma_k$ . Equation (15.29) is Cartan's first equation of structure, §16.14.1. 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

$$\boxed{\mathbf{D} \equiv \gamma^n D_n} . \quad (15.31)$$

The covariant spacetime derivative  $\mathbf{D}$  is a sum of a directed derivative  $\partial$  and a connection operator  $\hat{\Gamma}$ ,

$$\mathbf{D} = \partial + \hat{\Gamma}, \quad \partial \equiv \boldsymbol{\gamma}^n \partial_n, \quad \hat{\Gamma} \equiv \boldsymbol{\gamma}^n \hat{\Gamma}_n. \quad (15.32)$$

The action of the connection operator  $\hat{\Gamma}$  on a multivector  $\mathbf{a}$  is

$$\hat{\Gamma}\mathbf{a} = \frac{1}{2} \boldsymbol{\gamma}^n [\Gamma_n, \mathbf{a}] \quad (15.33)$$

(not  $\hat{\Gamma}\mathbf{a} = \frac{1}{2}[\Gamma, \mathbf{a}]$ ). The covariant spacetime derivative of a multivector  $\mathbf{a}$  is

$$\mathbf{D}\mathbf{a} = \boldsymbol{\gamma}^n D_n \mathbf{a} = \boldsymbol{\gamma}^n (\partial_n \mathbf{a} + \frac{1}{2} [\Gamma_n, \mathbf{a}]). \quad (15.34)$$

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  $\mathbf{a}$ ,

$$\mathbf{D}\mathbf{a} = \mathbf{D} \cdot \mathbf{a} + \mathbf{D} \wedge \mathbf{a} \quad \text{multivector}. \quad (15.37)$$

The dot and wedge notation for the covariant divergence and curl conforms to the convention for the dot and wedge product of two multivectors, §13.5. If torsion vanishes, the curl  $\mathbf{D} \wedge \mathbf{a}$  is essentially the same as the exterior derivative in the mathematics of differential forms, §15.8.

The covariant spacetime divergence and curl of a grade- $p$  multivector  $\mathbf{a} = (1/p!) \boldsymbol{\gamma}^{lm\dots n} a_{lm\dots n}$  are

$$\mathbf{D} \cdot \mathbf{a} = \frac{1}{(p-1)!} \boldsymbol{\gamma}^{m\dots n} (\mathbf{D} \cdot \mathbf{a})_{m\dots n}, \quad (\mathbf{D} \cdot \mathbf{a})_{m\dots n} = D^l a_{lm\dots n}, \quad (15.38a)$$

$$\mathbf{D} \wedge \mathbf{a} = \frac{1}{(p+1)!} \boldsymbol{\gamma}^{klm\dots n} (\mathbf{D} \wedge \mathbf{a})_{klm\dots n}, \quad (\mathbf{D} \wedge \mathbf{a})_{klm\dots n} = (p+1) D_{[k} a_{lm\dots n]}. \quad (15.38b)$$

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  $\varphi$ , a vector  $\mathbf{A} = \boldsymbol{\gamma}^n A_n$ , and a bivector  $\mathbf{F} = \frac{1}{2} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n F_{mn}$ , are respectively

$$\mathbf{D} \cdot \varphi = 0, \quad (\mathbf{D} \wedge \varphi)_n = D_n \varphi = \partial_n \varphi, \quad (15.39a)$$

$$\mathbf{D} \cdot \mathbf{A} = D^n A_n, \quad (\mathbf{D} \wedge \mathbf{A})_{mn} = D_m A_n - D_n A_m, \quad (15.39b)$$

$$(\mathbf{D} \cdot \mathbf{F})_n = D^m F_{mn}, \quad (\mathbf{D} \wedge \mathbf{F})_{lmn} = D_l F_{mn} + D_m F_{nl} + D_n F_{lm}. \quad (15.39c)$$

Notice that the spacetime derivative of a scalar contains only a curl part: the spacetime divergence of a scalar is zero in accordance with the convention (13.34).

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

$$\mathbf{DD} = \mathbf{D} \cdot \mathbf{D} + \mathbf{D} \wedge \mathbf{D} = D_k D^k + \frac{1}{2} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l [D_k, D_l], \quad (15.42)$$

which is a sum of the scalar d'Alembertian wave operator  $\square \equiv D_n D^n$ , 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  $\mathbf{a} = a^n \boldsymbol{\gamma}_n$ ,

$$\mathbf{D} \wedge \mathbf{D} \wedge \mathbf{a} = \frac{1}{2} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l \wedge \boldsymbol{\gamma}^m R_{klmn} a^n = 0, \quad (15.43)$$

which vanishes thanks to the cyclic symmetry of the Riemann tensor,  $R_{[klm]n} = 0$ , valid for vanishing torsion.

## 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  $\boldsymbol{\Gamma}_n$  and  $\mathring{\boldsymbol{\Gamma}}_n$  are related by

$$\boldsymbol{\Gamma}_n = \mathring{\boldsymbol{\Gamma}}_n + \mathbf{K}_n, \quad (15.44)$$

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

$$\mathbf{K}_n = \frac{1}{2} K_{kln} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l. \quad (15.45)$$

Acting on a multivector  $\mathbf{a}$ , the torsion-full and torsion-free covariant spacetime derivatives  $\mathbf{D}$  and  $\mathring{\mathbf{D}}$  are related by

$$\mathbf{D}\mathbf{a} = \mathring{\mathbf{D}}\mathbf{a} + \frac{1}{2} \boldsymbol{\gamma}^n [\mathbf{K}_n, \mathbf{a}]. \quad (15.46)$$

From equation (15.25), the Riemann tensor of bivectors  $\mathbf{R}_{\kappa\lambda}$  splits into torsion-free and contortion parts,

$$\begin{aligned} \mathbf{R}_{\kappa\lambda} &= \frac{\partial(\mathring{\boldsymbol{\Gamma}}_\lambda + \mathbf{K}_\lambda)}{\partial x^\kappa} - \frac{\partial(\mathring{\boldsymbol{\Gamma}}_\kappa + \mathbf{K}_\kappa)}{\partial x^\lambda} + \frac{1}{2} [\mathring{\boldsymbol{\Gamma}}_\kappa + \mathbf{K}_\kappa, \mathring{\boldsymbol{\Gamma}}_\lambda + \mathbf{K}_\lambda] \\ &= \mathring{\mathbf{R}}_{\kappa\lambda} + \mathring{D}_\kappa \mathbf{K}_\lambda - \mathring{D}_\lambda \mathbf{K}_\kappa + \frac{1}{2} [\mathbf{K}_\kappa, \mathbf{K}_\lambda]. \end{aligned} \quad (15.47)$$

In the remainder of this chapter,  $\mathbf{D}$  denotes the torsion-free covariant spacetime derivative, even when torsion is present, since it is the torsion-free covariant derivative to which integration is inverse.

## 15.6 Differential forms

Differential forms, or  $p$ -forms, 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  $x^\mu$ , a 1-form expressed in the coordinate frame is

$$\mathbf{a} = a_\mu dx^\mu . \quad (15.48)$$

By definition, the differential  $dx^\mu$  transforms under coordinate transformations like a contravariant coordinate vector. Requiring that the 1-form  $\mathbf{a}$  defined by equation (15.48) be a coordinate scalar imposes that  $a_\mu$  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

$$\mathbf{a} = \frac{1}{2} a_{\mu\nu} dx^\mu \wedge dx^\nu , \quad (15.49)$$

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 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,

$$dx^\mu \wedge dx^\nu = -dx^\nu \wedge dx^\mu . \quad (15.50)$$

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 (15.49) 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

$$dx' \wedge dy' = \left( \frac{\partial x'}{\partial x} dx + \frac{\partial x'}{\partial y} dy \right) \wedge \left( \frac{\partial y'}{\partial x} dx + \frac{\partial y'}{\partial y} dy \right) = \left( \frac{\partial x'}{\partial x} \frac{\partial y'}{\partial y} - \frac{\partial x'}{\partial y} \frac{\partial y'}{\partial x} \right) dx \wedge dy . \quad (15.51)$$

The factor relating the two areal elements is the familiar Jacobian determinant  $|\partial\{x', y'\}/\partial\{x, y\}|$ . The definition (15.49) 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

$$\mathbf{a} = \frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} . \quad (15.52)$$

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 (15.52) could also be written

$$\mathbf{a} = a_A dx^A , \quad (15.53)$$

where the sum is only over distinct antisymmetric sequences  $A$  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 (15.52) 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 (15.52) of a  $p$ -form extends to the case  $p = 0$ . A 0-form is simply a scalar.

## 15.7 Wedge product of differential forms

The wedge product of differential forms is defined consistent with the wedge product of multivectors, equation (13.29). The wedge product of a 1-form  $\mathbf{a}$  with a 1-form  $\mathbf{b}$  defines a 2-form

$$\mathbf{a} \wedge \mathbf{b} = a_\mu dx^\mu \wedge b_\nu dx^\nu = a_{[\mu} b_{\nu]} dx^\mu \wedge dx^\nu = \frac{1}{2} (a_\mu b_\nu - a_\nu b_\mu) dx^\mu \wedge dx^\nu . \quad (15.54)$$

The factor of  $\frac{1}{2}$  in the last expression of equations (15.54) could be omitted if the implicit sum were taken over distinct antisymmetric pairs  $\mu\nu$  of indices. In general, the wedge product of a  $p$ -form  $\mathbf{a}$  with a  $q$ -form  $\mathbf{b}$  defines a  $(p+q)$ -form  $\mathbf{a} \wedge \mathbf{b}$ ,

$$\begin{aligned} \mathbf{a} \wedge \mathbf{b} &= \left( \frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \right) \wedge \left( \frac{1}{q!} b_{\nu_1 \dots \nu_q} dx^{\nu_1} \wedge \dots \wedge dx^{\nu_q} \right) \\ &= \frac{1}{p!q!} a_{[\mu_1 \dots \mu_p} b_{\nu_1 \dots \nu_q]} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \wedge dx^{\nu_1} \wedge \dots \wedge dx^{\nu_q} . \end{aligned} \quad (15.55)$$

If the forms are expressed as sums  $\mathbf{a} \equiv a_A d^p x^A$  and  $\mathbf{b} \equiv b_B d^q x^B$  over distinct antisymmetric sequences  $A$  and  $B$  of respectively  $p$  and  $q$  indices, then their wedge product is

$$\mathbf{a} \wedge \mathbf{b} = a_A b_B d^{p+q} x^{AB} = \frac{(p+q)!}{p!q!} a_{[A} b_{B]} d^{p+q} x^{AB} , \quad (15.56)$$

where the second expression is implicitly summed over distinct antisymmetric sequences  $A$  and  $B$  of  $p$  and  $q$  indices, while the last expression is implicitly summed over distinct antisymmetric sequences  $AB$  of  $p+q$  indices.

The wedge product is symmetric or antisymmetric as  $pq$  is even or odd,

$$\mathbf{a} \wedge \mathbf{b} = (-)^{pq} \mathbf{b} \wedge \mathbf{a}, \quad (15.57)$$

consistent with the wedge product (13.29) of two multivectors.

The wedge product of a 0-form (scalar)  $a$  with a differential form  $\mathbf{b}$  is just their ordinary product,

$$a \wedge \mathbf{b} = ab \quad \text{if } a \text{ is a scalar}, \quad (15.58)$$

consistent with the result (13.32) for multivectors.

## 15.8 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

$$d\mathbf{a} \equiv d \left( \frac{1}{p!} a_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \right) \equiv \frac{1}{p!} \frac{\partial a_{\mu_1 \dots \mu_p}}{\partial x^\nu} dx^\nu \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}. \quad (15.59)$$

Equation (15.59) 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 (15.59) may be rewritten

$$d\mathbf{a} = \frac{1}{p!} \partial_{[\nu} a_{\mu_1 \dots \mu_p]} dx^\nu \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}, \quad (15.60)$$

where  $\partial_\nu \equiv \partial/\partial x^\nu$ . But the antisymmetrized coordinate derivative is just equal to the antisymmetrized torsion-free covariant derivative (Exercise 2.5),

$$\partial_{[\nu} a_{\mu_1 \dots \mu_p]} = D_{[\nu} a_{\mu_1 \dots \mu_p]}, \quad (15.61)$$

which is true even when torsion is present (that is, the antisymmetrized coordinate derivative equals the anti-symmetrized 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. If the  $p$ -form is expressed as a sum  $\mathbf{a} \equiv a_{\textcolor{brown}{A}} d^p x^{\textcolor{brown}{A}}$  over distinct antisymmetric sequences  $\textcolor{brown}{A}$  of  $p$  indices, then its exterior derivative is the  $(p+1)$ -form

$$d\mathbf{a} = \partial_{\kappa} a_{\textcolor{brown}{A}} d^{p+1} x^{\kappa \textcolor{brown}{A}} = (p+1) \partial_{[\kappa} a_{\textcolor{brown}{A}]} d^{p+1} x^{\kappa \textcolor{brown}{A}}, \quad (15.62)$$

where the second expression is implicitly summed over distinct antisymmetric sequences  $\textcolor{brown}{A}$  of  $p$  indices, while the last expression is implicitly summed over distinct antisymmetric sequences  $\kappa \textcolor{brown}{A}$  of  $p+1$  indices,

The simplest case is the exterior derivative of a 0-form (scalar)  $\varphi$ , which according to the definition (15.59) is the one-form

$$d\varphi \equiv \frac{\partial \varphi}{\partial x^\nu} dx^\nu. \quad (15.63)$$

The next simplest case is the exterior derivative of a one-form  $\mathbf{a}$ , which according to the definition (15.59) is the 2-form

$$\begin{aligned} d\mathbf{a} &= d(a_\nu dx^\nu) = \frac{\partial a_\nu}{\partial x^\mu} dx^\mu \wedge dx^\nu \\ &= \frac{1}{2} \left( \frac{\partial a_\nu}{\partial x^\mu} - \frac{\partial a_\mu}{\partial x^\nu} \right) dx^\mu \wedge dx^\nu \end{aligned} \quad (15.64)$$

$$= \frac{1}{2} (D_\mu a_\nu - D_\nu a_\mu) dx^\mu \wedge dx^\nu , \quad (15.65)$$

implicitly summed over both indices  $\mu$  and  $\nu$ . The factor 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,

$$d(\mathbf{a} \wedge \mathbf{b}) \equiv (da) \wedge \mathbf{b} + (-)^p \mathbf{a} \wedge (db) . \quad (15.66)$$

### 15.8.1 The square of the exterior derivative vanishes

The exterior derivative has the notable property that its square vanishes,

$$dd\mathbf{a} = \frac{1}{p!} \partial_{[\nu_1 \nu_2} a_{\mu_1 \dots \mu_p]} dx^{\nu_1} \wedge dx^{\nu_2} \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}] = 0 , \quad (15.67)$$

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),

$$\mathbf{D} \wedge \mathbf{D} \wedge \mathbf{a} = 0 . \quad (15.68)$$

## 15.9 An alternative notation for differential forms

The mathematicians' abstract notation for differential forms, where the  $p$ -volume element is absorbed into the definition of a  $p$ -form, is elegant, but can be hard to parse. On the other hand the wedge notation  $dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}$  is unwieldy, and seemingly wedded to coordinate frames. It is useful to introduce a compromise notation, equation (15.74), that is more in line with what physicists commonly use. The notation keeps the  $p$ -volume element of integration explicit, but is compact, and flexible enough to work in tetrad as well as coordinate frames.

Recall that an interval  $dx^\mu$  between two points of spacetime can be written as the abstract vector interval  $\mathbf{dx}$ , equation (2.19),

$$\mathbf{dx} \equiv e_\mu dx^\mu = \gamma_m e^m_\mu dx^\mu . \quad (15.69)$$

Define the invariant  $p$ -volume element  $\mathbf{d}^p \mathbf{x}$  to be the normalized wedge product of  $p$  copies of  $\mathbf{dx}$ ,

$$\boxed{\mathbf{d}^p \mathbf{x} \equiv \frac{(-)^{[p/2]}}{p!} \overbrace{\mathbf{dx} \wedge \dots \wedge \mathbf{dx}}^{p \text{ terms}} = \frac{(-)^{[p/2]}}{p!} \mathbf{e}_{\mu_1} \wedge \dots \wedge \mathbf{e}_{\mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}}, \quad (15.70)$$

which is both a  $p$ -form and a grade- $p$  multivector. The  $(-)^{[p/2]}$  sign is introduced to cancel the sign of

$$\boldsymbol{\gamma}^A \cdot \boldsymbol{\gamma}_B = (-)^{[p/2]} \delta_B^A, \quad (15.71)$$

thus ensuring that equation (15.74) holds. The  $1/p!$  factor ensures that each distinct antisymmetric sequence  $\mu_1 \dots \mu_p$  of indices counts just once. The  $1/p!$  factor could be omitted from the rightmost expression of equations (15.70) if the implicit sum were only over distinct antisymmetric sequences of indices. Denote the components of the invariant  $p$ -volume element by  $d^p x^A$ , where  $A$  runs over distinct antisymmetric sequences of  $p$  indices (in  $N$  dimensions, there are  $N!/[p!(N-p)!]$  such distinct sequences). Expanded in components, the invariant  $p$ -volume element is

$$\mathbf{d}^p \mathbf{x} = (-)^{[p/2]} \boldsymbol{\gamma}_A d^p x^A, \quad (15.72)$$

with  $A$  implicitly summed over distinct antisymmetric sequences of  $p$  indices. Notice that the notation has been switched from coordinate-frame to tetrad-frame, which is fine because  $\mathbf{d}^p \mathbf{x}$  is a (coordinate and tetrad) gauge-invariant object, so can be expanded in any frame. Equation (15.72) remains valid in a coordinate frame, where the tetrad basis elements  $\boldsymbol{\gamma}_m$  are the coordinate tangent vectors  $\mathbf{e}_\mu$ . The components  $d^p x^A$  of the  $p$ -volume element form a contravariant antisymmetric tensor of rank  $p$ . The notation  $d^2 x^{kl}$  for an areal element is to be distinguished from the notation  $ds^2 = \mathbf{dx} \cdot \mathbf{dx}$  for the scalar interval squared.

If  $\mathbf{a} = a_A \boldsymbol{\gamma}^A$  is a grade- $p$  multivector, then the corresponding  $p$ -form is defined to be the scalar product

$$\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}, \quad (15.73)$$

the dot signifying that the scalar part of the geometric product of the two grade- $p$  multivectors  $\mathbf{a}$  and  $\mathbf{d}^p \mathbf{x}$  is to be taken. In components, the  $p$ -form is  $\mathbf{a} \cdot \mathbf{d}^p \mathbf{x} = (-)^{[p/2]} a_A \boldsymbol{\gamma}^A \cdot \boldsymbol{\gamma}_B d^p x^B$ , which simplifies to

$$\boxed{\mathbf{a} \cdot \mathbf{d}^p \mathbf{x} = a_A d^p x^A}, \quad (15.74)$$

with  $A$  implicitly summed over distinct antisymmetric sequences of  $p$  indices. In a coordinate frame, the right hand side of equation (15.74) coincides with the right hand side of equation (15.52).

In this notation, the exterior derivative of a  $p$ -form  $\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}$  is the  $(p+1)$ -form corresponding to the torsion-free covariant spacetime curl  $\mathbf{D} \wedge \mathbf{a}$  of the grade- $p$  multivector  $\mathbf{a}$ ,

$$\boxed{d(\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}) \equiv (\mathbf{D} \wedge \mathbf{a}) \cdot \mathbf{d}^{p+1} \mathbf{x} = (\mathbf{D} \wedge \mathbf{a})_A d^{p+1} x^A}, \quad (15.75)$$

with  $A$  implicitly summed over distinct antisymmetric sequences of  $p+1$  indices. The components of the torsion-free covariant curl in equation (15.75) are

$$(\mathbf{D} \wedge \mathbf{a})_{kl\dots m} = (p+1) D_{[k} a_{l\dots m]} . \quad (15.76)$$

In a coordinate frame, the right hand side of equation (15.75) coincides with the right hand side of equation (15.60).

## 15.10 Hodge dual form

The Hodge dual  ${}^*\mathbf{a}$  of a differential form  $\mathbf{a}$  is most easily defined using the geometric algebra and the notation of §15.9. Given a  $p$ -form  $\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}$ , a dual  $q$ -form  ${}^*(\mathbf{a} \cdot \mathbf{d}^p \mathbf{x})$  with  $q = N - p$  can be defined by either of the two equivalent expressions

$${}^*(\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}) \equiv (\mathbf{a} I_N) \cdot \mathbf{d}^q \mathbf{x} = \mathbf{a} \cdot (I_N \mathbf{d}^q \mathbf{x}), \quad (15.77)$$

where  $I_N$  is the pseudoscalar of the geometric algebra in  $N$  dimensions. The dual  $q$ -volume element on the right hand side of equation (15.77) can be written in the notation of equation (13.22)

$${}^*\mathbf{d}^q \mathbf{x} \equiv I_N \mathbf{d}^q \mathbf{x}. \quad (15.78)$$

This is a  $q$ -volume element, but a multivector of grade  $p = N - q$ . Expanded in components, the invariant dual  $q$ -volume element is

$${}^*\mathbf{d}^q \mathbf{x} = (-)^{[q/2]} I_N \boldsymbol{\gamma}_B d^q x^B = \boldsymbol{\gamma}_A \varepsilon^A{}_B d^q x^B = (-)^{[p/2]} \boldsymbol{\gamma}_A {}^*d^q x^A, \quad (15.79)$$

where  $A$  and  $B$  run over distinct antisymmetric sequences of respectively  $p$  and  $q$  indices,  $\varepsilon^{AB}$  is the totally antisymmetric tensor normalized to  $\varepsilon^{1\dots N} = 1$  in a locally inertial tetrad frame, as is the convention of this book, and the components  ${}^*d^q x^A$  of the dual  $q$ -volume element are defined to be

$${}^*d^q x^A \equiv (-)^{[p/2]} \varepsilon^A{}_B d^q x^B, \quad (15.80)$$

implicitly summed over distinct antisymmetric sequences  $B$  of  $q$  indices. The components  ${}^*d^q x^A$  of the dual  $q$ -volume element form a contravariant antisymmetric tensor of rank  $p$ .

In components, the dual  $q$ -form (15.77) is

$$\boxed{\mathbf{a} \cdot {}^*\mathbf{d}^q \mathbf{x} = a_A {}^*d^q x^A = (-)^{[p/2]} a_A \varepsilon^A{}_B d^q x^B}. \quad (15.81)$$

The  $q$ -form corresponding to the Hodge dual  $I_N \mathbf{a}$  of a grade- $p$  multivector  $\mathbf{a}$  differs from the dual  $q$ -form by a factor of  $(-)^{pq}$  coming from  $I_N \mathbf{a} = (-)^{pq} \mathbf{a} I_N$ , equation (13.26),

$$(I_N \mathbf{a}) \cdot \mathbf{d}^q \mathbf{x} = (-)^{pq} \mathbf{a} \cdot {}^*\mathbf{d}^q \mathbf{x}. \quad (15.82)$$

It would have been possible to define the  $q$ -form dual to the  $p$ -form  $\mathbf{a} \cdot \mathbf{d}^p \mathbf{x}$  to be  $(I_N \mathbf{a}) \cdot \mathbf{d}^q \mathbf{x}$ , but this book adopts the convention that the dual  $q$ -form is  $(\mathbf{a} I_N) \cdot \mathbf{d}^q \mathbf{x}$ , equation (15.77).

In the mathematicians' notation, the dual  ${}^*\mathbf{a}$  of a  $p$ -form  $\mathbf{a} = a_A \mathbf{d}^p x^A$  is the  $q$ -form, from equation (15.81),

$${}^*\mathbf{a} = (-)^{[p/2]} a_A \varepsilon^A{}_B d^q x^B, \quad (15.83)$$

implicitly summed over distinct antisymmetric sequences  $A$  and  $B$  of respectively  $p$  and  $q$  indices.

### 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 coordinate- and tetrad-frame  $p$ -volume elements, which are tensors, are related by the vielbein in the usual way, leading to the result that

$$(1/e) d^p x^{\mu_1 \dots \mu_p} = d^p x^{1 \dots p}, \quad (15.84)$$

where  $e$  is the determinant of the  $p \times N$  vielbein matrix  $e_m^{\mu}$  with  $m$  running from 1 to  $p$  and  $\mu$  running from 1 to  $N$ ,

$$e \equiv \frac{N!}{(N-p)!} e_1^{[\mu_1} \dots e_p^{\mu_p]} . \quad (15.85)$$

The factor of  $N!/(N-p)!$  in equation (15.85) arises because there are that many distinct terms in the determinant, one for each choice of  $\mu_1 \dots \mu_p$ , and antisymmetrization over indices conventionally denotes the average, not the sum. Thus the  $N!/(N-p)!$  factor ensures that each term in the determinant appears with coefficient  $\pm 1$ , as it should. Equation (15.84) implies that  $(1/e) d^p x^{\mu_1 \dots \mu_p}$  is an invariant measure of the  $p$ -volume element. Despite the lack of indices,  $e$  is not a scalar, but rather is the antisymmetric coordinate and tetrad tensor defined by the right hand side of equation (15.85).

Dual  $q$ -volume elements are related similarly,

$$(1/e) {}^*d^q x^{\mu_1 \dots \mu_p} = {}^*d^q x^{1 \dots p}, \quad (15.86)$$

with the same  $e$ , equation (15.85).

### 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. 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  $\partial V$  of the hypersurface:

$$\int_V d\mathbf{a} = \oint_{\partial V} \mathbf{a} . \quad (15.87)$$

In the more explicit notation of §15.9, if  $\mathbf{a} = a_A \gamma^A$  is a grade- $p$  multivector, Stokes' theorem states

$$\boxed{\int_V (\mathbf{D} \wedge \mathbf{a})_B d^{p+1}x^B = \oint_{\partial V} a_A d^p x^A}, \quad (15.88)$$

where  $(\mathbf{D} \wedge \mathbf{a})_B = (p+1) D_{[n} a_{m_1 \dots m_p]} \gamma^B$  denotes the components of the torsion-free covariant curl  $\mathbf{D} \wedge \mathbf{a}$ , equation (15.38b), and  $A$  and  $B$  are implicitly summed over distinct antisymmetric sequences of  $p$  and  $p+1$  indices respectively.

In the case of a 0-form (scalar)  $\varphi$ , the exterior derivative  $d\varphi$ , equation (15.63), 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  $\lambda$ , from initial value  $\lambda_0$  to final value  $\lambda_1$ , is

$$\int_{\lambda_0}^{\lambda_1} d\varphi = \int_{\lambda_0}^{\lambda_1} \frac{\partial \varphi}{\partial x^\nu} dx^\nu = \int_{\lambda_0}^{\lambda_1} \frac{\partial \varphi}{\partial x^\nu} \frac{dx^\nu}{d\lambda} d\lambda = \int_{\lambda_0}^{\lambda_1} \frac{d\varphi}{d\lambda} d\lambda = \varphi(\lambda_1) - \varphi(\lambda_0). \quad (15.89)$$

Equation (15.89) can be recognized as the fundamental theorem of calculus. Equation (15.89) is equation (15.87) or (15.88) for the case where  $\mathbf{a}$  is the 0-form (scalar)  $\varphi$ . The hypersurface  $V$  is the 1-dimensional path of integration. The boundary  $\partial V$  is the two endpoints of the path.

Here is a sketch of a proof of the generalized Stokes' theorem (15.87). 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

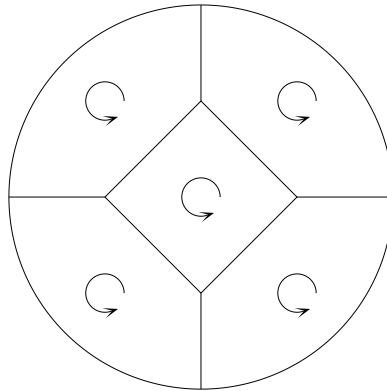


Figure 15.1 Partition of a disk into five rectangular patches. The arrowed circles show the direction of circulation of the integral over the boundary of each patch.

$(p+1)$ -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 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

$$\begin{aligned}
\int_{\text{patch}} d\mathbf{a} &= \int_{y_0}^{y_1} \int_{x_0}^{x_1} \left( \frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y} \right) dx \wedge dy \\
&= \int_{y_0}^{y_1} \left( \int_{x_0}^{x_1} \frac{\partial a_y}{\partial x} dx \right) \wedge dy - \int_{x_0}^{x_1} \left( \int_{y_0}^{y_1} \frac{\partial a_x}{\partial y} dy \right) \wedge dx \\
&= \int_{y_0}^{y_1} \left( \int_{x_0}^{x_1} \frac{\partial a_y}{\partial x} dx \right) dy - \int_{x_0}^{x_1} \left( \int_{y_0}^{y_1} \frac{\partial a_x}{\partial y} dy \right) dx \\
&= \int_{y_0}^{y_1} [a_y(x_1) - a_y(x_0)] dy - \int_{x_0}^{x_1} [a_x(y_1) - a_x(y_0)] dx \\
&= \oint_{\partial\text{patch}} a_\mu dx^\mu = \oint_{\partial\text{patch}} \mathbf{a} . \tag{15.90}
\end{aligned}$$

The first line of equations (15.90) is the standard expression (15.65) 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 (15.90) rearranges the first. The third line of equations (15.90) 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 (15.58). The fourth line of equations (15.90) follows from the fundamental theorem of calculus, equation (15.89). 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 (15.90) expresses the fourth in more compact notation, with  $\partial\text{patch}$  denoting the boundary, the 4 edges, of the patch. Equations (15.90) 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  $\phi$  around a circle. The integral of the angle around the circle is

$$\int_0^{2\pi} d\phi = 2\pi. \quad (15.91)$$

But since the circle has no boundary, should not Stokes' theorem imply that the integral vanishes? The resolution of the problem is that  $\phi$  is not a single-valued scalar. The 1-form  $d\phi$  constructed from  $\phi$  is well-defined, being single-valued and continuous everywhere on the circle, but  $\phi$  itself is not. The circle can be cut at any point, and a single-valued scalar  $\phi$  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  $2\pi$ .

A differential form  $\mathbf{F}$  is said to be **exact** if it can be expressed as the exterior derivative of a differential form  $\mathbf{A}$ ,

$$\mathbf{F} = d\mathbf{A}. \quad (15.92)$$

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  $d\phi$  is not exact, because  $\phi$  is not a single-valued 0-form (scalar).

A differential form  $\mathbf{F}$  is said to be **closed** if its exterior derivative vanishes,

$$d\mathbf{F} = 0. \quad (15.93)$$

The rule  $dd = 0$  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  $x$ - $y$  plane, the 1-form  $d\phi$  is

$$d\phi = d\tan^{-1}(y/x) = \frac{x dy - y dx}{x^2 + y^2}, \quad (15.94)$$

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 (15.88) 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 (15.88) can be converted to a covariant divergence by post-multiplying by the pseudoscalar  $I_N$ ,

$$\mathbf{D} \wedge (\mathbf{a} I_N) = (\mathbf{D} \cdot \mathbf{a}) I_N . \quad (15.95)$$

If  $\mathbf{a} = a_A \boldsymbol{\gamma}^A$  is a grade- $p$  multivector, substituting equation (15.95) into Stokes' theorem (15.88) gives the generalized Gauss' theorem, with  $q \equiv N - p$ ,

$$\boxed{\int_V (\mathbf{D} \cdot \mathbf{a})_B {}^* d^{q+1}x^B = \oint_{\partial V} a_A {}^* d^q x^A} , \quad (15.96)$$

where  $(\mathbf{D} \cdot \mathbf{a})_B = D^{m_1} a_{m_1 \dots m_p}$  denotes the components of the torsion-free covariant divergence  $\mathbf{D} \cdot \mathbf{a}$ , equation (15.38a),  ${}^* d^q x^A$  denotes the components of the dual  $q$ -volume element as defined in §15.10, and  $A$  and  $B$  are implicitly summed over distinct antisymmetric sequences of  $p$  and  $p-1$  indices respectively.

In the remainder of this book, the dual  $q$ -volume element  ${}^* d^q x^{k \dots n}$  is often abbreviated to  $d^q x^{k \dots n}$  without the Hodge star symbol, since the dual nature is usually evident from the number of indices  $k \dots n$ , 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  $(1/e) d^N x$  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. However, this book does use the convention  $d^p \mathbf{x}$  in boldface for the abstract  $p$ -volume element, equation (15.70).

The simplest and most common application of Gauss' theorem is where  $\mathbf{a} = a_m \boldsymbol{\gamma}^m$  is a vector (a grade-1 multivector), in which case

$$\boxed{\int_V D^m a_m d^N x = \oint_{\partial V} a_m d^{N-1} x^m} , \quad (15.97)$$

where, as just remarked,  $d^N x$  and  $d^{N-1} x^m$  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  $\delta^p(x) \equiv \boldsymbol{\gamma}^A \delta_A^p(x)$  is defined such that for any scalar function  $f(x)$ ,

the integral

$$\int f(x) \delta^p(x) \cdot d^p x = \int f(x) \delta_A^p(x) d^p x^A = f(0) , \quad (15.98)$$

yields the value  $f(0)$  of the function at the origin when integrated over any  $p$ -dimensional volume  $V$  containing the origin  $x = 0$ . The Dirac delta-function  $\delta^p(x)$  is a grade- $p$  multivector with components  $\delta_A^p(x)$ , where  $A$  runs over distinct antisymmetric sequences of  $p$  indices.

The dual  $q$ -dimensional Dirac delta-function  ${}^*\delta^q(x) = \gamma^A {}^*\delta_A^q(x)$  with  $q \equiv N - p$ , is defined to behave similarly when integrated over the dual  $q$ -volume element  ${}^*d^q x$  defined by equation (15.78),

$$\int f(x) {}^*\delta^q(x) \cdot {}^*d^q x = \int f(x) {}^*\delta_A^q(x) {}^*d^q x^A = f(0) . \quad (15.99)$$

The dual  $q$ -dimensional Dirac delta-function  ${}^*\delta^q(x)$  is a grade- $p$  multivector with components  ${}^*\delta_A^q(x)$  where  $A$  runs over distinct antisymmetric sequences of  $p$  indices.

As with the dual  $q$ -volume, the dual Dirac delta-function  ${}^*\delta_{k\dots n}^q(x)$  will often be abbreviated in this book to  ${}^*\delta_{k\dots n}^q(x)$  without the Hodge star symbol, since the dual nature can usually be inferred from the number of indices.

The most common use of the Dirac delta-function is in integration over  $N$ -dimensional space,

$$\int f(x) \delta^N(x) d^N x = f(0) ,$$

(15.100)

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.

# 16

---

## Action principle for electromagnetism and gravity

One of the profound realizations 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  $U(1)$  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  $U_Y(1) \times SU(2)$ , which through interactions with a scalar field called the Higgs field breaks down to the electromagnetic group  $U(1)$  at collision energies less than the electroweak scale of about 1 TeV (the  $U_Y(1)$  electroweak hypercharge group is not the same as the  $U(1)$  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  $SU(3)$ .

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 and the 4-dimensional group of translations.

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) 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 string theories. 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  $\varphi$ , 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, §16.13, uses multivector-valued differential forms. The multivector forms notation provides an elegant formulation of the definitions of curvature and torsion, equations (16.203) and (16.207), first formulated by Cartan (1904), and elegant versions of the equations of motion (16.234) 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 (15.97), 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.

## 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  $L(\varphi, \dot{D}_\mu \varphi)$  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  $\dot{D}_\mu \varphi$  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, it is precisely the torsion-free covariant derivative that can be integrated to yield surface terms.

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  $\delta$  symbol. For example, the derivative of the functional  $L$  with respect to the function  $\varphi$  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  $\lambda_i$  and  $\lambda_f$ . The variation  $\delta S$  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  $\pi^\mu$  conjugate to the field  $\varphi$  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

$$\delta S = \left[ \oint \pi_\mu \delta\varphi d^3x^\mu \right]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left( \frac{\delta L}{\delta \varphi} - \mathring{D}_\mu \pi^\mu \right) \delta\varphi d^4x = 0. \quad (16.6)$$

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,

$$\mathring{D}_\mu \pi^\mu = \frac{\delta L}{\delta \varphi}. \quad (16.7)$$

All of the above derivations carry through with the field  $\varphi$  replaced by a set of fields  $\varphi_i$ , 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  $\varphi$  and its conjugate momenta  $\pi^\mu$ , defined in terms of the Lagrangian by equation (16.4).

Varying the action with Lagrangian (16.8) with respect to the field  $\varphi$  and its conjugate momenta  $\pi^\mu$  gives

$$\delta S = \int_{\lambda_i}^{\lambda_f} \left( \pi^\mu \dot{D}_\mu \delta \varphi + \delta \pi^\mu \dot{D}_\mu \varphi - \frac{\delta H}{\delta \varphi} \delta \varphi - \delta \pi^\mu \frac{\delta H}{\delta \pi^\mu} \right) d^4x . \quad (16.9)$$

Integrating the first term in the integrand by parts brings the variation of the action to

$$\delta S = \left[ \oint \pi^\mu \delta \varphi d^3x_\mu \right]_{\lambda_i}^{\lambda_f} + \int_{\lambda_i}^{\lambda_f} \left[ - \left( \dot{D}_\mu \pi^\mu + \frac{\delta H}{\delta \varphi} \right) \delta \varphi + \delta \pi^\mu \left( \dot{D}_\mu \varphi - \frac{\delta H}{\delta \pi^\mu} \right) \right] d^4x . \quad (16.10)$$

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  $\varphi$  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  $x^\mu$  are split into a time coordinate  $t$  and spatial coordinates  $x^\alpha$ . The momentum  $\pi$  conjugate to the field  $\varphi$  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

$$H = \pi \dot{D}_t \varphi - L . \quad (16.13)$$

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 spin angular-momentum (§16.10.1), and general coordinate transformations implying conservation of energy-momentum (§16.10.2).

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  $\epsilon$  is an infinitesimal parameter. The torsion-free covariant derivatives  $\mathring{D}_n\varphi$  of the field transform correspondingly as

$$\mathring{D}_n\varphi \rightarrow \mathring{D}_n\varphi + \epsilon \mathring{D}_n(\delta\varphi) . \quad (16.15)$$

The variation (16.14) is a symmetry of the field if the Lagrangian  $L(\varphi, \mathring{D}_n\varphi)$  is unchanged by it. The vanishing

of the variation of the Lagrangian implies

$$\begin{aligned}
0 &= \frac{\delta L}{\delta \epsilon} \\
&= \frac{\delta L}{\delta \varphi} \delta \varphi + \frac{\delta L}{\delta (\mathring{D}_n \varphi)} \mathring{D}_n(\delta \varphi) \\
&= \mathring{D}_n \left( \frac{\delta L}{\delta (\mathring{D}_n \varphi)} \delta \varphi \right) + \left( \frac{\delta L}{\delta \varphi} - \mathring{D}_n \frac{\delta L}{\delta (\mathring{D}_n \varphi)} \right) \delta \varphi \\
&= \mathring{D}_n (\pi^n \delta \varphi) + \left( \frac{\delta L}{\delta \varphi} - \mathring{D}_n \pi^n \right) \delta \varphi,
\end{aligned} \tag{16.16}$$

with  $\pi^n$  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  $j^n$  defined by

$$j^n \equiv \pi^n \delta \varphi \tag{16.17}$$

is covariantly conserved,

$$\mathring{D}_n j^n = 0. \tag{16.18}$$

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  $U(1)$  of rotations about a circle.

### 16.5.1 Electromagnetic gauge transformations

Under an electromagnetic gauge transformation, a field  $\varphi$  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  $\varphi$  involves the torsion-free derivative  $\mathring{D}_\nu \varphi$  of the field. The torsion-free covariant derivative is prescribed because (a) only covariant derivatives transform as tensors under arbitrary coordinate transformations, and (b) least action involves integration by parts, and it is the torsion-free covariant derivative that integrates to yield surface terms. To ensure that the Lagrangian remains invariant also under an electromagnetic gauge transformation (16.19), the derivative  $\mathring{D}_\nu$  must be augmented by an electromagnetic connection  $A_\nu$ , which equals the thing historically known as the **electromagnetic potential**. The result is an electromagnetic gauge-covariant derivative  $\mathring{D}_\nu + ieA_\nu$  with the

defining property that, when acting on the charged field  $\varphi$ , it transforms under the electromagnetic gauge transformation (16.19) as

$$(\mathring{D}_\nu + ieA_\nu)\varphi \rightarrow e^{-ie\theta}(\mathring{D}_\nu + ieA_\nu)\varphi . \quad (16.20)$$

In other words, the gauge-covariant derivative of the field  $\varphi$  is required to transform under electromagnetic gauge transformations in the same way as the field  $\varphi$ . The gauge-covariant derivative  $\mathring{D}_\nu + ieA_\nu$  transforms correctly provided that the gauge field  $A_\nu$  transforms under the electromagnetic gauge transformation (16.19) as

$$A_\nu \rightarrow A_\nu + \mathring{D}_\nu\theta . \quad (16.21)$$

Since  $\theta$  is a scalar phase, its covariant derivative reduces to its partial derivative,  $\mathring{D}_\nu\theta = \partial\theta/\partial x^\nu$ .

### 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  $A_\nu$  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 (2.71). 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  $D_\mu$ , 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

$L_e \equiv -\frac{1}{16\pi} F^{\mu\nu} F_{\mu\nu} ,$

(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  $\epsilon_0 = \mu_0 = 1$ ), for which the normalization factor is 1/4 instead of 1/(16 $\pi$ ).

The momenta conjugate to the electromagnetic coordinates  $A_\nu$  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 $\pi$ , which explains why high-energy theorists prefer Heaviside units.

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  $j^\nu$  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  $A_\nu$  and their velocities  $\mathring{D}_\mu A_\nu$ ,

$$S = \int_{\lambda_i}^{\lambda_f} \left[ -\frac{1}{16\pi} (\mathring{D}^\mu A^\nu - \mathring{D}^\nu A^\mu) (\mathring{D}_\mu A_\nu - \mathring{D}_\nu A_\mu) + j^\nu A_\nu \right] d^4x . \quad (16.32)$$

Varying the action (16.32) with respect to the electromagnetic coordinates  $A_\nu$  and their velocities  $\mathring{D}_\mu A_\nu$ , along the same lines as equations (16.2)–(16.6) for the template field  $\varphi$ , yields

$$\delta S = -\frac{1}{4\pi} \left[ \oint F^{\mu\nu} \delta A_\nu d^3x_\mu \right]_{\lambda_i}^{\lambda_f} + \frac{1}{4\pi} \int_{\lambda_i}^{\lambda_f} (\mathring{D}_\mu F^{\mu\nu} + 4\pi j^\nu) \delta A_\nu d^4x . \quad (16.33)$$

Least action requires that the variation of the action with respect to arbitrary variations  $\delta A_\nu$  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

$$\mathring{D}_\mu F^{\nu\mu} = 4\pi j^\nu . \quad (16.34)$$

The factor  $4\pi$  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.

### 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  $p^\mu = -F^{\mu\nu}/4\pi$ ,

$$L_e = -\frac{1}{4\pi} F^{\mu\nu} \mathring{D}_\mu A_\nu - H , \quad (16.35)$$

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  $A_\nu$  and the momenta  $F^{\mu\nu}$ .

Varying the action with Lagrangian (16.35) with respect to the coordinates  $A_\nu$  and momenta  $F^{\mu\nu}$  gives

$$\delta S = \frac{1}{4\pi} \int (-F^{\mu\nu} \mathring{D}_\mu \delta A_\nu - \delta F^{\mu\nu} \mathring{D}_\mu A_\nu + \frac{1}{2} \delta F^{\mu\nu} F_{\mu\nu} + 4\pi j^\nu \delta A_\nu) d^4x . \quad (16.37)$$

Integrating the first term in the integrand of equation (16.37) by parts yields

$$\delta S = -\frac{1}{4\pi} \oint F^{\mu\nu} \delta A_\nu d^3x_\mu + \frac{1}{4\pi} \int [(\mathring{D}_\mu F^{\mu\nu} + 4\pi j^\nu) \delta A_\nu - \frac{1}{2} (\mathring{D}_\mu A_\nu - \mathring{D}_\nu A_\mu + F_{\mu\nu}) \delta F^{\mu\nu}] d^4x . \quad (16.38)$$

The surface term vanishes provided that the electromagnetic coordinates  $A_\nu$  are held fixed on the boundary.

Requiring that the variation of the action vanish with respect to arbitrary variations  $\delta A_\nu$  and  $\delta F^{\mu\nu}$  of the coordinates and momenta then yields Hamilton's equations,

$$\mathring{D}_\mu F^{\nu\mu} = 4\pi j^\nu , \quad (16.39a)$$

$$\mathring{D}_\mu A_\nu - \mathring{D}_\nu A_\mu = F_{\mu\nu} . \quad (16.39b)$$

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  $A_\nu$  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  $j^\nu$ ,

$$\mathring{D}_\nu j^\nu = 0 . \quad (16.40)$$

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  $A_\nu$  varies as, equation (16.19),

$$\delta A_\nu = \mathring{D}_\nu \theta . \quad (16.41)$$

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

$$\delta S_q = \int j^\nu \mathring{D}_\nu \theta d^4x . \quad (16.42)$$

Integrating equation (16.42) by parts gives

$$\delta S_q = \oint j^\nu \theta d^3x_\nu - \int (\mathring{D}_\nu j^\nu) \theta d^4x . \quad (16.43)$$

Electromagnetic gauge-invariance requires that the variation vanish with respect to arbitrary choices of the gauge parameter  $\theta$ , subject to the condition that  $\theta$  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  $A_\nu$ ,

$$-\mathring{\square}A_\nu + \mathring{R}_{\nu\lambda}A^\lambda + \mathring{D}_\nu\mathring{D}_\mu A^\mu = 4\pi j_\nu , \quad (16.45)$$

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  $j_\nu$  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  $A_\nu$ . 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  $A_\nu$ , 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  $A_\nu$ , equation (16.45) (see §26.6 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

$$3 \text{ equations: } \dot{D}_t F^{\alpha t} + \dot{D}_\beta F^{\alpha\beta} = 4\pi j^\alpha , \quad (16.46a)$$

$$3 \text{ equations: } \dot{D}_t A_\alpha - \dot{D}_\alpha A_t = F_{t\alpha} . \quad (16.46b)$$

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  $A_\alpha$ . The physical degrees of freedom are thus identified as the 3 spatial coordinates  $A_\alpha$  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  $B^\alpha \equiv \epsilon^{t\alpha\beta\gamma} F_{\beta\gamma}$  of the magnetic field, to the spatial curl of the spatial coordinates  $A_\alpha$ . Since the equations of motion (16.46b) already determine completely the spatial coordinates  $A_\alpha$ , 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  $A_\alpha$  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 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  $\mathbf{D} \equiv (\mathring{D}_{\mu} + ieA_{\mu})dx^{\mu}$  for brevity, the field 2-form  $\mathbf{F}$  is defined by the commutator of the gauge-covariant derivative,

$$[\mathbf{D}, \mathbf{D}] \equiv ie\mathbf{F}. \quad (16.49)$$

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  $\mathbf{A}$ ,

$$\mathbf{F} = \mathring{\mathbf{D}} \wedge \mathbf{A}. \quad (16.52)$$

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  $\mathbf{j} \equiv j_n \gamma^n$  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 (15.78). 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

$$(\mathbf{a} \cdot \mathbf{b}) \cdot *d^4x \equiv (\mathbf{a} \cdot \mathbf{b}) \cdot (Id^4x) = ((\mathbf{a} \cdot \mathbf{b})I) \cdot d^4x = (I(\mathbf{a} \cdot \mathbf{b})) \cdot d^4x = ((I\mathbf{a}) \wedge \mathbf{b}) \cdot d^4x , \quad (16.55)$$

where the second expression is the definition (15.78) of the dual volume element, the third expression is an application of the multivector triple-product relation (13.37), 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 (13.37). 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, in accordance with equation (15.74), 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 (16.51). The symbol  $*$  denotes the form dual, equation (15.83). The dual  $*F$  is a 2-form, while the dual  $*j$  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  $p \wedge dq - H$  with coordinates  $q = A$  and momenta  $p = *F/4\pi$ ,

$$L_e + L_q = \frac{1}{4\pi} *F \wedge dA - H , \quad (16.59)$$

and super-Hamiltonian 4-form

$$H = \frac{1}{8\pi} *F \wedge F - *j \wedge A . \quad (16.60)$$

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

$$\dot{\mathbf{D}} \cdot \mathbf{F} = -4\pi \mathbf{j} , \quad (16.66a)$$

$$\dot{\mathbf{D}} \wedge \mathbf{A} = \mathbf{F} . \quad (16.66b)$$

Applying the multivector triple-product relation (13.38) gives the multivector identities (the torsion-free curl of  $\mathbf{A}$  vanishes, equation (15.43), so  $\dot{\mathbf{D}}\dot{\mathbf{D}}\mathbf{A}$  has only a vector part, no trivector part)

$$\begin{aligned} \dot{\mathbf{D}}\dot{\mathbf{D}}\mathbf{A} &= \dot{\mathbf{D}}(\dot{\mathbf{D}}\mathbf{A}) = \dot{\mathbf{D}} \cdot (\dot{\mathbf{D}} \wedge \mathbf{A}) + \dot{\mathbf{D}} \wedge (\dot{\mathbf{D}} \cdot \mathbf{A}) \\ &= (\dot{\mathbf{D}}\dot{\mathbf{D}})\mathbf{A} = (\dot{\mathbf{D}} \cdot \dot{\mathbf{D}})\mathbf{A} + (\dot{\mathbf{D}} \wedge \dot{\mathbf{D}}) \cdot \mathbf{A} . \end{aligned} \quad (16.67)$$

Eliminating  $\mathbf{F}$  from Hamilton's equations (16.66) then yields a second order differential equation for the electromagnetic potential  $\mathbf{A}$ ,

$$-\ddot{\square}\mathbf{A} - (\dot{\mathbf{D}} \wedge \dot{\mathbf{D}}) \cdot \mathbf{A} + \dot{\mathbf{D}}(\dot{\mathbf{D}} \cdot \mathbf{A}) = 4\pi\mathbf{j}, \quad (16.68)$$

where  $\ddot{\square} \equiv \dot{\mathbf{D}} \cdot \dot{\mathbf{D}}$  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  $\dot{\mathbf{D}} \cdot \mathbf{A} = 0$ , in which case equation (16.68) reduces to

$$-\ddot{\square}\mathbf{A} - (\dot{\mathbf{D}} \wedge \dot{\mathbf{D}}) \cdot \mathbf{A} = 4\pi\mathbf{j}, \quad (16.69)$$

or more simply

$$-(\dot{\mathbf{D}} \dot{\mathbf{D}}) \mathbf{A} = 4\pi\mathbf{j}. \quad (16.70)$$

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 to break general covariance by splitting spacetime into time and space coordinates,  $x^{\textcolor{brown}{\mu}} = \{t, x^{\alpha}\}$ , 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  $\mathbf{a}$  represents the remaining all-spatial components (the spatial indices being suppressed for brevity). The bar on the time index  $\bar{t}$  serves to distinguish the form  $\mathbf{a}_{\bar{t}} \equiv \mathbf{a}_{t,A} d^p x^{\textcolor{brown}{tA}}$  from its components  $\mathbf{a}_{t,\alpha}$ . Thus a 1-form  $\mathbf{a} \equiv a_{\kappa} dx^{\kappa}$  splits into

$$\mathbf{a} \rightarrow \mathbf{a}_{\bar{t}} + \mathbf{a} \equiv a_{\bar{t}} dt + a_{\alpha} dx^{\alpha}, \quad (16.71)$$

while a 2-form  $\mathbf{a} \equiv a_{\kappa\lambda} dx^{\kappa\lambda}$  splits into

$$\mathbf{a} \rightarrow \mathbf{a}_{\bar{t}} + \mathbf{a} \equiv a_{t\alpha} d^2 x^{t\alpha} + a_{\alpha\beta} d^2 x^{\alpha\beta}, \quad (16.72)$$

implicitly summed over distinct sequences of indices. The time component of the exterior product of two forms  $\mathbf{a}$  and  $\mathbf{b}$  is

$$(\mathbf{a} \wedge \mathbf{b})_{\bar{t}} = \mathbf{a}_{\bar{t}} \wedge \mathbf{b} + \mathbf{a} \wedge \mathbf{b}_{\bar{t}} \quad (16.73)$$

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

$$\mathbf{F} \rightarrow \mathbf{F}_{\bar{t}} + \mathbf{F} = F_{t\alpha} dx^{t\alpha} + F_{\beta\gamma} dx^{\beta\gamma}, \quad (16.74)$$

whose time and space parts encode the electric and magnetic fields. The dual electromagnetic field 2-form  ${}^*\mathbf{F}$  splits as

$${}^*\mathbf{F} \rightarrow {}^*\mathbf{F}_{\bar{t}} + {}^*\mathbf{F} = \varepsilon_{t\alpha\beta\gamma} F^{\beta\gamma} dx^{t\alpha} + \varepsilon_{t\alpha\beta\gamma} F^{t\alpha} dx^{\beta\gamma}, \quad (16.75)$$

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 (14.61).

The time components of Hamilton's equations (16.63) comprise 3 equations of motion for the 3 spatial components  ${}^*\mathbf{F}$  of the momenta, which is the electric field, and 3 equations of motion for the 3 spatial components  $\mathbf{A}$  of the coordinates,

$$3 \text{ equations: } (d {}^*\mathbf{F})_{\bar{t}} \equiv d_{\bar{t}} {}^*\mathbf{F} + d {}^*\mathbf{F}_{\bar{t}} = 4\pi {}^*\mathbf{j}_{\bar{t}}, \quad (16.76a)$$

$$3 \text{ equations: } (d\mathbf{A})_{\bar{t}} \equiv d_{\bar{t}} \mathbf{A} + d\mathbf{A}_{\bar{t}} = \mathbf{F}_{\bar{t}}. \quad (16.76b)$$

The exterior time derivative here is the 1-form  $d_{\bar{t}} = dt \partial/\partial t$ . 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  $d\mathbf{A}_{\bar{t}} = dx^\alpha \partial/\partial x^\alpha dt A_{\bar{t}} = -\partial A_{\bar{t}}/\partial x^\alpha d^2x^{t\alpha}$ . The remaining Hamilton's equations (16.63), those not involving any time derivatives, are

$$1 \text{ constraint: } d {}^*\mathbf{F} = 4\pi {}^*\mathbf{j}, \quad (16.77a)$$

$$3 \text{ identities: } d\mathbf{A} = \mathbf{F}. \quad (16.77b)$$

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,

$$d(d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j}) = -4\pi d {}^*\mathbf{j} = 0, \quad (16.78)$$

becomes in a 3+1 split

$$d_{\bar{t}}(d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j}) + d(d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j})_{\bar{t}} = 0. \quad (16.79)$$

The second term on the right hand side of equation (16.79) vanishes on the equation of motion (16.76a), so equation (16.79) reduces to

$$d_{\bar{t}}(d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j}) = 0. \quad (16.80)$$

If the spatial components  $d {}^*\mathbf{F} - 4\pi {}^*\mathbf{j}$  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  $\mathbf{F}$  as the curl of the potential  $\mathbf{A}$ , is a constraint in Dirac's (1964) sense, but not in the relativists' sense, since it is not guaranteed by any conservation law.

### 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,

$$\begin{aligned} \delta S = & \left[ \frac{1}{4\pi} \oint {}^* \mathbf{F} \wedge \delta \mathbf{A} \right]_{t_i}^{t_f} \\ & + \frac{1}{4\pi} \int_{t_i}^{t_f} -(\mathrm{d} {}^* \mathbf{F} - 4\pi {}^* \mathbf{j})_{\bar{t}} \wedge \delta \mathbf{A} - (\mathrm{d} {}^* \mathbf{F} - 4\pi {}^* \mathbf{j}) \wedge \delta \mathbf{A}_{\bar{t}} + {}^* \mathbf{F} \wedge (\mathrm{d} \mathbf{A} - \mathbf{F})_{\bar{t}} + {}^* \mathbf{F}_{\bar{t}} \wedge (\mathrm{d} \mathbf{A} - \mathbf{F}) . \end{aligned} \quad (16.81)$$

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  $\mathbf{A}$  and 3 spatial momenta  ${}^* \mathbf{F}$ . The 1 constraint (16.77a) arises from minimizing the action with respect to the 1 time component  $\mathbf{A}_{\bar{t}}$  of the coordinates, and the 3 identities (16.77b) from minimizing with respect to the 3 time components  ${}^* \mathbf{F}_{\bar{t}}$  of the momenta. Now  $\mathbf{A}_{\bar{t}}$  is a gauge variable: it can be adjusted arbitrarily by an electromagnetic gauge transformation,

$$\mathbf{A}_{\bar{t}} \rightarrow \mathbf{A}_{\bar{t}} + \mathrm{d}_{\bar{t}} \theta . \quad (16.82)$$

Minimizing the action with respect to the gauge variable  $\mathbf{A}_{\bar{t}}$  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 §26.6 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  $\mathbf{F}$ . The magnetic field  ${}^* \mathbf{F}_{\bar{t}}$ , or equivalently its dual  $\mathbf{F}$ , 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 with respect to the magnetic field  ${}^* \mathbf{F}_{\bar{t}}$  determines its own components through the identities (16.77b).

### 16.6.8 Conventional electromagnetic Hamiltonian

The conventional Hamiltonian  $H$  is defined by

$$H \equiv \frac{1}{4\pi} {}^* \mathbf{F} \wedge \mathrm{d}_{\bar{t}} \mathbf{A} - L . \quad (16.83)$$

The combined electromagnetic and charged interaction Lagrangian (16.59) can be written

$$L = \frac{1}{4\pi} [d(*\mathbf{F} \wedge \mathbf{A}_{\bar{t}}) + *\mathbf{F} \wedge d_{\bar{t}}\mathbf{A} - (d*\mathbf{F} - 4\pi *j) \wedge \mathbf{A}_{\bar{t}} + *\mathbf{F}_{\bar{t}} \wedge (d\mathbf{A} - \mathbf{F}) - \frac{1}{2} *\mathbf{F} \wedge \mathbf{F}_{\bar{t}} + \frac{1}{2} *\mathbf{F}_{\bar{t}} \wedge \mathbf{F} + 4\pi *j_{\bar{t}} \wedge \mathbf{A}] . \quad (16.84)$$

Dropping the total derivative term  $d(*\mathbf{F} \wedge \mathbf{A}_{\bar{t}})$  from the Lagrangian (16.84), and inserting the rest into the defining equation (16.83) yields the conventional Hamiltonian

$$H = \frac{1}{4\pi} [(d*\mathbf{F} - 4\pi *j) \wedge \mathbf{A}_{\bar{t}} - *\mathbf{F}_{\bar{t}} \wedge (d\mathbf{A} - \mathbf{F}) + \frac{1}{2} *\mathbf{F} \wedge \mathbf{F}_{\bar{t}} - \frac{1}{2} *\mathbf{F}_{\bar{t}} \wedge \mathbf{F} - 4\pi *j_{\bar{t}} \wedge \mathbf{A}] . \quad (16.85)$$

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  $(*\mathbf{F} \wedge \mathbf{F}_{\bar{t}} - *\mathbf{F}_{\bar{t}} \wedge \mathbf{F})/(8\pi)$  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  $-j \cdot \mathbf{A} d^4x$  is an interaction term.

The conventional Hamiltonian (16.85) is a function of spatial coordinates  $\mathbf{A}$  and their conjugate spatial momenta  $*\mathbf{F}$ , 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  $d\mathbf{A}$  and  $d*\mathbf{F}$  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  $\mathbf{A}(x^\alpha)$  and  $*\mathbf{F}(x^\alpha)$  as being infinite collections of fields indexed by the spatial position  $x^\alpha$ ; the spatial derivatives of the fields are then effectively linear combinations of those fields.

Varying the conventional Hamiltonian (16.85) with respect to  $\mathbf{A}$ ,  $\mathbf{A}_{\bar{t}}$ ,  $*\mathbf{F}$ , and  $*\mathbf{F}_{\bar{t}}$  recovers Hamilton's equations (16.76) and (16.77). In executing the variation, the terms involving the varied derivatives  $d\delta\mathbf{A}$  and  $d\delta*\mathbf{F}$  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**

$L_g \equiv \frac{1}{16\pi G} R ,$

(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  $c = G = 1$ )

$$L_g = \frac{1}{16\pi} e^{m\kappa} e^{n\lambda} R_{\kappa\lambda mn} = \frac{1}{16\pi} e^{m\kappa} e^{n\lambda} \left( \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{mn\lambda}}{\partial x^\lambda} + \Gamma_m^p \Gamma_{pn\lambda} - \Gamma_{m\kappa}^p \Gamma_{pn\lambda} \right). \quad (16.88)$$

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  $\gamma_{mn}$ . If the tetrad is orthonormal, then the tetrad metric is Minkowski,  $\gamma_{mn} = \eta_{mn}$ , but any tetrad with constant metric  $\gamma_{mn}$ , 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 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  $\kappa$ ) 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 (2.109), 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  $x^\mu$  of spacetime. The coordinates  $x^\mu$  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  $e_\mu$  and the associated coordinate connections, but it is not possible to do without coordinates.

If the Lorentz connections  $\Gamma_{mn\lambda}$  are taken to be the coordinates of the gravitational field, then the corresponding canonical momenta are (a factor of  $8\pi$  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)

$$\varepsilon_{\kappa\lambda\mu\nu} e^{kl\kappa\lambda} e^{mn\mu\nu} = \varepsilon^{klmn} , \quad (16.90)$$

which is a symmetric  $6 \times 6$  matrix of conditions, or 21 conditions, except that the normalization of  $\varepsilon_{\kappa\lambda\mu\nu} = -(1/e)[\kappa\lambda\mu\nu]$ , 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 (2.71), 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  $\dot{D}_\kappa$ , 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.

### 16.7.1 The Lorentz connection is not a tetrad scalar, but any variation of it is

The Lorentz connection  $\Gamma_{mn\lambda} \equiv e^l_\lambda \Gamma_{ml}$  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  $\epsilon_{nm}$ , 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  $\delta a_n$  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  $\epsilon_{nm}$  is

$$\begin{aligned} \delta \Gamma_{mn\lambda} &= \delta \left( \gamma_m \cdot \frac{\partial \gamma_n}{\partial x^\lambda} \right) = \gamma_m \cdot \frac{\partial (\epsilon_n{}^p \gamma_p)}{\partial x^\lambda} + \epsilon_m{}^p \gamma_p \cdot \frac{\partial \gamma_n}{\partial x^\lambda} = \frac{\partial \epsilon_{nm}}{\partial x^\lambda} + \epsilon_n{}^p \Gamma_{mp\lambda} + \epsilon_m{}^p \Gamma_{pn\lambda} \\ &= D_\lambda \epsilon_{nm} \quad \text{a coordinate and tetrad tensor .} \end{aligned} \quad (16.96)$$

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  $\delta A_\lambda = D_\lambda \theta$  of the electromagnetic potential  $A_\lambda$  by the gradient of a scalar  $\theta$  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

$$S_g = \frac{1}{8\pi} \int e^{mn\kappa\lambda} \left( \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) e^{-1} d^4x^{0123} . \quad (16.97)$$

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  $\Gamma_{mnk}$ , 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  $x^\mu$ , 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,  $d^4x = e^{-1}d^4x^{0123}$ , equation (15.86), 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  $\Gamma_{mn\lambda}$  yields

$$\delta S_g = \frac{1}{8\pi} \int \left[ e^{mn\kappa\lambda} \frac{\partial \delta \Gamma_{mn\lambda}}{\partial x^\kappa} + e^{mn\kappa\lambda} \delta(\Gamma_{m\lambda}^p \Gamma_{pn\kappa}) + \left( \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) e \delta(e^{-1}e^{mn\kappa\lambda}) \right] e^{-1} d^4x^{0123}. \quad (16.98)$$

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

$$e^{mn\kappa\lambda} \frac{\partial \delta \Gamma_{mn\lambda}}{\partial x^\kappa} = \frac{e \partial(e^{-1}e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda})}{\partial x^\kappa} - \frac{e \partial(e^{-1}e^{mn\kappa\lambda})}{\partial x^\kappa} \delta \Gamma_{mn\lambda}. \quad (16.99)$$

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 (2.73) (the  $\circ$  atop  $\mathring{D}_\kappa$  is a reminder that it is torsion-free),

$$\frac{e \partial(e^{-1}e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda})}{\partial x^\kappa} = \mathring{D}_\kappa (e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda}), \quad (16.100)$$

and therefore integrates to a surface term in accordance with Gauss' theorem (15.97). 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

$$e^{mn\kappa\lambda} \delta(\Gamma_{m\lambda}^p \Gamma_{pn\kappa}) = 2 e^{mn\kappa\lambda} \Gamma_{m\lambda}^p \delta \Gamma_{pn\kappa}. \quad (16.101)$$

The variation  $\delta \ln e$  of the vierbein determinant  $e$  may be written in terms of the variation  $\delta e^{m\kappa}$  of the vierbein, equation (2.76),

$$\delta \ln e = e_{m\kappa} \delta e^{m\kappa}. \quad (16.102)$$

The last term in the integrand on the right hand side of equation (16.98) is then

$$\left( \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) e \delta(e^{-1}e^{mn\kappa\lambda}) = \frac{1}{2} R_{\kappa\lambda mn} e \delta(e^{-1}e^{m\kappa} e^{n\lambda}) = (R_{\kappa m} - \frac{1}{2} e_{m\kappa} R) \delta e^{m\kappa} = G_{km} e^k{}_\kappa \delta e^{m\kappa}, \quad (16.103)$$

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

$$8\pi \delta S_g = \oint e^{mn\kappa\lambda} \delta \Gamma_{mn\lambda} d^3x_\kappa + \int \left[ \left( - \frac{e \partial(e^{-1}e^{mn\kappa\lambda})}{\partial x^\kappa} + 2 \Gamma_{p\kappa}^{[m} e^{n]p\kappa\lambda} \right) \delta \Gamma_{mn\lambda} + G_{\kappa m} \delta e^{m\kappa} \right] d^4x. \quad (16.104)$$

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

$$\frac{e \partial(e^{-1} e^{mn\kappa\lambda})}{\partial x^\kappa} = 2 \Gamma_{p\kappa}^{[m} e^{n]p\kappa\lambda} . \quad (16.106)$$

Abbreviate the left hand side of equation (16.106) by

$$f^{lmn} \equiv e^l \frac{e \partial(e^{-1} e^{mn\kappa\lambda})}{\partial x^\kappa} , \quad (16.107)$$

which is antisymmetric in its last two indices,  $f_{lmn} = f_{l[mn]}$ . In terms of the vierbein derivatives  $d_{lmn}$  defined by equation (11.32), 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  $\Gamma_{mnl}$  in terms of  $f_{lmn}$ ,

$$\Gamma_{mnl} = 2f_{lmn} - 3f_{[lmn]} + 2\gamma_{l[m} f^p_{n]p} . \quad (16.109)$$

Inserting the expression (16.108) into equations (16.109) yields the standard torsion-free expression (11.54) for the tetrad-frame connection  $\Gamma_{mnl}$  in terms of vierbein derivatives  $d_{mnl}$ ,

$$\Gamma_{mnl} = \mathring{\Gamma}_{mnl} = 2d_{l[mn]} - 3d_{[lmn]} . \quad (16.110)$$

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, §16.10.

## 16.9 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,

$$\delta S_m = - \int T_{\kappa m} \delta e^{m\kappa} d^4x . \quad (16.111)$$

Adding the variation (16.104) of the gravitational action and the variation (16.111) of the matter action gives

$$8\pi (\delta S_g + \delta S_m) = \int (G_{\kappa m} - 8\pi T_{\kappa m}) \delta e^{m\kappa} d^4x , \quad (16.112)$$

extremization of which implies Einstein's equations in the presence of matter

$$\boxed{G_{km} = 8\pi T_{km}} . \quad (16.113)$$

The Einstein equations (16.113) 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 §§16.10.1 and 16.10.2 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 in the scalar volume element  $d^4x \equiv e^{-1} d^4x^{0123}$ ,

$$T_{\kappa m} = -\frac{\delta L_m}{\delta e^{m\kappa}} + L_m e_{m\kappa} . \quad (16.114)$$

## 16.10 Spin angular-momentum

In the standard  $U_Y(1) \times SU(2) \times SU(3)$  model of physics, the connections associated with the gauge groups are dynamical fields, the gauge bosons, which include 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 (16.113) 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 (15.47). 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  $\Sigma^{\lambda mn}$ ,

$$\delta S_m = \frac{1}{2} \int \Sigma^{\lambda mn} \delta \Gamma_{mn\lambda} d^4x \quad (16.115)$$

(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 and Majorana fields, Exercises 16.4 and 16.5. The spin angular-momentum vanishes for gauge fields such as the electromagnetic field, Exercise 16.3. 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 (16.115) of the matter action gives

$$8\pi(\delta S_g + \delta S_m) = \int \left( -\frac{e \partial(e^{-1} e^{mn\kappa\lambda})}{\partial x^\kappa} + 2\Gamma_{p\kappa}^{[m} e^{n]p\kappa\lambda} + 4\pi \Sigma^{\lambda mn} \right) \delta \Gamma_{mn\lambda} d^4x , \quad (16.116)$$

extremization of which implies

$$\frac{e \partial(e^{-1} e^{mn\kappa\lambda})}{\partial x^\kappa} = 2\Gamma_{p\kappa}^{[m} e^{n]p\kappa\lambda} + 4\pi \Sigma^{\lambda mn} . \quad (16.117)$$

Inverting equation (16.117) 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  $K_{mn\lambda}$ ,

$$\Gamma_{mn\lambda} = \mathring{\Gamma}_{mn\lambda} + K_{mn\lambda} , \quad (16.118)$$

with the contortion tensor  $K_{mn\lambda}$  being related to the spin angular-momentum  $\Sigma_{lmn}$  by

$$K_{mn\lambda} = 8\pi \left( -\Sigma_{lmn} + \frac{3}{2}\Sigma_{[lmn]} - \gamma_{l[m} \Sigma^p_{n]p} \right) . \quad (16.119)$$

The contortion  $K_{mn\lambda}$  is related to the torsion  $S_{mn\lambda}$  by equations (11.56). Equation (16.119) implies that the torsion  $S_{\lambda mn}$  is related to the spin angular-momentum  $\Sigma_{\lambda mn}$  by

$$S_{mn\lambda} = 8\pi \left( \Sigma_{mn}^\lambda + e_{[m}^\lambda \Sigma_{n]k}^k \right) . \quad (16.120)$$

Equation (16.120) inverts to

$$\boxed{S_{mn}^\lambda + 2e_{[m}^\lambda S_{n]k}^k = 8\pi \Sigma_{mn}^\lambda} . \quad (16.121)$$

Equation (16.121) relating the torsion to the spin angular-momentum is the analogue of Einstein's equations (16.113) relating the Einstein tensor to the matter energy-momentum. Whereas the Einstein equations (16.113) 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 (16.121) 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  $\Gamma_{mn\lambda}$ ,

$$\delta S_m = \frac{1}{2} \int \tilde{\Sigma}^{\lambda mn} \delta \pi_{mn\lambda} d^4x . \quad (16.122)$$

The relation between the torsion  $S_{mn}^\lambda$  and the modified spin angular-momentum  $\tilde{\Sigma}_{mn}^\lambda$  is

$$S_{mn}^\lambda = 8\pi \tilde{\Sigma}_{mn}^\lambda . \quad (16.123)$$

Comparing equation (16.123) to equation (16.120) shows that the modified and original spin angular-momenta  $\tilde{\Sigma}_{mn}^{\lambda}$  and  $\Sigma_{mn}^{\lambda}$  differ by a trace term,

$$\tilde{\Sigma}_{mn}^{\lambda} = \Sigma_{mn}^{\lambda} + e_{[m}{}^{\lambda} \Sigma_{n]k}^k, \quad \Sigma_{mn}^{\lambda} = \tilde{\Sigma}_{mn}^{\lambda} + 2 e_{[m}{}^{\lambda} \tilde{\Sigma}_{n]k}^k. \quad (16.124)$$

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  $\Sigma_{lmn}$  package these parts with different weights. The three parts are related by

$$\begin{aligned} S_{np}^p &= K_{np}^p = -4\pi\Sigma_{np}^p && \text{trace part ,} \\ S_{[lmn]} &= 2K_{[mln]} = 8\pi\Sigma_{[lmn]} && \text{totally antisymmetric part ,} \\ S_{lmn} &= -K_{mln} = 8\pi\Sigma_{lmn} && \text{trace-free, antisymmetry-free part .} \end{aligned} \quad (16.125)$$

### 16.10.1 Conservation of spin 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 (16.130) for the spin angular-momentum  $\Sigma^{\lambda mn}$  of the field. If torsion vanishes, the conservation law (16.130) implies that the energy-momentum tensor  $T^{mn}$  of the field is symmetric, equation (16.131).

Equation (16.95) gives the variation of a tetrad vector under a local Lorentz transformation generated by the infinitesimal antisymmetric tensor  $\epsilon_{mn}$ . Under such an infinitesimal Lorentz transformation, the vierbein tensor  $e^{m\kappa}$  varies as

$$\delta e^{m\kappa} = \epsilon^m{}_n e^{n\kappa} = \epsilon^{m\kappa}. \quad (16.126)$$

Equation (16.96) gives the variation of the Lorentz connection under an infinitesimal Lorentz transformation generated by  $\epsilon_{mn}$ ,

$$\delta\Gamma_{mn\lambda} = D_\lambda \epsilon_{mn}. \quad (16.127)$$

The coefficients of the variation  $\delta S_m$  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 (16.115) and (16.111). Inserting the variations (16.126) and (16.127) with respect to Lorentz transformations yields the variation of the matter action under a Lorentz transformation,

$$\delta S_m = \int \left( \frac{1}{2} \Sigma^{\lambda mn} D_\lambda \epsilon_{mn} - T_{\kappa m} \epsilon^{m\kappa} \right) d^4x. \quad (16.128)$$

An integration by parts brings the variation to

$$\delta S_m = \oint \frac{1}{2} \Sigma^{\lambda mn} \epsilon_{mn} d^3x^\lambda - \int \left( \frac{1}{2} D_\lambda \Sigma^{\lambda mn} + T^{[mn]} \right) \epsilon_{mn} d^4x. \quad (16.129)$$

Requiring that the matter action be invariant under Lorentz transformations imposes that the variation (16.129) must vanish under arbitrary variations of the antisymmetric Lorentz generators  $\epsilon_{mn}$ , subject to the generators being fixed on the initial and final hypersurfaces of integration. Therefore the integrand of the rightmost integral in equation (16.129) must vanish, implying the conservation law

$$\boxed{\frac{1}{2}D_{\lambda}\Sigma^{\lambda mn} + T^{[mn]} = 0} . \quad (16.130)$$

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,

$$T^{mn} = T^{nm} . \quad (16.131)$$

### 16.10.2 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 (16.139) for the energy-momentum  $T^{mn}$  of the field.

Under a coordinate transformation generated by the coordinate shift  $\delta x^\mu = \epsilon^\mu$ , the variation of any quantity is given by minus its Lie derivative with respect to the coordinate shift  $\epsilon^\mu$ , equation (7.108). The Lie derivative of a coordinate tensor is given by equation (7.131), 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 25.2). 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  $\mathcal{L}_\epsilon$  of the vierbein is

$$\begin{aligned} \mathcal{L}_\epsilon e^{m\kappa} &= -e^{m\lambda} \frac{\partial \epsilon^\kappa}{\partial x^\lambda} + \epsilon^\lambda \left( \frac{\partial e^{m\kappa}}{\partial x^\lambda} + \Gamma_{n\lambda}^m e^{n\kappa} \right) \\ &= -\mathring{D}^m \epsilon^\kappa - \epsilon_\lambda K^{m\kappa\lambda} \\ &= -D^m \epsilon^\kappa - \epsilon_l S^{\kappa ml} , \end{aligned} \quad (16.132)$$

which differs from equation (25.18) 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 (16.132) 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 (25.18),

$$\delta e^{m\kappa} = -\mathcal{L}_\epsilon e^{m\kappa} = \mathring{D}^m \epsilon^\kappa + \epsilon_\lambda K^{m\kappa\lambda} . \quad (16.133)$$

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,

$$\delta D_\lambda a_n - D_\lambda \delta a_n = \delta(\partial_\lambda a_n - \Gamma_{n\lambda}^m a_m) - (\partial_\lambda \delta a_n - \Gamma_{n\lambda}^m \delta a_m) = -(\delta\Gamma_{n\lambda}^m) a_m . \quad (16.134)$$

Thus the variation of the Lorentz connection under a coordinate transformation by  $\epsilon^\kappa$  satisfies

$$\begin{aligned} (\delta\Gamma_{mn\lambda}) a^m &= \mathcal{L}_{\epsilon}(D_\lambda a_n) - D_\lambda \mathcal{L}_\epsilon a_n = \epsilon^\kappa \left( \frac{\partial}{\partial x^\kappa} D_\lambda a_n - \Gamma_{n\kappa}^m D_\lambda a_m \right) + (D_\kappa a_n) \frac{\partial \epsilon^\kappa}{\partial x^\lambda} - D_\lambda (\epsilon^\kappa D_\kappa a_n) \\ &= \epsilon^\kappa a^m R_{\lambda\kappa mn} . \end{aligned} \quad (16.135)$$

Equation (16.136) is true for arbitrary  $a^m$ , so

$$\delta\Gamma_{mn\lambda} = \epsilon^\kappa R_{\lambda\kappa mn} . \quad (16.136)$$

Inserting the variations (16.133) and (16.136) of the vierbein and Lorentz connection into the variations (16.111) and (16.115) of the matter action yields the variation of the matter action under a coordinate transformation by  $\epsilon^\kappa$ ,

$$\delta S_m = \int \left[ -T_{\kappa m} \left( \mathring{D}^m \epsilon^\kappa + \epsilon_\lambda K^{m\kappa\lambda} \right) + \frac{1}{2} \Sigma^{\lambda mn} \epsilon^\kappa R_{\lambda\kappa mn} \right] d^4x . \quad (16.137)$$

An integration by parts brings the variation of the matter action to

$$\delta S_m = - \oint T_{\kappa\lambda} \epsilon^\kappa d^3x^\lambda + \int \left( \mathring{D}^m T_{\kappa m} + T^{mn} K_{mn\kappa} + \frac{1}{2} \Sigma^{\lambda mn} R_{\lambda\kappa mn} \right) \epsilon^\kappa d^4x . \quad (16.138)$$

Invariance of the action under coordinate transformations requires that the variation (16.138) vanish for arbitrary coordinate shifts  $\epsilon^\kappa$  that vanish on the boundary. Therefore the integrand of rightmost integral in equation (16.138) must vanish, implying the law of conservation of matter energy-momentum,

$$\boxed{\mathring{D}^m T_{\kappa m} + T^{mn} K_{mn\kappa} + \frac{1}{2} \Sigma^{\lambda mn} R_{\lambda\kappa mn} = 0} . \quad (16.139)$$

Since the contortion  $K_{mn\kappa}$  is antisymmetric in its first two indices  $mn$ , the second term of the conservation law (16.279) depends on the antisymmetric part  $T^{[mn]}$  of the energy-momentum tensor.

If the spin angular-momentum of the matter component vanishes,  $\Sigma^{\lambda mn} = 0$ , then its energy-momentum tensor is symmetric, equation (16.130), and the energy-momentum conservation equation (16.279) of the matter component simplifies to

$$\mathring{D}_m T^{nm} = 0 . \quad (16.140)$$

**Concept question 16.1 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 2.4, 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 16.2 Why the names energy-momentum and spin angular-momentum?** What is the justification for calling  $T_{\kappa m}$  the 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 includes Lorentz boosts as well as spatial rotations). The energy-momentum tensor  $T_{\kappa m}$  satisfies a conservation law (16.139) that arises as a result of symmetry under coordinate transformations. The spin angular-momentum tensor  $\Sigma^{\lambda mn}$  satisfies a conservation law (16.130) that arises as a result of symmetry under local Lorentz transformations. The reason for the adjective “spin” is that, as seen in Exercises 16.3–16.5, spin angular-momentum vanishes for bosonic fields such as electromagnetism, but is non-vanishing for fermionic (half-integral spin) fields.

### Exercise 16.3 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 (16.111) and (16.115).

**Solution.**

1. **Energy-momentum of the electromagnetic field.** The Lagrangian of the electromagnetic field is, equation (16.28),

$$L \equiv -\frac{1}{16\pi} g^{\kappa\mu} g^{\lambda\nu} F_{\kappa\lambda} F_{\mu\nu}, \quad (16.141)$$

where the inverse metric is in terms of the vierbein,

$$g^{\kappa\mu} = \eta_{km} e^{k\alpha} e^{m\mu}. \quad (16.142)$$

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 (16.141) with respect to the vierbein is

$$\delta L = -\frac{1}{4\pi} F_{\kappa\lambda} F_{\kappa}^{\lambda} \delta e^{k\alpha}. \quad (16.143)$$

An additional contribution to the energy-momentum comes from variation of the vierbein determinant in the volume element, equation (16.114). The resulting tetrad-frame energy-momentum tensor  $T_{kl}$  of the electromagnetic field is the symmetric tensor

$$T_{kl} = \frac{1}{4\pi} \left( F_{km} F_l^m - \frac{1}{4} \gamma_{kl} F_{mn} F^{mn} \right). \quad (16.144)$$

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,

$$\Sigma_{lmn} = 0. \quad (16.145)$$

**Exercise 16.4 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 (38.3),

$$L = \frac{1}{2}\bar{\psi} \cdot (e^{k\lambda} \boldsymbol{\gamma}_k \mathbf{D}_\lambda + m) \psi - \frac{1}{2}\psi \cdot (e^{k\lambda} \boldsymbol{\gamma}_k \mathbf{D}_\lambda + m) \bar{\psi}, \quad (16.146)$$

where the covariant derivative is  $\mathbf{D}_\lambda = \partial_\lambda + \frac{1}{4}\Gamma_{mn\lambda} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n$  (implicit sum over both indices  $m$  and  $n$ ). 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  $lk$ ,

$$T_{lk} = \frac{1}{2}\bar{\psi} \cdot \boldsymbol{\gamma}_k \mathbf{D}_l \psi - \frac{1}{2}\psi \cdot \boldsymbol{\gamma}_k \mathbf{D}_l \bar{\psi}. \quad (16.147)$$

The Dirac Lagrangian  $L$  vanishes on the equations of motion, so the contribution to the energy-momentum, equation (16.114), arising from variation of the vierbein determinant in the scalar volume element  $d^4x \equiv e^{-1}d^4x^{0123}$  vanishes.

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  $lmn$ ,

$$\Sigma_{lmn} = \frac{1}{2}\bar{\psi} \cdot \boldsymbol{\gamma}_l \wedge \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n \psi. \quad (16.148)$$

The possible vector current contribution cancels between the two terms on the right hand side of equation (16.146).

**Exercise 16.5 Energy-momentum and spin angular-momentum of a Majorana field.** Find the energy-momentum and spin angular-momentum of a Majorana spinor field.

**Solution.** The Lagrangian of a Majorana field is given by equation (38.70). The energy-momentum and spin angular-momentum are given by the same expressions (16.147) and (16.148) as for the Dirac field, but with  $\bar{\psi}$  replaced by  $I\bar{\psi}$ . Like the Dirac field, the energy-momentum tensor is *not* symmetric, while the spin angular-momentum tensor is totally antisymmetric.

**Exercise 16.6 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 16.7 Dirac (or Majorana) spinor field in the presence of torsion.** How does torsion affect the propagation of a massive Dirac (or Majorana) 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  $\Sigma_{mnk}$  is uniform.

**Solution.** Dirac and Majorana spinors satisfy the same equation of motion, so the arguments are the same for both spinor types. 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 or Majorana) spinor field of rest mass  $m$  is

$$[\boldsymbol{\gamma}^k(\partial_k + \frac{1}{4}K_{mnk}\boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n) + m]\psi = 0. \quad (16.149)$$

For simplicity, go to the rest frame of the spinor field, where the particle is in a time-up and spin-up eigenstate  $\psi \propto \gamma_{\uparrow\uparrow}$ , equation (14.97), which means that the particle is a particle, not an antiparticle, and its spin is along the positive 3-direction. The only Dirac  $\gamma$ -matrices that are non-vanishing when acting on a spinor  $\psi$  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

$$(\partial_0 + \frac{3}{2}K_{[012]} + K_{0a}^a + m)\psi = 0. \quad (16.150)$$

The solutions are

$$\psi \propto e^{-i(m+\delta m)t}, \quad (16.151)$$

where the mass change  $\delta m$  is

$$\delta m = \frac{3}{2}K_{[012]} + K_{0a}^a = 4\pi G (\Sigma_{[012]} - \Sigma_{0a}^a), \quad (16.152)$$

the contortion being related to the spin angular-momentum  $\Sigma_{mnk}$  by equations (16.125). 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 Exercises 16.4 and 16.5, a Dirac or Majorana 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.

---

## 16.11 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  $\delta\Gamma_{mn\lambda}$ ),

$$e^{mn\kappa\lambda} \frac{\partial\Gamma_{mn\lambda}}{\partial x^\kappa} = \frac{e\partial(e^{-1}e^{mn\kappa\lambda}\Gamma_{mn\lambda})}{\partial x^\kappa} - \frac{e\partial(e^{-1}e^{mn\kappa\lambda})}{\partial x^\kappa} \Gamma_{mn\lambda}. \quad (16.153)$$

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 (16.153) is a torsion-free covariant divergence  $\mathring{D}_\kappa\delta(e^{mn\kappa\lambda}\Gamma_{mn\lambda})$ , which can be discarded from the Lagrangian without changing the equations of motion. The resulting alternative gravitational Lagrangian is

$$8\pi L'_g = -\frac{e\partial(e^{-1}e^{mn\kappa\lambda})}{\partial x^\kappa}\Gamma_{mn\lambda} + e^{mn\kappa\lambda}\Gamma_{m\lambda}^p\Gamma_{pn\kappa}. \quad (16.154)$$

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 (16.154), the coordinates are the vierbein  $e^{n\lambda}$ , and the corresponding canonically conjugate momenta are

$$\pi_n{}^\kappa{}_\lambda \equiv \frac{8\pi\delta L'_g}{\partial(\partial e^{n\lambda}/\partial x^\kappa)} = e^{m\kappa}\pi_{nm\lambda}, \quad (16.155)$$

where  $\pi_{nm\lambda}$  and  $\Gamma_{nm\lambda}$  are related by

$$\pi_{nm\lambda} \equiv \Gamma_{nm\lambda} - e_{n\lambda}\Gamma_{mp}^p + e_{m\lambda}\Gamma_{np}^p, \quad \Gamma_{nm\lambda} = \pi_{nm\lambda} - \frac{1}{2}e_{n\lambda}\pi_{mp}^p + \frac{1}{2}e_{m\lambda}\pi_{np}^p. \quad (16.156)$$

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 (16.154) is in Hamiltonian form  $L'_g = p^\kappa\partial_\kappa q - H_g$  with coordinates  $q = e^{n\lambda}$  and momenta  $p^\kappa = \pi_n{}^\kappa{}_\lambda/8\pi$ ,

$$L'_g = \frac{1}{8\pi}\pi_n{}^\kappa{}_\lambda\frac{\partial e^{n\lambda}}{\partial x^\kappa} - H_g, \quad (16.157)$$

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,

$$8\pi\delta S'_g = \oint\pi_{nm\lambda}\delta e^{n\lambda}d^3x^m + \int \text{as eq. (16.104)}. \quad (16.158)$$

The surface term vanishes provided that the vierbein  $e^{n\lambda}$  is held fixed on the boundary.

## 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 §16.11. In the Lagrangian approach, the alternative Lagrangian  $L'_g$  given by equation (16.154) provides a satisfactory Lagrangian, once the connections  $\Gamma_{mn\lambda}$  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 (16.154) 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  $\mathring{\Gamma}_{pn\kappa}^m$ ) brings the alternative Lagrangian (16.154) to

$$8\pi L'_g = e^{mn\kappa\lambda} \left( -2\mathring{\Gamma}_{m\lambda}^p \Gamma_{pn\kappa} + \Gamma_{m\lambda}^p \Gamma_{pn\kappa} \right) = e^{mn\kappa\lambda} \left( -\mathring{\Gamma}_{m\lambda}^p \mathring{\Gamma}_{pn\kappa} + K_{m\lambda}^p K_{pn\kappa} \right) , \quad (16.159)$$

the last step of which follows from expanding the torsion-full connection as a sum of the torsion-free connection and the contortion tensor,  $\Gamma_{mn\kappa} = \mathring{\Gamma}_{mn\kappa} + K_{mn\kappa}$ , equation (11.55). The torsion-free connections  $\mathring{\Gamma}_{pn\kappa} \equiv e^k{}_\kappa \mathring{\Gamma}_{pnk}$  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 (16.159) 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

$$8\pi L'_g = -e^{mn\kappa\lambda} \Gamma_{m\lambda}^p \Gamma_{pn\kappa} . \quad (16.160)$$

Thus, for vanishing torsion, the first (“surface”) term in the original alternative Lagrangian (16.154) 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 (16.159) 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

$$R = 2\mathring{D}_\kappa (e^{mn\kappa\lambda} \Gamma_{mn\lambda}) + 2e^{mn\kappa\lambda} (-\mathring{\Gamma}_{m\lambda}^p \mathring{\Gamma}_{pn\kappa} + K_{m\lambda}^p K_{pn\kappa}) . \quad (16.161)$$

Discarding the torsion-free covariant divergence recovers the quadratic gravitational Lagrangian (16.159). 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

$$L_g \equiv \frac{1}{16\pi} (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \mathbf{R}_{\kappa\lambda} = \frac{1}{16\pi} (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \left( \frac{\partial \mathbf{\Gamma}_\lambda}{\partial x^\kappa} - \frac{\partial \mathbf{\Gamma}_\kappa}{\partial x^\lambda} + \frac{1}{2} [\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] \right), \quad (16.162)$$

implicitly summed over both indices  $\kappa$  and  $\lambda$ . In equation (16.162),  $\mathbf{e}^\kappa = e^{m\kappa} \gamma_m$  are the usual coordinate (co)tangent vectors, equation (11.6), and the bivectors  $\mathbf{\Gamma}_\kappa$  and  $\mathbf{R}_{\kappa\lambda}$  are given by equations (15.20) and (15.25). The dot in equation (16.162) signifies the multivector dot product, equation (13.33), which here is a scalar product of bivectors. The order of  $\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa$  is flipped to cancel a minus sign from taking a scalar product of bivectors, equation (13.36).

Applying the multivector triple-product relation (13.37) to the derivative term in the rightmost expression of equation (16.162) brings the Hilbert Lagrangian to

$$L_g = \frac{1}{16\pi} (2 \mathbf{e}^\lambda \cdot (\partial \cdot \mathbf{\Gamma}_\lambda) + \frac{1}{2} (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot [\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda]), \quad (16.163)$$

where  $\partial \equiv \mathbf{e}^\kappa \partial / \partial x^\kappa$ . The form of the Lagrangian (16.163) indicates that the “velocities” corresponding to the “coordinates”  $\mathbf{\Gamma}_\lambda$  are  $\partial \cdot \mathbf{\Gamma}_\lambda$ . The Lagrangian (16.163) is in (super-)Hamiltonian form with bivector coordinates  $\mathbf{\Gamma}_\lambda$ , vector velocities  $\partial \cdot \mathbf{\Gamma}_\lambda$ , and vector momenta  $\mathbf{e}^\lambda / 8\pi$ ,

$$L_g = \frac{1}{8\pi} \mathbf{e}^\lambda \cdot (\partial \cdot \mathbf{\Gamma}_\lambda) - H_g, \quad (16.164)$$

and (super-)Hamiltonian  $H_g(\mathbf{\Gamma}_\lambda, \mathbf{e}^\lambda)$  (compare (16.93))

$$H_g = -\frac{1}{32\pi} (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot [\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda]. \quad (16.165)$$

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  $\mathbf{\Gamma}_\lambda$  and momenta  $\mathbf{e}^\lambda$  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  $\mathbf{e}^\kappa$ . 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  $\boldsymbol{\gamma}_k$  being considered fixed. Thus the variation  $\delta\mathbf{\Gamma}_\lambda$  of the Lorentz connections is

$$\delta\mathbf{\Gamma}_\lambda \equiv \frac{1}{2}(\delta\Gamma_{kl\lambda}) \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l \quad (16.166)$$

(implicitly summed over all indices; the factor of  $\frac{1}{2}$  would disappear if the sum were over distinct antisymmetric indices  $kl$ ). The variation  $\delta\mathbf{e}^\kappa$  of the coordinate vectors is

$$\delta\mathbf{e}^\kappa \equiv (\delta e_k{}^\kappa) \boldsymbol{\gamma}^k . \quad (16.167)$$

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  $d^4x = e^{-1}d^4x^{0123}$  must be taken into account. The variation of the vierbein determinant is related to the variation  $\delta\mathbf{e}^\kappa$  of the coordinate vectors by

$$\delta \ln e = e^k{}_\kappa \delta e_k{}^\kappa = \mathbf{e}_\kappa \cdot \delta\mathbf{e}^\kappa . \quad (16.168)$$

The variation of the gravitational action with the multivector Lagrangian (16.169) is

$$\delta S_g = \frac{1}{16\pi} \int \left[ (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \left( 2 \frac{\partial \delta\mathbf{\Gamma}_\lambda}{\partial x^\kappa} + \frac{1}{2} \delta[\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] \right) + e \delta(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \mathbf{R}_{\kappa\lambda} \right] d^4x . \quad (16.169)$$

The first term in the integrand of equation (16.169) integrates by parts to

$$(\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \frac{\partial \delta\mathbf{\Gamma}_\lambda}{\partial x^\kappa} = \mathring{D}_\kappa ((\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \delta\mathbf{\Gamma}_\lambda) - \frac{e \partial(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa)}{\partial x^\kappa} \cdot \delta\mathbf{\Gamma}_\lambda . \quad (16.170)$$

The second term in the integrand of equation (16.169) is

$$\frac{1}{2}(\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \delta[\mathbf{\Gamma}_\kappa, \mathbf{\Gamma}_\lambda] = (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot [\mathbf{\Gamma}_\kappa, \delta\mathbf{\Gamma}_\lambda] = [\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa, \mathbf{\Gamma}_\kappa] \cdot \delta\mathbf{\Gamma}_\lambda , \quad (16.171)$$

the last step of which follows from the multivector triple-product relation (13.37) 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 (16.169) is

$$\begin{aligned} \mathbf{R}_{\kappa\lambda} \cdot e \delta(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) &= 2 \mathbf{R}_{\kappa\lambda} \cdot (\mathbf{e}^\lambda \wedge \delta\mathbf{e}^\kappa) - (\mathbf{R}_{\kappa\lambda} \cdot (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa)) \mathbf{e}_\mu \cdot \delta\mathbf{e}^\mu \\ &= (2 \mathbf{R}_{\kappa\lambda} \cdot \mathbf{e}^\lambda - R \mathbf{e}_\kappa) \cdot \delta\mathbf{e}^\kappa \\ &= 2 \mathbf{G}_\kappa \cdot \delta\mathbf{e}^\kappa , \end{aligned} \quad (16.172)$$

where the second line again follows from the multivector triple-product relation (13.37), and  $\mathbf{G}_\kappa$  is the Einstein vector

$$\mathbf{G}_\kappa \equiv \mathbf{R}_{\kappa\lambda} \cdot \mathbf{e}^\lambda - \frac{1}{2} R \mathbf{e}_\kappa = (R_{\kappa m} - \frac{1}{2} R e_{m\kappa}) \boldsymbol{\gamma}^m . \quad (16.173)$$

The manipulations (16.170)–(16.172) bring the variation (16.169) of the action to

$$\delta S_g = \frac{1}{8\pi} \oint (\mathbf{e}_\lambda \wedge \mathbf{e}_\kappa) \cdot \delta \Gamma^\lambda d^3x^\kappa + \frac{1}{8\pi} \int \left[ \left( -\frac{e \partial(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa)}{\partial x^\kappa} + \frac{1}{2} [\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa, \Gamma_\kappa] \right) \cdot \delta \Gamma^\lambda + \mathbf{G}_\kappa \cdot \delta \mathbf{e}^\kappa \right] d^4x . \quad (16.174)$$

The surface term vanishes provided that  $\Gamma^\lambda$  is held fixed on the boundaries of integration. Extremizing the action (16.174) with respect to the variation  $\delta \mathbf{e}^\kappa$  of coordinate vectors yields the Einstein equations in vacuo,

$$\mathbf{G}_\kappa = 0 . \quad (16.175)$$

Extremizing the action (16.174) with respect to the variation  $\delta \Gamma_\lambda$  of the Lorentz connections yields the multivector equivalent of equation (16.106),

$$\frac{e \partial(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa)}{\partial x^\kappa} = \frac{1}{2} [\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa, \Gamma_\kappa] . \quad (16.176)$$

The left hand side of equation (16.176) is

$$\frac{e \partial(e^{-1} \mathbf{e}^\lambda \wedge \mathbf{e}^\kappa)}{\partial x^\kappa} = \partial \wedge \mathbf{e}^\lambda - \mathbf{e}^\lambda \wedge (\mathbf{e}_\mu \cdot (\partial \wedge \mathbf{e}^\mu)) . \quad (16.177)$$

The “velocities” of the coordinate vectors are their curls  $\partial \wedge \mathbf{e}^\lambda$ ,

$$\partial \wedge \mathbf{e}^\lambda \equiv \mathbf{e}^\kappa \wedge \frac{\partial \mathbf{e}^\lambda}{\partial x^\kappa} = d_{[mn]}^\lambda \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n \quad (16.178)$$

implicitly summed over both indices  $m$  and  $n$ . The  $d_{mn}^\lambda \equiv e^{l\lambda} d_{lmn}$  in equation (16.178) are the vierbein derivatives defined by equation (11.32). Equation (16.176) solves to yield the torsion-free relation between the connections  $\Gamma^\lambda$  and the velocities  $\partial \wedge \mathbf{e}^\lambda$  of the coordinate vectors,

$$\Gamma^\lambda = \mathring{\Gamma}^\lambda = \partial \wedge \mathbf{e}^\lambda - \mathbf{e}^\lambda \cdot (\mathbf{e}_\mu \wedge (\partial \wedge \mathbf{e}^\mu)) , \quad \partial \wedge \mathbf{e}^\lambda = \mathring{\Gamma}^\lambda - 2 \mathbf{e}^\lambda \cdot (\mathbf{e}_\mu \wedge \mathring{\Gamma}^\mu) . \quad (16.179)$$

### 16.13.3 Alternative multivector gravitational action

As in §16.11, since the Lagrangian (16.162) is in Hamiltonian form, the coordinates  $\Gamma_\lambda$  and momenta  $\mathbf{e}^\lambda$  can be traded without changing the equations of motion. Integrating the Lagrangian (16.162) by parts gives

$$8\pi L_g = \frac{e \partial(e^{-1}(\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \Gamma_\lambda)}{\partial x^\kappa} - \frac{e \partial(e^{-1}(\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \Gamma_\lambda)}{\partial x^\kappa} + \frac{1}{4} (\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot [\Gamma_\kappa, \Gamma_\lambda] . \quad (16.180)$$

As in §16.11, the connection  $\Gamma_\lambda$  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 (16.180) is a covariant divergence  $\mathring{D}_\kappa \delta ((\mathbf{e}^\lambda \wedge \mathbf{e}^\kappa) \cdot \Gamma_\lambda)$ , which can be discarded from the Lagrangian without changing the equations of motion.

The middle term on the right hand side of equation (16.180) can be written

$$-\mathbf{\Gamma}_\lambda \cdot \frac{e \partial(e^{-1} e^\lambda \wedge e^\kappa)}{\partial x^\kappa} = \boldsymbol{\pi}_\lambda \cdot (\boldsymbol{\partial} \wedge e^\lambda), \quad (16.181)$$

where  $\boldsymbol{\partial} \wedge e^\lambda$  is given by equation (16.178), and  $\boldsymbol{\pi}_\lambda$  is the trace-modified Lorentz connection bivector

$$\boldsymbol{\pi}_\lambda = \mathbf{\Gamma}_\lambda - \mathbf{e}_\lambda \wedge (e^\mu \cdot \mathbf{\Gamma}_\mu), \quad \mathbf{\Gamma}_\lambda = \boldsymbol{\pi}_\lambda - \frac{1}{2} \mathbf{e}_\lambda \wedge (e^\mu \cdot \boldsymbol{\pi}_\mu), \quad (16.182)$$

with components

$$\boldsymbol{\pi}_\lambda = \frac{1}{2} \pi_{mn\lambda} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n. \quad (16.183)$$

The components  $\pi_{mn\lambda}$  are as given by equation (16.156).

Discarding the torsion-free divergence from the Lagrangian (16.180) yields the alternative Lagrangian

$$L'_g = \frac{1}{8\pi} \boldsymbol{\pi}_\lambda \cdot (\boldsymbol{\partial} \wedge e^\lambda) - H_g, \quad (16.184)$$

with the same (super-)Hamiltonian (16.165) as before. The alternative Lagrangian (16.184) is in Hamiltonian form with coordinates  $e^\lambda$ , velocities  $\boldsymbol{\partial} \wedge e^\lambda$ , and corresponding canonically conjugate momenta  $\boldsymbol{\pi}_\lambda/(8\pi)$ . As with the alternative Lagrangian (16.154) in index notation, the alternative Lagrangian (16.184) 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 (16.184) 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  $\delta e^\kappa$  of the coordinate vectors. The variation defines the matter energy-momentum vector  $\mathbf{T}_\kappa$ ,

$$\delta S_m = - \int \mathbf{T}_\kappa \cdot \delta e^\kappa d^4x, \quad (16.185)$$

with

$$\mathbf{T}_\kappa = T_{\kappa m} \boldsymbol{\gamma}^m. \quad (16.186)$$

The combined variation of the gravitational and matter actions with respect to  $\delta e^\kappa$  is

$$8\pi(\delta S_g + \delta S_m) = \int (\mathbf{G}_\kappa - 8\pi \mathbf{T}_\kappa) \cdot \delta e^\kappa d^4x, \quad (16.187)$$

extremization of which yields Einstein's equations with matter,

$$\mathbf{G}_\kappa = 8\pi \mathbf{T}_\kappa. \quad (16.188)$$

### 16.13.5 Spin angular-momentum in multivector notation

Just as the variation of the matter action with respect to the the variation  $\delta e^\kappa$  of the coordinate vectors defines the matter energy-momentum vector  $T_\kappa$ , so also the variation of the matter action with respect to the the variation  $\delta \Gamma_\lambda$  of the Lorentz connection bivectors defines the spin angular-momentum bivector  $\Sigma^\lambda$ ,

$$\delta S_m = \int \Sigma^\lambda \cdot \delta \Gamma_\lambda d^4x , \quad (16.189)$$

with (the minus sign is introduced for the same reason as the minus in equation (15.27))

$$\Sigma^\lambda \equiv -\frac{1}{2} \Sigma_{mn}^\lambda \gamma^m \wedge \gamma^n , \quad (16.190)$$

implicitly summed over both indices  $m$  and  $n$ . As in §16.10, the usual expression (15.44) for the torsion-full tetrad connections  $\Gamma_\lambda$  as a sum of the torsion-free connection  $\tilde{\Gamma}_\lambda$  and the contortion  $K_\lambda$  is recovered,

$$\Gamma_\lambda = \tilde{\Gamma}_\lambda + K_\lambda , \quad (16.191)$$

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  $\Sigma^\lambda$  by

$$S^\lambda = 8\pi (\Sigma^\lambda - \frac{1}{2} e^\lambda \wedge (e_\mu \cdot \Sigma^\mu)) , \quad S^\lambda - e^\lambda \wedge (e_\mu \cdot S^\mu) = 8\pi \Sigma^\lambda . \quad (16.192)$$

## 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  $\gamma_m$ , is kept distinct from the symmetry with respect to coordinate transformations, encoded in the tangent vectors  $e_\mu$ . 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 product<sup>1</sup>. Similarly, it is convenient to reserve the Hodge duality symbol  $*$  to signify the dual of a form, equation (15.81), not the dual of a multivector, and to write  $Ia$  for the Hodge dual of a multivector  $a$ , equation (13.22). The form dual of a  $p$ -form  $a = a_\Lambda d^p x^\Lambda$

<sup>1</sup> 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 (16.279).

with multivector coefficients  $\mathbf{a}_\Lambda$  is the multivector  $q$ -form  ${}^*\mathbf{a}$  given by (this is equation (15.81) generalized to allow multivector coefficients)

$${}^*\mathbf{a} \equiv \mathbf{a}_\Lambda {}^*d^q x^\Lambda = (-)^{[p/2]} \varepsilon_{\Lambda\Pi} \mathbf{a}^\Lambda d^q x^\Pi , \quad (16.193)$$

implicitly summed over distinct sequences  $\Lambda$  and  $\Pi$  of respectively  $p$  and  $q \equiv N - p$  (in  $N$  dimensional spacetime) coordinate indices. The dual (16.193) 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  $^{**}$ ,

$${}^{**}\mathbf{a} \equiv I {}^*\mathbf{a} . \quad (16.194)$$

In this section §16.14 and in the remainder of this chapter, implicit sums are over distinct antisymmetric sequences of indices, since this removes the ubiquitous factorial factors that otherwise appear.

### Exercise 16.8 Commutation of multivector forms.

1. Argue that if  $\mathbf{a} \equiv \mathbf{a}_{AB} \gamma^A d^p x^B$  is a multivector form of grade  $k$  and form index  $p$ , and  $\mathbf{b} \equiv \mathbf{a}_{AB} \gamma^A d^q x^B$  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

$$\langle \mathbf{ab} \rangle_{k+l-2n} = (-)^{kl-n^2+pq} \langle \mathbf{ba} \rangle_{k+l-2n} . \quad (16.195)$$

As particular cases of equation (16.195), conclude that

$$\mathbf{a} \cdot \mathbf{a} = 0 \quad p \text{ odd} , \quad (16.196a)$$

$$\mathbf{a} \wedge \mathbf{a} = 0 \quad k + p \text{ odd} . \quad (16.196b)$$

2. What is the form index of the product  $\mathbf{ab}$  of multivector forms  $\mathbf{a}$  and  $\mathbf{b}$  of form index  $p$  and  $q$ ?  
 3. Argue 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,

$$[\mathbf{a}, \mathbf{b}] = \begin{cases} [\mathbf{b}, \mathbf{a}] & p \text{ and } q \text{ odd} , \\ -[\mathbf{b}, \mathbf{a}] & \text{otherwise} . \end{cases} \quad (16.197)$$

As a corollary, conclude that the (anti-)commutator of a  $p$ -form  $\mathbf{a}$  with itself vanishes if  $p$  is (odd) even,

$$\{\mathbf{a}, \mathbf{a}\} = 0 \quad p \text{ odd} , \quad (16.198a)$$

$$[\mathbf{a}, \mathbf{a}] = 0 \quad p \text{ even} . \quad (16.198b)$$

### Solution.

1. This is a combination of equations (13.26) and (15.57).
2. A product of forms is always their exterior product, so the form index of the product  $\mathbf{ab}$  is  $p + q$ .

3. The anticommutation of the multivectors cancels the anticommutation of the forms when  $p$  and  $q$  are both odd.
- 

### 16.14.1 Interval, connection, curvature, and torsion forms

In multivector notation, the gravitational coordinates and momenta proved to be  $\Gamma_\kappa$  and  $e^\kappa$  (or vice versa). In forms notation, the corresponding coordinates and momenta are the Lorentz connection bivector 1-form  $\Gamma$  and the line interval vector 1-form  $dx$  defined by

$$\Gamma \equiv \Gamma_\kappa dx^\kappa = \Gamma_{kl\kappa} \gamma^k \wedge \gamma^l dx^\kappa , \quad (16.199a)$$

$$dx \equiv e_\kappa dx^\kappa = e_{k\kappa} \gamma^k dx^\kappa , \quad (16.199b)$$

with, for  $\Gamma$ , implicit summation over distinct antisymmetric sets of indices  $kl$ . The Lorentz connection 1-form  $\Gamma$  and coordinate interval 1-form  $dx$  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 symbol  $dx$ , first introduced in this book in equation (2.19), should be understood as denoting a single holistic object, not a composition of  $d$  and  $x$ . It is tempting to use an abbreviated notation  $e$  for the 1-form  $e_\kappa dx^\kappa$ , but the notation  $dx$  is used here to emphasize that using forms language does not require switching to a whole new set of symbols. The adopted notation has the advantage that the  $p$ -volume element can be denoted by the abbreviated symbol  $d^p x$  introduced in equation (15.70), instead of by (a factor times) the wedge product of  $p$  copies of  $dx$  (or  $e$ ).

The Riemann bivector 2-form  $R$  is defined by, in the forms notation of equation (15.74),

$$R \equiv R_{\kappa\lambda} d^2x^{\kappa\lambda} = R_{\kappa\lambda mn} \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda} , \quad (16.200)$$

again implicitly summed over distinct antisymmetric indices  $mn$  and  $\kappa\lambda$ . The exterior derivative of the Lorentz connection 1-form is, equation (15.65), the 2-form

$$d\Gamma = \left( \frac{\partial \Gamma_\lambda}{\partial x^\kappa} - \frac{\partial \Gamma_\kappa}{\partial x^\lambda} \right) d^2x^{\kappa\lambda} = \left( \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{mn\kappa}}{\partial x^\lambda} \right) \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda} , \quad (16.201)$$

implicitly summed over distinct antisymmetric indices  $mn$  and  $\kappa\lambda$ . The commutator  $\frac{1}{4}[\Gamma, \Gamma]$  of the 1-form  $\Gamma$  with itself is the bivector 2-form

$$\frac{1}{4}[\Gamma, \Gamma] = \frac{1}{2}[\Gamma_\kappa, \Gamma_\lambda] d^2x^{\kappa\lambda} = (\Gamma_{m\lambda}^p \Gamma_{pn\kappa} - \Gamma_{m\kappa}^p \Gamma_{pn\lambda}) \gamma^m \wedge \gamma^n d^2x^{\kappa\lambda} , \quad (16.202)$$

again implicitly summed over distinct antisymmetric indices  $mn$  and  $\kappa\lambda$ . The commutator  $[\Gamma, \Gamma]$  of the bivector 1-form  $\Gamma$  is symmetric, the anticommutation of multivectors cancelling against the anticommutation of 1-forms, equation (16.197). Equations (16.201) and (16.202) imply that the Riemann 2-form  $R$  is related to the Lorentz connection 1-form  $\Gamma$  by

$$R \equiv d\Gamma + \frac{1}{4}[\Gamma, \Gamma] . \quad (16.203)$$

Equation (16.203) is **Cartan's second equation of structure**. It constitutes the definition of Riemann curvature  $R$  in terms of the Lorentz connection  $\Gamma$ .

The torsion vector 2-form  $\mathbf{S}$  is defined by (the minus sign ensures that Cartan's equation (16.207) takes conventional form)

$$\mathbf{S} \equiv -S_{m\kappa\lambda} \boldsymbol{\gamma}^m d^2x^{\kappa\lambda}, \quad (16.204)$$

implicitly summed over distinct antisymmetric indices  $\kappa\lambda$ . The exterior derivative of the line interval 1-form  $d\mathbf{x}$  is the 2-form

$$d\mathbf{dx} \equiv \left( \frac{\partial e_\lambda}{\partial x^\kappa} - \frac{\partial e_\kappa}{\partial x^\lambda} \right) d^2x^{\kappa\lambda} = \left( \frac{\partial e_{m\lambda}}{\partial x^\kappa} - \frac{\partial e_{m\kappa}}{\partial x^\lambda} \right) \boldsymbol{\gamma}^m d^2x^{\kappa\lambda} = -2 d_{m[\kappa\lambda]} \boldsymbol{\gamma}^m d^2x^{\kappa\lambda}, \quad (16.205)$$

again implicitly summed over distinct antisymmetric indices  $\kappa\lambda$ . The  $d_{m\kappa\lambda}$  in the rightmost expression of equations (16.205) are the vierbein derivatives defined by equation (11.33). The commutator  $\frac{1}{2}[\Gamma, d\mathbf{x}]$  of the 1-forms  $\Gamma$  and  $d\mathbf{x}$  is the vector 2-form

$$\frac{1}{2}[\Gamma, d\mathbf{x}] = [\Gamma_\kappa, e_\lambda] d^2x^{\kappa\lambda} = -2 \Gamma_{m\kappa\lambda} \boldsymbol{\gamma}^m d^2x^{\kappa\lambda}, \quad (16.206)$$

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 multivector forms language to, from equations (16.204)–(16.206),

$$\boxed{\mathbf{S} \equiv d\mathbf{dx} + \frac{1}{2}[\Gamma, d\mathbf{x}].} \quad (16.207)$$

Equation (16.207) is **Cartan's first equation of structure**. Cartan's equations of structure (16.207) and (16.203), 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 (16.234) for the torsion and curvature are obtained from extremizing the Hilbert action.

### 16.14.2 Area and volume forms

The other factor in the gravitational Lagrangian (16.162) is  $e^\lambda \wedge e^\kappa$ . The 2-form corresponding to  $e^\lambda \wedge e^\kappa$  is the element of area  $d^2x$  defined by equation (15.70),

$$d^2x = -\frac{1}{2} d\mathbf{x} \wedge d\mathbf{x} = -e_\kappa \wedge e_\lambda d^2x^{\kappa\lambda} = -(e_{m\kappa} e_{n\lambda} - e_{m\lambda} e_{n\kappa}) \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n d^2x^{\kappa\lambda}, \quad (16.208)$$

implicitly summed over distinct antisymmetric pairs  $mn$  and  $\kappa\lambda$  of indices. The minus sign on the right hand side of equation (16.208) cancels the minus sign in the multivector Lagrangian (16.162) (both minuses come from the same  $(-)^{[p/2]}$  factor with  $p = 2$ ).

The exterior derivative  $dd^p\mathbf{x}$  of the  $p$ -volume element is

$$dd^p\mathbf{x} = \mathbf{d}^{p-1}\mathbf{x} \wedge d\mathbf{dx}. \quad (16.209)$$

The  $1/p!$  factor in the definition (15.70) of the  $p$ -volume element absorbs the factor of  $p$  from differentiating  $p$  products of  $d\mathbf{x}$ , and the  $(-)^{[p/2]-[(p-1)/2]} = (-)^{p-1}$  factor cancels the sign from commuting  $d\mathbf{dx}$  past  $\mathbf{d}^{p-1}\mathbf{x}$ .

The form dual of the  $p$ -volume  $\mathbf{d}^p \mathbf{x}$  is the dual  $q$ -volume,  ${}^*(\mathbf{d}^p \mathbf{x}) \equiv {}^*\mathbf{d}^q \mathbf{x}$ , which in turn equals the pseudoscalar  $I$  times the  $q$ -volume, equation (15.78),

$${}^*(\mathbf{d}^p \mathbf{x}) = {}^*\mathbf{d}^q \mathbf{x} = I \mathbf{d}^q \mathbf{x}. \quad (16.210)$$

The  $p$ -volume and its  $q$ -form dual are both multivectors of grade  $p$ . For example, the form dual, equation (16.193), of the area element  $\mathbf{d}^2 \mathbf{x}$  is the dual area element  ${}^*\mathbf{d}^2 \mathbf{x}$ ,

$${}^*\mathbf{d}^2 \mathbf{x} = \varepsilon_{\kappa\lambda\mu\nu} e^\kappa \wedge e^\lambda d^2 x^{\mu\nu} = 2 \varepsilon_{\kappa\lambda\mu\nu} e_m{}^\kappa e_n{}^\lambda \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n d^2 x^{\mu\nu} = 2 \varepsilon_{klmn} e^m{}_\kappa e^n{}_\lambda \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l d^2 x^{\kappa\lambda}, \quad (16.211)$$

implicitly summed over distinct antisymmetric indices  $\kappa\lambda$ ,  $\mu\nu$ ,  $kl$ , and  $mn$ . The exterior derivative  $d {}^*\mathbf{d}^q \mathbf{x}$  of the dual  $q$ -volume element is

$$d {}^*\mathbf{d}^q \mathbf{x} = I d d^q \mathbf{x} = I (\mathbf{d}^{q-1} \mathbf{x} \wedge d \mathbf{d} \mathbf{x}) = (I \mathbf{d}^{q-1} \mathbf{x}) \cdot d \mathbf{d} \mathbf{x} = {}^*\mathbf{d}^{q-1} \mathbf{x} \cdot d \mathbf{d} \mathbf{x}, \quad (16.212)$$

the third equality following from the duality relation (13.39).

**Concept question 16.9 Scalar product of the interval form  $d\mathbf{x}$ .** 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  $d\mathbf{x}$  with itself is

$$d\mathbf{x} \cdot d\mathbf{x} = 2 g_{\mu\nu} d^2 x^{\mu\nu} = 0 \quad (16.213)$$

(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  $d^2 x^{\mu\nu}$ .

**Exercise 16.10 Triple products involving products of the interval form  $d\mathbf{x}$ .** Show that for any multivector form  $\mathbf{a}$  (of any grade and any form index),

$$d\mathbf{x} \wedge (d\mathbf{x} \cdot \mathbf{a}) = d\mathbf{x} \cdot (d\mathbf{x} \wedge \mathbf{a}) = -\frac{1}{2} [\mathbf{d}^2 \mathbf{x}, \mathbf{a}]. \quad (16.214)$$

**Solution.** The first equality in equations (16.214) can be proved by expanding the multivector forms in components. The final equality is an application of the triple-product relation (13.38), along with  $d\mathbf{x} \cdot d\mathbf{x} = 0$ .

### 16.14.3 Gravitational Lagrangian 4-form

Recall that the scalar volume element  $d^4 x$  that goes into the action is really the dual scalar 4-volume  ${}^*d^4 x$ , equation (15.78). 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

$$S_g = \int L_g , \quad (16.215)$$

where  $L_g$  is the gravitational Lagrangian scalar 4-form corresponding to the Lagrangian (16.162),

$$L_g \equiv \frac{1}{8\pi} {}^*d^2x \cdot R = \frac{1}{8\pi} {}^*d^2x \cdot (d\Gamma + \frac{1}{4}[\Gamma, \Gamma]) . \quad (16.216)$$

The dot in  ${}^*d^2x \cdot R$  signifies a scalar product of the bivectors  ${}^*d^2x$  and  $R$ . The product  ${}^*d^2x \cdot R$  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 (16.216) is in Hamiltonian form  $L_g = p \cdot dq - H_g$  with coordinates  $q = \Gamma$  and momenta  $p = {}^*d^2x/(8\pi)$ , and (super-)Hamiltonian 4-form

$$H_g = -\frac{1}{32\pi} {}^*d^2x \cdot [\Gamma, \Gamma] . \quad (16.217)$$

The Lagrangian scalar 4-form (16.216) can also be written elegantly, from the expression (16.210) for the dual volume  ${}^*d^2x$ , and the duality relation (13.39),

$$\boxed{L_g \equiv \frac{I}{8\pi} d^2x \wedge R = \frac{I}{8\pi} d^2x \wedge (d\Gamma + \frac{1}{4}[\Gamma, \Gamma])} . \quad (16.218)$$

Expanded in components, the gravitational Lagrangian 4-form (16.216) or (16.218) is

$$L_g = \frac{1}{8\pi} \varepsilon_{\pi\rho\mu\nu} (e^\rho \wedge e^\pi) \cdot R_{\kappa\lambda} d^4x^{\kappa\lambda\mu\nu} = \frac{I}{8\pi} e_\nu \wedge e_\mu \wedge R_{\kappa\lambda} d^4x^{\kappa\lambda\mu\nu} , \quad (16.219)$$

implicitly summed over distinct antisymmetric indices  $\kappa\lambda$ ,  $\mu\nu$ , and  $\pi\rho$ . The Lagrangian 4-form (16.218) is in Hamiltonian form  $L_g = I(p \wedge dq) - H_g$  with coordinates  $q = \Gamma$  and momenta  $p = d^2x/(8\pi)$ , and (super-)Hamiltonian scalar 4-form

$$\boxed{H_g = -\frac{I}{32\pi} d^2x \wedge [\Gamma, \Gamma]} . \quad (16.220)$$

#### 16.14.4 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  $\Gamma$  and the line element  $dx$ . In forms notation, when the fields are varied, it is the coefficients  $\Gamma_{kl\kappa}$  and  $e_{k\kappa}$  that are varied, the tetrad  $\gamma_k$  and the line interval  $dx^\kappa$  being considered fixed. Thus the variation  $\delta\Gamma$  of the Lorentz connection is

$$\delta\Gamma \equiv (\delta\Gamma_\kappa) dx^\kappa \equiv (\delta\Gamma_{kl\kappa}) \gamma^k \wedge \gamma^l dx^\kappa , \quad (16.221)$$

implicitly summed over distinct antisymmetric indices  $kl$ , and the variation  $\delta dx$  of the line interval is

$$\delta dx \equiv (\delta e_\kappa) dx^\kappa \equiv (\delta e_{k\kappa}) \gamma^k dx^\kappa . \quad (16.222)$$

The variation  $\delta \mathbf{d}^p \mathbf{x}$  of the  $p$ -volume element defined by equation (15.70) is

$$\delta \mathbf{d}^p \mathbf{x} = (-)^{p-1} \mathbf{d}^{p-1} \mathbf{x} \wedge \delta \mathbf{d} \mathbf{x} . \quad (16.223)$$

The variation  $\delta^* \mathbf{d}^q \mathbf{x}$  of the dual  $q$ -volume element is

$$\delta^* \mathbf{d}^q \mathbf{x} = I \delta \mathbf{d}^q \mathbf{x} = (-)^{q-1} I (\mathbf{d}^{q-1} \mathbf{x} \wedge \delta \mathbf{d} \mathbf{x}) = (-)^{q-1} (I \mathbf{d}^{q-1} \mathbf{x}) \cdot \delta \mathbf{d} \mathbf{x} = (-)^{q-1} {}^* \mathbf{d}^{q-1} \mathbf{x} \cdot \delta \mathbf{d} \mathbf{x} , \quad (16.224)$$

the third equality following from the duality relation (13.39).

The variation of the action with gravitational Lagrangian (16.219) with respect to the fields  $\mathbf{\Gamma}$  and  $\mathbf{d} \mathbf{x}$  is

$$\delta S_g = \frac{I}{8\pi} \int \mathbf{d}^2 \mathbf{x} \wedge d(\delta \mathbf{\Gamma}) + \frac{1}{4} \mathbf{d}^2 \mathbf{x} \wedge \delta [\mathbf{\Gamma}, \mathbf{\Gamma}] + \delta(\mathbf{d}^2 \mathbf{x}) \wedge \mathbf{R} . \quad (16.225)$$

The first term integrates by parts to

$$\mathbf{d}^2 \mathbf{x} \wedge d(\delta \mathbf{\Gamma}) = d(\mathbf{d}^2 \mathbf{x} \wedge \delta \mathbf{\Gamma}) - d(\mathbf{d}^2 \mathbf{x}) \wedge \delta \mathbf{\Gamma} . \quad (16.226)$$

The second term in the integrand of (16.225) is

$$\frac{1}{4} \mathbf{d}^2 \mathbf{x} \wedge \delta [\mathbf{\Gamma}, \mathbf{\Gamma}] = \frac{1}{2} \mathbf{d}^2 \mathbf{x} \wedge [\mathbf{\Gamma}, \delta \mathbf{\Gamma}] = \frac{1}{2} [\mathbf{d}^2 \mathbf{x}, \mathbf{\Gamma}] \wedge \delta \mathbf{\Gamma} = -\frac{1}{2} [\mathbf{\Gamma}, \mathbf{d}^2 \mathbf{x}] \wedge \delta \mathbf{\Gamma} , \quad (16.227)$$

the second step of which applies the multivector triple-product relation (13.37). The coefficients of the  $\wedge \delta \mathbf{\Gamma}$  terms in equations (16.226) and (16.227) combine to

$$- d(\mathbf{d}^2 \mathbf{x}) - \frac{1}{2} [\mathbf{\Gamma}, \mathbf{d}^2 \mathbf{x}] = - \mathbf{d} \mathbf{x} \wedge \mathbf{S} , \quad (16.228)$$

the torsion  $\mathbf{S}$  being defined by equation (16.207). To switch between commutators  $[\mathbf{\Gamma}, \mathbf{d} \mathbf{x}]$  and commutators  $[\mathbf{\Gamma}, \mathbf{d}^2 \mathbf{x}]$ , use the result (16.214) along with the fact that  $\mathbf{d} \mathbf{x} \cdot \mathbf{a} = \frac{1}{2} [\mathbf{d} \mathbf{x}, \mathbf{a}]$  for any bivector form  $\mathbf{a}$ . The third term in the integrand of (16.225) is

$$\delta(\mathbf{d}^2 \mathbf{x}) \wedge \mathbf{R} = - \delta \mathbf{d} \mathbf{x} \wedge \mathbf{d} \mathbf{x} \wedge \mathbf{R} = \delta \mathbf{d} \mathbf{x} \wedge {}^* \mathbf{G} , \quad (16.229)$$

where  ${}^* \mathbf{G} \equiv - \mathbf{d} \mathbf{x} \wedge \mathbf{R}$  is the double-dual, equation (16.194), of the Einstein vector 1-form  $\mathbf{G} \equiv G_{\nu n} \boldsymbol{\gamma}^n dx^\nu$ ,

$$\begin{aligned} \mathbf{G} &\equiv - I^* (\mathbf{d} \mathbf{x} \wedge \mathbf{R}) \\ &= \varepsilon^{klmn} \varepsilon_{\kappa \lambda \mu \nu} R^{\kappa \lambda}_{kl} e_m{}^\mu \boldsymbol{\gamma}_n dx^\nu \\ &= -6 e^{[k}{}_\kappa e^{l]}{}_\lambda e^{n]}{}_\nu R^{\kappa \lambda}_{kl} \boldsymbol{\gamma}_n dx^\nu \\ &= (R_{\nu n} - \frac{1}{2} R e_{n \nu}) \boldsymbol{\gamma}^n dx^\nu , \end{aligned} \quad (16.230)$$

implicitly summed over distinct antisymmetric sequences  $kl$  and  $\kappa \lambda$ , and over all  $m, n, \mu$ , and  $\nu$ . Combining equations (16.226)–(16.229) brings the variation (16.225) of the gravitational action to

$$\delta S_g = \frac{I}{8\pi} \oint \mathbf{d}^2 \mathbf{x} \wedge \delta \mathbf{\Gamma} + \frac{I}{8\pi} \int - (\mathbf{d} \mathbf{x} \wedge \mathbf{S}) \wedge \delta \mathbf{\Gamma} - \delta \mathbf{d} \mathbf{x} \wedge (\mathbf{d} \mathbf{x} \wedge \mathbf{R}) . \quad (16.231)$$

The variation of the matter action  $S_m$  with respect to  $\delta\Gamma$  and  $\delta\mathbf{dx}$  defines the spin angular-momentum  $\Sigma$  (compare equation (16.115)), and the matter energy-momentum  $\mathbf{T}$  (compare equations (16.111)),

$$\delta S_m = \int {}^*\Sigma \cdot \delta\Gamma - \delta\mathbf{dx} \cdot {}^*\mathbf{T} = -I \int \overset{**}{\Sigma} \wedge \delta\Gamma + \delta\mathbf{dx} \wedge \overset{**}{\mathbf{T}} , \quad (16.232)$$

where  $\overset{**}{\Sigma}$  and  $\overset{**}{\mathbf{T}}$  are the double-duals, equation (16.194), of the spin angular-momentum  $\Sigma$  and energy-momentum  $\mathbf{T}$  of the matter. The components of the spin angular-momentum bivector 1-form  $\Sigma$  and the energy-momentum vector 1-form  $\mathbf{T}$  are (the minus sign in the definition of  $\Sigma$  conforms to the convention for the definition of torsion  $\mathbf{S}$ , equation (16.204))

$$\Sigma \equiv \Sigma_{\kappa} dx^{\kappa} = -\Sigma_{klm} \gamma^l \wedge \gamma^m dx^{\kappa} , \quad (16.233a)$$

$$\mathbf{T} \equiv \mathbf{T}_{\kappa} dx^{\kappa} = T_{km} \gamma^m dx^{\kappa} , \quad (16.233b)$$

with, for  $\Sigma$ , implicit summation over distinct antisymmetric sets of indices  $kl$ . Extremizing the combined gravitational and matter actions with respect to  $\delta\Gamma$  and  $\delta\mathbf{dx}$  yields the torsion and Einstein equations of motion in the form

$$\boxed{\mathbf{dx} \wedge \mathbf{S} = -8\pi \overset{**}{\Sigma}} , \quad (16.234a)$$

$$\boxed{\mathbf{dx} \wedge \mathbf{R} = -8\pi \overset{**}{\mathbf{T}}} . \quad (16.234b)$$

The torsion equation of motion (16.234a) is a bivector 3-form with  $6 \times 4 = 24$  components, while the Einstein equation of motion (16.234b) is a pseudovector 3-form with  $4 \times 4 = 16$  components. The Einstein equation (16.234b) is equivalent to the traditional expression

$$\mathbf{G} = 8\pi \mathbf{T} . \quad (16.235)$$

The contracted Bianchi identities (16.379) enforce conservation laws for the total spin angular-momentum  $\overset{**}{\Sigma}$  and total energy-momentum  $\overset{**}{\mathbf{T}}$ , §16.14.7 and §16.14.8.

#### 16.14.5 Alternative gravitational action in multivector forms notation

As in §16.11 and §16.13.3, the coordinates and momenta can be traded without changing the equations of motion. Integrating the  $d^2\mathbf{x} \wedge d\Gamma$  term in the Lagrangian (16.218) by parts gives

$$d^2\mathbf{x} \wedge d\Gamma = d(d^2\mathbf{x} \wedge \Gamma) + dd\mathbf{x} \wedge (\mathbf{dx} \wedge \Gamma) . \quad (16.236)$$

Discarding the total divergence yields the alternative Lagrangian

$$L'_g = \frac{I}{8\pi} d\mathbf{dx} \wedge \boldsymbol{\pi} - H_g , \quad (16.237)$$

where  $\boldsymbol{\pi}$  is defined to be

$$\boxed{\boldsymbol{\pi} \equiv \mathbf{dx} \wedge \Gamma} , \quad (16.238)$$

and  $H_g$  is the same (super-)Hamiltonian as before, equation (16.220). The alternative Lagrangian (16.237) is in Hamiltonian form with coordinates  $\mathbf{dx}$  and momenta  $\boldsymbol{\pi}/(8\pi)$ .

The Lorentz connection  $\boldsymbol{\Gamma}$ , which is a bivector 1-form, and the momentum  $\boldsymbol{\pi}$ , which is a pseudovector 2-form, both have the same number of components, 24. The components are invertibly related to each other. A manipulation similar to equations (16.230) gives the double-dual of the momentum  $\boldsymbol{\pi}$  as

$$\overset{**}{\boldsymbol{\pi}} = (\Gamma_{\mu\nu l} - 2e_{l\mu}\Gamma_{\nu p}^p) \boldsymbol{\gamma}^l d^2x^{\mu\nu}, \quad (16.239)$$

the components of which, when the coordinate and tetrad indices are transposed, match the  $\pi_{mn\lambda}$  given by equation (16.156). In multivector forms language, the Lorentz connection  $\boldsymbol{\Gamma}$  is given in terms of the momentum  $\boldsymbol{\pi}$  by

$$\boldsymbol{\Gamma} = (\overset{**}{\boldsymbol{\pi}} - \mathbf{dx} \wedge (\mathbf{dx} \wedge \overset{**}{\boldsymbol{\pi}}))^\top, \quad (16.240)$$

where  $^\top$  denotes the transpose operation (16.251).

Variation of the gravitational action  $S'_g$  with the alternative Lagrangian (16.237) yields

$$\delta S'_g = \frac{I}{8\pi} \oint \delta \mathbf{dx} \wedge \boldsymbol{\pi} + \frac{I}{8\pi} \int \mathbf{S} \wedge \delta \boldsymbol{\pi} + \delta \mathbf{dx} \wedge \boldsymbol{\Pi}, \quad (16.241)$$

where the curvature pseudovector 3-form  $\boldsymbol{\Pi}$  is defined to be

$$\boxed{\boldsymbol{\Pi} \equiv -\mathbf{dx} \wedge \mathbf{R} + \mathbf{S} \wedge \boldsymbol{\Gamma} = d\boldsymbol{\pi} + \frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{\pi}] + \frac{1}{4}\mathbf{dx} \wedge [\boldsymbol{\Gamma}, \boldsymbol{\Gamma}]} . \quad (16.242)$$

Previously, variation of the matter action  $S_m$  with respect to  $\delta\boldsymbol{\Gamma}$  and  $\delta\mathbf{dx}$  defined the (double duals of the) spin angular-momentum  $\boldsymbol{\Sigma}$  and matter energy-momentum  $\mathbf{T}$ , equation (16.232). Variation of the matter action  $S_m$  instead with respect to  $\delta\boldsymbol{\pi}$  and  $\delta\mathbf{dx}$  defines modified versions  $\tilde{\boldsymbol{\Sigma}}$  and  $\tilde{\mathbf{T}}$  of the spin angular-momentum and energy-momentum (beware that the components  $\pi_{mn\lambda}$  in the earlier variation (16.122) are the components of the double-dual  $\overset{**}{\boldsymbol{\pi}}$ ),

$$\delta S_m = -I \int \tilde{\boldsymbol{\Sigma}} \wedge \delta \boldsymbol{\pi} + \delta \mathbf{dx} \wedge \tilde{\mathbf{T}} . \quad (16.243)$$

where the vector 2-form  $\tilde{\boldsymbol{\Sigma}}$  is (the minus sign conforms to the convention for the torsion  $\mathbf{S}$  and spin angular-momentum  $\boldsymbol{\Sigma}$ , equations (16.204) and (16.233a))

$$\tilde{\boldsymbol{\Sigma}} = -\tilde{\Sigma}_{\lambda mn} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n dx^\lambda . \quad (16.244)$$

The original  $\boldsymbol{\Sigma}$  and modified  $\tilde{\boldsymbol{\Sigma}}$  spin angular-momenta are invertibly related to each other, while the (double-dual of the) original  $\mathbf{T}$  and modified  $\tilde{\mathbf{T}}$  energy-momenta differ by a term depending on the spin angular-momentum,

$$\overset{**}{\boldsymbol{\Sigma}} = -\mathbf{dx} \wedge \tilde{\boldsymbol{\Sigma}}, \quad (16.245a)$$

$$\overset{**}{\mathbf{T}} = \tilde{\mathbf{T}} - \tilde{\boldsymbol{\Sigma}} \wedge \boldsymbol{\Gamma} . \quad (16.245b)$$

In components, the relation (16.245a) between the original  $\Sigma$  and modified  $\tilde{\Sigma}$  spin angular-momenta is equation (16.124). The equations of motion for the torsion  $S$  and curvature  $\Pi$  are

$$\boxed{S = 8\pi\tilde{\Sigma}} , \quad (16.246a)$$

$$\boxed{\Pi = 8\pi\tilde{T}} . \quad (16.246b)$$

The contraction of  $\pi$  defines the expansion  $\vartheta$ , a pseudoscalar 3-form with 4 components,

$$\vartheta \equiv dx \wedge \Gamma = -\frac{1}{2} dx \wedge \pi . \quad (16.247)$$

The double-dual of the expansion is the scalar 1-form

$$\overset{**}{\vartheta} = \Gamma_{\kappa p}^p dx^\kappa . \quad (16.248)$$

The transpose (16.251) of the double dual of the expansion is

$$\overset{**}{\vartheta}^\top = \Gamma_{kp}^p \gamma^k , \quad (16.249)$$

whose tetrad time component  $\Gamma_{0p}^p$  is what is commonly called the expansion, §18.3, justifying the nomenclature. The exterior derivative of  $\vartheta$  is the pseudoscalar 4-form

$$d\vartheta = dx \wedge (-\frac{1}{2} dx \wedge (R + \frac{1}{4}[\Gamma, \Gamma]) + S \wedge \Gamma) . \quad (16.250)$$

If  $dx \wedge R$  and  $dx \wedge S$  are replaced by their matter energy-momentum and spin angular-momentum sources in accordance with equation (16.234), then equation (16.250) looks like a covariant version of the Raychaudhuri equation (§18.2) expressed in multivector forms language.

#### 16.14.6 Transpose of a multivector form

The transpose  $a^\top$  of a multivector form  $a \equiv a_{A\Lambda} \gamma^A dx^\Lambda$  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,

$$a^\top = (a_{A\Lambda} \gamma^A dx^\Lambda)^\top \equiv e^A{}_\Pi e_B{}^\Lambda a_{A\Lambda} \gamma^B dx^\Pi = a_{\Pi B} \gamma^B dk x^\Pi , \quad (16.251)$$

implicitly summed over distinct sequences  $A, B, \Lambda, \Pi$  of indices. For example, the transpose of a vector 2-form is the bivector 1-form

$$(a_{k\lambda\mu} \gamma^k dx^\lambda)^\top \equiv e^k{}_\kappa e_l{}^\lambda e_m{}^\mu a_{k\lambda\mu} \gamma^l \wedge \gamma^m dx^\kappa = a_{\kappa l m} \gamma^l \wedge \gamma^m dx^\kappa . \quad (16.252)$$

The transpose of a symmetric tensor  $a$ , one satisfying,  $a_{k\lambda} \equiv a_{kl} e^l{}_\lambda = a_{lk} e^l{}_\lambda \equiv a_{\lambda k}$ , is itself,

$$a^\top = (a_{k\lambda} \gamma^k dx^\lambda)^\top = a_{\lambda k} \gamma^k dx^\lambda = a_{k\lambda} \gamma^k dx^\lambda = a . \quad (16.253)$$

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  $dx$  is itself,

$$dx^\top = (e_{k\lambda} \gamma^k dx^\lambda)^\top = e^k{}_\kappa e_l{}^\lambda e_{k\lambda} \gamma^l dx^\kappa = e_{l\kappa} \gamma^l dx^\kappa = dx . \quad (16.254)$$

The transpose of a wedge product of multivector forms  $\mathbf{a}$  and  $\mathbf{b}$  is the wedge product of their transposes,

$$(\mathbf{a} \wedge \mathbf{b})^\top = \mathbf{a}^\top \wedge \mathbf{b}^\top . \quad (16.255)$$

The transpose of the double-dual of a multivector form  $\mathbf{a}$  is the double dual of its transpose

$$(\mathbf{a}^{**})^\top = I^*(\mathbf{a}^\top) = (\mathbf{a}^\top)^{**} . \quad (16.256)$$

#### 16.14.7 Conservation of spin angular-momentum in multivector forms language

The action  $S_m$  of any individual matter field is Lorentz invariant. Lorentz symmetry implies a conservation law (16.261) for the spin angular-momentum of the field.

Under an infinitesimal Lorentz transformation generated by the bivector  $\epsilon = \epsilon_{kl} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l$ , any multivector form  $\mathbf{a}$  whose multivector components transform like a tensor varies as, equation (16.95),

$$\delta \mathbf{a} = \frac{1}{2}[\epsilon, \mathbf{a}] . \quad (16.257)$$

In particular, since the vierbein  $e_{k\kappa}$  is a tetrad vector, the variation of the line interval  $d\mathbf{x} \equiv e_{k\kappa} \boldsymbol{\gamma}^k dx^\kappa$  under an infinitesimal Lorentz transformation is

$$\delta d\mathbf{x} = \frac{1}{2}[\epsilon, d\mathbf{x}] . \quad (16.258)$$

The components of the Lorentz connection  $\mathbf{\Gamma} \equiv \Gamma_{mn\lambda} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n dx^\lambda$  do not constitute a tetrad tensor, so do not transform like equation (16.257). Rather, the Lorentz connection transforms as equation (16.96), which in multivector forms language translates to

$$\delta \mathbf{\Gamma} = d\epsilon + \frac{1}{2}[\mathbf{\Gamma}, \epsilon] . \quad (16.259)$$

Under an infinitesimal Lorentz transformation generated by the bivector  $\epsilon$ , the matter action  $S_m$  varies as

$$\begin{aligned} \delta S_m &= -I \int \mathbf{\Sigma}^{**} \wedge (d\epsilon + \frac{1}{2}[\mathbf{\Gamma}, \epsilon]) + \frac{1}{2}[\epsilon, d\mathbf{x}] \wedge \mathbf{T}^{**} \\ &= I \oint \mathbf{\Sigma}^{**} \wedge \epsilon - I \int (d\mathbf{\Sigma}^{**} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\Sigma}^{**}] + d\mathbf{x} \cdot \mathbf{T}^{**}) \wedge \epsilon . \end{aligned} \quad (16.260)$$

Invariance of the action under local Lorentz transformations requires that the variation (16.260) must vanish for arbitrary choices of the bivector  $\epsilon$  vanishing on the initial and final hypersurfaces. Consequently the spin angular-momentum  $\mathbf{\Sigma}^{**}$  must satisfy the covariant conservation equation

$$\boxed{d\mathbf{\Sigma}^{**} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\Sigma}^{**}] + d\mathbf{x} \cdot \mathbf{T}^{**} = 0} . \quad (16.261)$$

Equation (16.261) is the same as the conservation equation (16.130) derived previously in index notation. Equation (16.261) is consistent with the contracted torsion Bianchi identity (16.379a), which enforces the spin angular-momentum conservation equation (16.261) summed over all species. If the spin angular-momentum

of a matter component vanishes, then the conservation equation (16.261) implies that the energy-momentum tensor of that matter component is symmetric,

$$dx \cdot \overset{**}{T} = 0 . \quad (16.262)$$

#### 16.14.8 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 (16.279) for the energy-momentum of the field.

Any infinitesimal 1-form  $\epsilon \equiv \epsilon^\kappa dx^\kappa$  generates an infinitesimal coordinate transformation

$$x^\kappa \rightarrow x^\kappa + \epsilon^\kappa . \quad (16.263)$$

As discussed in §7.34, the variation of any quantity  $a$  with respect to an infinitesimal coordinate transformation  $\epsilon$  is, by construction, minus its Lie derivative,  $-\mathcal{L}_\epsilon a$ , with respect to the vector  $\epsilon^\kappa$ . 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  $\epsilon$  with a  $p$ -form  $a \equiv a_{\textcolor{brown}{A}} d^p x^{\textcolor{brown}{A}}$  is, mirroring the multivector dot product (13.33) (the form dot  $\cdot$  is written slightly larger than the multivector dot  $\cdot$  to distinguish the two),

$$\epsilon \cdot a \equiv p \epsilon^\kappa a_{\kappa \textcolor{brown}{A}} d^{p-1} x^{\textcolor{brown}{A}} , \quad (16.264)$$

implicitly summed over distinct antisymmetric sets of indices  $\kappa \textcolor{brown}{A}$ . The form dot product of a 1-form  $\epsilon$  and a 0-form is zero, consistent with the convention (13.34) for multivectors. A useful result is that for any two multivector forms  $a$  and  $b$  with product  $ab$  (a geometric product of multivectors and an exterior product of forms), the dot product of the 1-form  $\epsilon$  with the product  $ab$  is

$$\epsilon \cdot (ab) = (\epsilon \cdot a)b + (-)^p a(\epsilon \cdot b) , \quad (16.265)$$

where  $p$  is the form index of  $a$ .

From the definition (7.129) of the Lie derivative of a coordinate tensor, it can be shown (Exercise 16.11 asks you to do this) that the Lie derivative of a  $p$ -form  $a$  with respect to a 1-form  $\epsilon$  is given by the elegant expression

$$\mathcal{L}_\epsilon a = \epsilon \cdot (da) + d(\epsilon \cdot a) , \quad (16.266)$$

which is known as **Cartan's magic formula**. Cartan's magic formula (16.266), 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  $d$ ,

$$\mathcal{L}_\epsilon da - d\mathcal{L}_\epsilon a = 0 . \quad (16.267)$$

Cartan's magic formula (16.266) holds also for multivector-valued forms, since multivectors are coordinate scalars (they are unchanged by coordinate transformations). However, a difficulty arises because the Lie derivative of a tetrad tensor is not a tetrad tensor (see Concept Question 25.2). Consequently the Lie derivative of neither the line interval nor the Lorentz connection is a tetrad tensor. However, as pointed

out at the beginning of §5.2.1 of Hehl et al. (1995), the Lagrangian is a Lorentz scalar, so in varying the Lagrangian 4-form  $L$ , the exterior derivative can be replaced by the Lorentz-covariant exterior derivative,  $da \rightarrow D_\Gamma a \equiv da + \frac{1}{2}[\Gamma, a]$  (see §16.17.1),

$$\mathcal{L}_\epsilon L = \mathcal{L}_{\Gamma\epsilon} L \equiv \epsilon \cdot (D_\Gamma L) + D_\Gamma(\epsilon \cdot L) . \quad (16.268)$$

Thus the variation of the Lagrangian under a coordinate transformation can be carried out using the Lorentz-covariant Lie derivative

$$\mathcal{L}_{\Gamma\epsilon} a \equiv \epsilon \cdot (D_\Gamma a) + D_\Gamma(\epsilon \cdot a) \quad (16.269)$$

in place of the usual Lie derivative (16.266). The advantage of this replacement is that the Lorentz-covariant Lie derivatives of the line interval and Lorentz connection are then (coordinate and tetrad) tensors, and the resulting law of conservation of energy-momentum is manifestly tensorial, as it should be. The Lorentz-covariant derivative  $D_\Gamma$  is torsion-free acting on coordinate indices, but torsion-full acting on multivector (Lorentz) indices. An alternative version of the covariant magic formula (16.269) in terms of the torsion-free exterior derivative  $\mathring{D}$  and the contortion  $K$  is

$$\mathcal{L}_{\Gamma\epsilon} a = \epsilon \cdot (\mathring{D} a) + \mathring{D}(\epsilon \cdot a) + \frac{1}{2}[\epsilon \cdot K, a] , \quad (16.270)$$

which follows from  $\Gamma = \mathring{\Gamma} + K$  and the relation (16.265). As a particular example of the Lorentz-covariant magic formula (16.269), the variation  $\delta dx$  of the line interval under an infinitesimal coordinate transformation (16.263) generated by  $\epsilon$  is

$$\delta dx = -\mathcal{L}_{\Gamma\epsilon} dx = -\epsilon \cdot (D_\Gamma dx) - D_\Gamma(\epsilon \cdot dx) = -d(\epsilon \cdot dx) - \frac{1}{2}[\Gamma, \epsilon \cdot dx] - \epsilon \cdot S . \quad (16.271)$$

Alternatively, in terms of the torsion-free exterior derivative  $\mathring{D}$ ,

$$\delta dx = -\mathcal{L}_{\Gamma\epsilon} dx = -\mathring{D}(\epsilon \cdot dx) - \frac{1}{2}[\epsilon \cdot K, dx] = -d(\epsilon \cdot dx) - \frac{1}{2}[\mathring{\Gamma}, \epsilon \cdot dx] - \frac{1}{2}[\epsilon \cdot K, dx] . \quad (16.272)$$

The Lorentz connection  $\Gamma$  is a coordinate tensor but not a tetrad tensor, so the covariant magic formula (16.269) does not apply to the Lorentz connection. Rather, the variation  $\delta\Gamma$  of the Lorentz connection follows from the difference

$$\delta D_\Gamma a - D_\Gamma \delta a = \delta(da + \frac{1}{2}[\Gamma, a]) - (d\delta a + \frac{1}{2}[\Gamma, \delta a]) = \frac{1}{2}[\delta\Gamma, a] . \quad (16.273)$$

Thus the variation  $\delta\Gamma$  of the Lorentz connection under an infinitesimal coordinate transformation generated by  $\epsilon$  satisfies

$$\frac{1}{2}[\delta\Gamma, a] = -\mathcal{L}_{\Gamma\epsilon} D_\Gamma a + D_\Gamma \mathcal{L}_{\Gamma\epsilon} a = -\epsilon \cdot (D_\Gamma D_\Gamma a) + D_\Gamma D_\Gamma(\epsilon \cdot a) = -\frac{1}{2}\epsilon \cdot [R, a] + \frac{1}{2}[R, \epsilon \cdot a] = -\frac{1}{2}[\epsilon \cdot R, a] , \quad (16.274)$$

where  $R$  is the Riemann curvature bivector 2-form. Equation (16.273) holds for all multivector forms  $a$ , so

$$\delta\Gamma = -\mathcal{L}_{\Gamma\epsilon} \Gamma = -\epsilon \cdot R . \quad (16.275)$$

Inserting the variations (16.271) and (16.275) of the line interval  $dx$  and Lorentz connection  $\Gamma$  into the

variation (16.232) of the matter action yields the variation of the action under an infinitesimal coordinate transformation (16.263) generated by the 1-form  $\epsilon$ ,

$$\delta S_m = I \int (d(\epsilon \cdot dx) + \frac{1}{2}[\Gamma, \epsilon \cdot dx] + \epsilon \cdot S) \wedge \overset{**}{T} + \overset{**}{\Sigma} \wedge (\epsilon \cdot R) . \quad (16.276)$$

Integrating the  $d(\epsilon \cdot dx) \wedge \overset{**}{T}$  term by parts, and rearranging the  $\frac{1}{2}[\Gamma, \epsilon \cdot dx] \wedge \overset{**}{T}$  term using the multivector triple-product relation (13.37), yields

$$\delta S_m = I \oint (\epsilon \cdot dx) \wedge \overset{**}{T} + I \int -(\epsilon \cdot dx) \wedge (dT + \frac{1}{2}[\Gamma, \overset{**}{T}]) + (\epsilon \cdot S) \wedge \overset{**}{T} + \overset{**}{\Sigma} \wedge (\epsilon \cdot R) . \quad (16.277)$$

Invariance of the action under coordinate transformations requires that the variation (16.277) must vanish for arbitrary choices of the 1-form  $\epsilon$  vanishing on the initial and final hypersurfaces. Consequently the matter energy-momentum  $\overset{**}{T}$  must satisfy the conservation equation

$$(\epsilon \cdot dx) \wedge (dT + \frac{1}{2}[\Gamma, \overset{**}{T}]) - (\epsilon \cdot S) \wedge \overset{**}{T} - \overset{**}{\Sigma} \wedge (\epsilon \cdot R) = 0 . \quad (16.278)$$

Equivalently, in terms of the torsion-free connection  $\overset{\circ}{\Gamma}$  and the contortion  $K$ ,

$$(\epsilon \cdot dx) \wedge (dT + \frac{1}{2}[\overset{\circ}{\Gamma}, \overset{**}{T}]) - \frac{1}{2}[\epsilon \cdot K, dx] \wedge \overset{**}{T} - \overset{**}{\Sigma} \wedge (\epsilon \cdot R) = 0 . \quad (16.279)$$

I don't know a way to recast equations (16.278) or (16.279) in multivector forms notation with the arbitrary 1-form  $\epsilon$  factored out, but in components equation (16.279) reduces to equation (16.139) derived earlier.

If the spin angular-momentum of a matter component vanishes,  $\overset{**}{\Sigma} = 0$ , then the energy-momentum conservation law (16.279) for that matter component simplifies to

$$dT + \frac{1}{2}[\overset{\circ}{\Gamma}, \overset{**}{T}] = 0 . \quad (16.280)$$

If the energy-momentum conservation law (16.278) 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 (16.234), then the law of conservation of total energy-momentum becomes

$$(\epsilon \cdot dx) \wedge (d(dx \wedge R) + \frac{1}{2}[\Gamma, dx \wedge R]) - (\epsilon \cdot S) \wedge dx \wedge R - dx \wedge S \wedge (\epsilon \cdot R) = 0 , \quad (16.281)$$

which by the relation (16.265) rearranges to

$$(\epsilon \cdot dx) \wedge (d(dx \wedge R) + \frac{1}{2}[\Gamma, dx \wedge R] - S \wedge R) = 0 . \quad (16.282)$$

Equation (16.282) is true for arbitrary infinitesimal  $\epsilon$ , so the law of conservation of total energy-momentum is

$$d(dx \wedge R) + \frac{1}{2}[\Gamma, dx \wedge R] - S \wedge R = 0 , \quad (16.283)$$

which agrees with the contracted Bianchi identity (16.379b).

---

**Exercise 16.11 Lie derivative of a form.** Confirm from the definition (7.129) that the Lie derivative of a  $p$ -form is indeed given by Cartan's magic formula (16.266).

---

### 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  $\mathbf{a}$  that represents the remaining spatial-coordinate components (the spatial indices being suppressed for brevity),

$$\mathbf{a} \rightarrow \mathbf{a}_{\bar{t}} + \mathbf{a} \equiv a_{At\Lambda} \boldsymbol{\gamma}^A d^p x^{t\Lambda} + a_{A\Lambda} \boldsymbol{\gamma}^A d^p x^\Lambda , \quad (16.284)$$

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 §16.15.5, equation (16.308).

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

$$(\mathbf{ab})_{\bar{t}} = \mathbf{a}_{\bar{t}} \mathbf{b} + \mathbf{a} \mathbf{b}_{\bar{t}} \quad (16.285)$$

with no minus signs (minus signs from the antisymmetry of form indices cancel minus signs from commuting  $dt$  through a spatial form).

#### 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 (16.218). The gravitational coordinates in this case are the Lorentz connections  $\mathbf{\Gamma}$ , and their conjugate momenta are, depending on how one chooses to proceed, either the interval 1-form  $\mathbf{dx}$ , or else the area 2-form  $\mathbf{d}^2x$ . After the coordinate 3+1 split, the coordinates are the spatial components of the Lorentz connection  $\mathbf{\Gamma}$ , which is a bivector 1-form with  $6 \times 3 = 18$  components,

$$\mathbf{\Gamma} = \mathbf{\Gamma}_\alpha dx^\alpha = \Gamma_{kl\alpha} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l dx^\alpha . \quad (16.286)$$

The momenta are either the spatial components of the line interval  $\mathbf{dx}$ , which is a vector 1-form with  $4 \times 3 = 12$  components, or else the spatial components of the area element  $\mathbf{d}^2\mathbf{x}$ , which is a bivector 2-form with  $6 \times 3 = 18$  spatial components,

$$\mathbf{dx} = e_{k\alpha} dx^\alpha = e_{k\alpha} \boldsymbol{\gamma}^k dx^\alpha, \quad \mathbf{d}^2\mathbf{x} = -e_\alpha \wedge e_\beta d^2x^{\alpha\beta} = -2 e_{k\alpha} e_{l\beta} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l d^2x^{\alpha\beta}, \quad (16.287)$$

with in the case of  $\mathbf{d}^2\mathbf{x}$  implicit summation over distinct antisymmetric pairs  $kl$  and  $\alpha\beta$  of indices.

Either choice of conjugate momentum is problematic. The number 18 of components of the spatial area element  $\mathbf{d}^2\mathbf{x}$  matches the number 18 of components of the spatial connection  $\boldsymbol{\Gamma}$ , which is good. However, the 18-component spatial area element  $\mathbf{d}^2\mathbf{x}$  contains superfluous degrees of freedom compared to the 12-component spatial interval  $\mathbf{dx}$ .

A traditional solution, the one adopted so far in this chapter, is to treat the line interval  $\mathbf{dx}$  as the momentum conjugate to the coordinates  $\boldsymbol{\Gamma}$ . This approach encounters the difficulty that the number 12 of components of the spatial interval form  $\mathbf{dx}$  does not match the number 18 of components of the spatial connection form  $\boldsymbol{\Gamma}$ .

Despite the mismatch of coordinates and momenta, 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, provided that it is satisfied initially, it is guaranteed thereafter by conservation of electric charge. 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 (16.289) and constraint equations (16.290) follow directly from splitting equations (16.234) into time and space parts, but they can be derived at a more fundamental level by splitting the variation  $\delta S$  of the action into time and space parts. Splitting the variation  $\delta S_g$  of the gravitational action, equation (16.231), into time and space parts gives

$$\begin{aligned} \delta S_g = & \frac{I}{8\pi} \oint_{t_i}^{t_f} (\mathbf{d}^2\mathbf{x} \wedge \delta \boldsymbol{\Gamma})_{\bar{t}} + \left[ \oint \mathbf{d}^2\mathbf{x} \wedge \delta \boldsymbol{\Gamma} \right]_{t_i}^{t_f} \\ & + \frac{I}{8\pi} \int -(\mathbf{dx} \wedge \mathbf{S})_{\bar{t}} \wedge \delta \boldsymbol{\Gamma} - \delta \mathbf{dx} \wedge (\mathbf{dx} \wedge \mathbf{R})_{\bar{t}} - (\mathbf{dx} \wedge \mathbf{S}) \wedge \delta \boldsymbol{\Gamma}_{\bar{t}} - \delta \mathbf{dx}_{\bar{t}} \wedge (\mathbf{dx} \wedge \mathbf{R}). \end{aligned} \quad (16.288)$$

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  $\delta \boldsymbol{\Gamma}$  and  $\delta \mathbf{dx}$  of the spatial coordinates and momenta yields the equations of motion,

$$18 \text{ equations: } (\mathbf{dx} \wedge \mathbf{S})_{\bar{t}} = -8\pi \overset{**}{\boldsymbol{\Sigma}}_{\bar{t}}, \quad (16.289a)$$

$$12 \text{ equations: } (\mathbf{dx} \wedge \mathbf{R})_{\bar{t}} = -8\pi \overset{**}{\mathbf{T}}_{\bar{t}}. \quad (16.289b)$$

These are just the coordinate time components of the equations of motion (16.234). Variation with respect

to the variations  $\delta\Gamma_{\bar{t}}$  and  $\delta\mathbf{dx}_{\bar{t}}$  of the time components of the coordinates and momenta yields the Gaussian and Hamiltonian constraints,

$$6 \text{ Gaussian constraints: } \mathbf{dx} \wedge \mathbf{S} = -8\pi \overset{**}{\Sigma}_{\bar{t}}, \quad (16.290a)$$

$$4 \text{ Hamiltonian constraints: } \mathbf{dx} \wedge \mathbf{R} = -8\pi \overset{**}{\mathbf{T}}_{\bar{t}}. \quad (16.290b)$$

These are the purely spatial coordinate components of the equations of motion (16.234). Whereas the equations of motion (16.289) involve derivatives with respect to time  $t$ , the constraint equations (16.290) involve no time derivatives. More explicitly, the equations of motion (16.289) are

$$18 \text{ equations: } \mathbf{dx} \wedge (\mathbf{d}_{\bar{t}}\mathbf{dx} + \frac{1}{2}[\Gamma_{\bar{t}}, \mathbf{dx}] + \mathbf{d}\mathbf{dx}_{\bar{t}} + \frac{1}{2}[\Gamma, \mathbf{dx}_{\bar{t}}]) + \mathbf{dx}_{\bar{t}} \wedge \mathbf{S} = -8\pi \overset{**}{\Sigma}_{\bar{t}}, \quad (16.291a)$$

$$12 \text{ equations: } \mathbf{dx} \wedge (\mathbf{d}_{\bar{t}}\Gamma + \mathbf{d}\Gamma_{\bar{t}} + \frac{1}{2}[\Gamma, \Gamma_{\bar{t}}]) + \mathbf{dx}_{\bar{t}} \wedge \mathbf{R} = -8\pi \overset{**}{\mathbf{T}}_{\bar{t}}. \quad (16.291b)$$

The exterior time derivative here is the 1-form  $\mathbf{d}_{\bar{t}} \equiv dt \partial/\partial t$ . The equations of motion (16.291) are problematic not only because they remain unbalanced despite the 3+1 split, but also because the time derivative is not  $\mathbf{d}_{\bar{t}}$  but rather  $\mathbf{dx} \wedge \mathbf{d}_{\bar{t}}$ .

Both  $\Gamma_{\bar{t}}$  and  $\mathbf{dx}_{\bar{t}}$  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  $\mathbf{dx}_{\bar{t}}$  can be adjusted arbitrarily by a coordinate transformation. Thus the Gaussian and Hamiltonian constraint equations (16.290) can be interpreted as representing conserved Noether charges. The spin angular-momentum  $\overset{**}{\Sigma}$  on the right hand side of the Gaussian constraint equation (16.290a) satisfies the conservation law (16.261). The energy-momentum  $\overset{**}{\mathbf{T}}$  on the right hand side of the Hamiltonian constraint equation (16.290b) satisfies the conservation law (16.279). The left hand sides of the Gaussian and Hamiltonian constraints satisfy corresponding conservation laws enforced by the contracted Bianchi identities (16.379). 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

Split into time and spatial components, the gravitational Lagrangian 4-form (16.218) is

$$L_g = \frac{I}{8\pi} (\mathbf{d}^2\mathbf{x} \wedge (\mathbf{d}_{\bar{t}}\Gamma + \mathbf{d}\Gamma_{\bar{t}} + \frac{1}{2}[\Gamma, \Gamma_{\bar{t}}]) - \mathbf{dx}_{\bar{t}} \wedge \mathbf{dx} \wedge \mathbf{R}). \quad (16.292)$$

The  $\mathbf{d}^2\mathbf{x} \wedge \mathbf{d}_{\bar{t}}\Gamma$  term in the Lagrangian (16.292) indicates that the momentum conjugate to the 18-component spatial connection  $\Gamma$  is the 18-component spatial area element  $\mathbf{d}^2\mathbf{x}$ . But, as discussed in the §16.15.1 above, the spatial area element has excess degrees of freedom compared to the 12-component line interval  $\mathbf{dx}$ . The fix adopted in §16.15.1 was to regard the spatial line interval  $\mathbf{dx}$  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 (16.289b) would be replaced by an equation for  $\mathbf{R}$  in place of  $\mathbf{dx} \wedge \mathbf{R}$ , and the

result would not be general relativity, contradicting observation and experiment. To treat  $\Gamma$  and  $d\mathbf{x}$  as conjugate variables, the  $d^2\mathbf{x} \wedge d_{\bar{t}}\Gamma$  term may be rewritten

$$d^2\mathbf{x} \wedge d_{\bar{t}}\Gamma = -\frac{1}{2} d\mathbf{x} \wedge (d\mathbf{x} \wedge d_{\bar{t}}\Gamma). \quad (16.293)$$

Equation (16.293) effectively replaces the time derivative  $d_{\bar{t}}$  with  $d\mathbf{x} \wedge d_{\bar{t}}$ , consistent with the time derivative in the equations of motion (16.291). The remaining terms in the gravitational Lagrangian (16.292) rearrange as follows. The  $d\Gamma_{\bar{t}}$  term integrates by parts to

$$d^2\mathbf{x} \wedge d\Gamma_{\bar{t}} = d(d^2\mathbf{x} \wedge \Gamma_{\bar{t}}) - (dd^2\mathbf{x}) \wedge \Gamma_{\bar{t}}. \quad (16.294)$$

The  $\frac{1}{2}[\Gamma, \Gamma_{\bar{t}}]$  term rearranges by the multivector triple-product relation (13.37) to

$$\frac{1}{2} d^2\mathbf{x} \wedge [\Gamma, \Gamma_{\bar{t}}] = \frac{1}{2} [d^2\mathbf{x}, \Gamma] \wedge \Gamma_{\bar{t}} = -\frac{1}{2} [\Gamma, d^2\mathbf{x}] \wedge \Gamma_{\bar{t}}. \quad (16.295)$$

The coefficients of the  $\wedge \Gamma_{\bar{t}}$  terms in equations (16.294) and (16.295) are

$$-(dd^2\mathbf{x}) - \frac{1}{2}[\Gamma, d^2\mathbf{x}] = -d\mathbf{x} \wedge S, \quad (16.296)$$

where  $S$  is the torsion defined by equation (16.207). The manipulations (16.293)–(16.296) bring the gravitational action to

$$S_g = \frac{I}{8\pi} \oint_{t_i}^{t_f} d^2\mathbf{x} \wedge \Gamma_{\bar{t}} + \frac{I}{8\pi} \int -\frac{1}{2} d\mathbf{x} \wedge (d\mathbf{x} \wedge d_{\bar{t}}\Gamma) - (d\mathbf{x} \wedge S) \wedge \Gamma_{\bar{t}} - d\mathbf{x}_{\bar{t}} \wedge (d\mathbf{x} \wedge R). \quad (16.297)$$

With the surface term discarded, the gravitational action (16.297) is in conventional Hamiltonian form  $L_g = I(\mathbf{p} \wedge (d\mathbf{x} \wedge d_{\bar{t}}\mathbf{q})) - H_g$  with coordinates  $\mathbf{q} \equiv \Gamma$  and momenta  $\mathbf{p} \equiv d\mathbf{x}/(16\pi)$ , and a somewhat strange time derivative  $d\mathbf{x} \wedge d_{\bar{t}}$ . The conventional (not super-) Hamiltonian 4-form  $H_g$  is

$$H_g = \frac{I}{8\pi} ((d\mathbf{x} \wedge S) \wedge \Gamma_{\bar{t}} + d\mathbf{x}_{\bar{t}} \wedge (d\mathbf{x} \wedge R)). \quad (16.298)$$

The conventional Hamiltonian (16.298) is a sum of the Gaussian and Hamiltonian constraints (16.290) wedged with the gauge variables  $\Gamma_{\bar{t}}$  and  $d\mathbf{x}_{\bar{t}}$ .

The Hamiltonian (16.298) is fine as a conventional Hamiltonian in which the coordinates and momenta are the 18-component spatial connection  $\Gamma$  and the 12-component spatial line interval  $d\mathbf{x}$ . But the Hamiltonian cannot be satisfactory because it yields only 12 equations of motion (16.291b) for the 18 components of  $\Gamma$ , and because the time derivative in those equations is  $d\mathbf{x} \wedge d_{\bar{t}}$  rather than  $d_{\bar{t}}$ . Ultimately, these problems stem from the fact that there remain redundant degrees of freedom in  $\Gamma$  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 (16.237) yields a more promising result: a balanced set of equations of motion, and a time derivative that is just  $d_{\bar{t}}$  as opposed to  $d\mathbf{x} \wedge d_{\bar{t}}$ . In the alternative Lagrangian, the gravitational coordinates are the line interval  $d\mathbf{x}$ , and their conjugate momenta are  $\pi$  defined by equation (16.238). After the 3+1 split, the coordinates are the spatial components of  $d\mathbf{x}$ , which is a vector 1-form

with  $4 \times 3 = 12$  components, while the momenta are the spatial components of  $\boldsymbol{\pi}$ , which is a trivector 2-form also with  $4 \times 3 = 12$  components.

Once again, the equations of motion (16.300) and constraints and identities (16.304) follow directly from splitting equations (16.246) into time and space parts, but they can be derived more fundamentally by splitting the variation  $\delta S$  of the action into time and space parts. Splitting the variation  $\delta S_g$  of the alternative gravitational action (16.241) into time and space parts gives

$$\delta S'_g = \frac{I}{8\pi} \oint_{t_i}^{t_f} (\delta \mathbf{dx} \wedge \boldsymbol{\pi})_{\bar{t}} + \frac{I}{8\pi} \left[ \oint \delta \mathbf{dx} \wedge \boldsymbol{\pi} \right]_{t_i}^{t_f} + \frac{I}{8\pi} \int \mathbf{S}_{\bar{t}} \wedge \delta \boldsymbol{\pi} + \delta \mathbf{dx}_{\bar{t}} \wedge \boldsymbol{\Pi} + \mathbf{S} \wedge \delta \boldsymbol{\pi}_{\bar{t}} + \delta \mathbf{dx} \wedge \boldsymbol{\Pi}_{\bar{t}} . \quad (16.299)$$

Variation of the combined gravitational and matter actions with respect to the variations  $\delta \mathbf{dx}$  and  $\delta \boldsymbol{\pi}$  of the spatial coordinates and momenta yields  $12 + 12 = 24$  equations of motion involving time derivatives,

$$12 \text{ equations: } \mathbf{S}_{\bar{t}} = 8\pi \tilde{\boldsymbol{\Sigma}}_{\bar{t}} , \quad (16.300a)$$

$$12 \text{ equations: } \boldsymbol{\Pi}_{\bar{t}} = 8\pi \tilde{\mathbf{T}}_{\bar{t}} . \quad (16.300b)$$

Variation of the action with respect to the variations  $\delta \mathbf{dx}_{\bar{t}}$  and  $\delta \boldsymbol{\pi}_{\bar{t}}$  of the time components of the coordinates and momenta yields 6 identities and 10 constraint equations involving only spatial derivatives,

$$6 \text{ Gaussian constraints and 6 identities: } \mathbf{S} = 8\pi \tilde{\boldsymbol{\Sigma}} , \quad (16.301a)$$

$$4 \text{ Hamiltonian constraints: } \boldsymbol{\Pi} = 8\pi \tilde{\mathbf{T}} . \quad (16.301b)$$

The Gaussian constraints are the subset of equations (16.301a) comprising

$$6 \text{ Gaussian constraints: } \mathbf{dx} \wedge \mathbf{S} = 8\pi (\mathbf{dx} \wedge \tilde{\boldsymbol{\Sigma}}) . \quad (16.302)$$

More explicitly, the equations of motion (16.300) are

$$12 \text{ equations: } d_{\bar{t}} \mathbf{dx} + \frac{1}{2} [\boldsymbol{\Gamma}_{\bar{t}}, \mathbf{dx}] + d \mathbf{dx}_{\bar{t}} + \frac{1}{2} [\boldsymbol{\Gamma}, \mathbf{dx}_{\bar{t}}] = 8\pi \tilde{\boldsymbol{\Sigma}}_{\bar{t}} , \quad (16.303a)$$

$$12 \text{ equations: } d_{\bar{t}} \boldsymbol{\pi} + \frac{1}{2} [\boldsymbol{\Gamma}_{\bar{t}}, \boldsymbol{\pi}] + d \boldsymbol{\pi}_{\bar{t}} + \frac{1}{2} [\boldsymbol{\Gamma}, \boldsymbol{\pi}_{\bar{t}}] + \frac{1}{4} (\mathbf{dx} \wedge [\boldsymbol{\Gamma}, \boldsymbol{\Gamma}])_{\bar{t}} = 8\pi \tilde{\mathbf{T}}_{\bar{t}} , \quad (16.303b)$$

and the constraints and identities (16.301) are

$$6 \text{ Gaussian constraints and 6 identities: } d \mathbf{dx} + \frac{1}{2} [\boldsymbol{\Gamma}, \mathbf{dx}] = 8\pi \tilde{\boldsymbol{\Sigma}} , \quad (16.304a)$$

$$4 \text{ Hamiltonian constraints: } d \boldsymbol{\pi} + \frac{1}{2} [\boldsymbol{\Gamma}, \boldsymbol{\pi}] + \frac{1}{4} \mathbf{dx} \wedge [\boldsymbol{\Gamma}, \boldsymbol{\Gamma}] = 8\pi \tilde{\mathbf{T}} . \quad (16.304b)$$

Equations (16.303a) comprise 12 equations of motion for the 12 coordinates  $\mathbf{dx}$ , while equations (16.303b) comprise 12 equations of motion for the 12 momenta  $\boldsymbol{\pi}$ . The equations of motion (16.303) do not suffer from the peculiarities of the earlier equations of motion (16.291): the time evolution operator is  $d_{\bar{t}} \equiv dt \partial/\partial t$ ; and the number of equations of motion matches the number of dynamical variables.

In numerical calculations, the 6 Gaussian constraints and 6 identities packaged together in the 12 equations (16.301a) must be separated out, since the constraints can be dropped after being imposed in the initial conditions, while the identities must be imposed at every time step. As discussed in §16.15.8, equations (16.318) and (16.319), the Gaussian constraint constrains a set of 6 linear combinations of the 12

quantities  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$ . The null space of the 6 linear combinations defines another set of 6 linear combinations of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$ . The 6 identities are equations governing this null space of combinations of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$ .

The 12 + 12 gravitational equations of motion (16.300) together with their 10 constraints and 6 identities (16.301) may be compared to the 3 + 3 electromagnetic equations of motion (16.76) together with their 1 constraint and 3 identities (16.77).

#### 16.15.4 Alternative conventional Hamiltonian

Splitting the alternative Lagrangian  $L'_g$ , equation (16.237), into time and space components, and rearranging along lines similar to those leading to the gravitational action (16.297), brings the alternative gravitational action  $S'_g$  to

$$S'_g = \frac{I}{8\pi} \oint_{t_i}^{t_f} d\mathbf{x}_{\bar{t}} \wedge \boldsymbol{\pi} + \frac{I}{8\pi} \int d\bar{t} \mathbf{dx} \wedge \boldsymbol{\pi} + \mathbf{S} \wedge \boldsymbol{\pi}_{\bar{t}} + \mathbf{dx}_{\bar{t}} \wedge \boldsymbol{\Pi} . \quad (16.305)$$

With the surface term discarded, the gravitational action (16.305) is in conventional Hamiltonian form with coordinates  $\mathbf{dx}$  and momenta  $\boldsymbol{\pi}/(8\pi)$ . The alternative conventional gravitational Hamiltonian  $H'_g$  is

$$H'_g = -\frac{I}{8\pi} (\mathbf{S} \wedge \boldsymbol{\pi}_{\bar{t}} + \mathbf{dx}_{\bar{t}} \wedge \boldsymbol{\Pi}) . \quad (16.306)$$

The alternative conventional Hamiltonian (16.306) is a sum of identities and constraints, equations (16.301), wedged with time components  $\mathbf{dx}_{\bar{t}}$  and  $\boldsymbol{\pi}_{\bar{t}}$  of the coordinates and momenta. Unlike the standard conventional Hamiltonian (16.298), the multipliers are not all gauge variables. The 4-component multiplier  $\mathbf{dx}_{\bar{t}}$  can be considered a gauge variable as before, being arbitrarily adjustable by a coordinate transformation, and its partner  $\boldsymbol{\Pi}$  satisfies a constraint equation, the Hamiltonian constraint. But only 6 degrees of freedom of the 12-component multiplier  $\boldsymbol{\pi}_{\bar{t}}$  can be adjusted by a gauge (Lorentz) transformation. The remaining 6 degrees of freedom of  $\boldsymbol{\pi}_{\bar{t}}$  must be considered as constituting an auxiliary field, analogous to the magnetic components of the electromagnetic field, variation with respect to which effectively determines the components of the auxiliary field itself.

In contrast to the conventional Hamiltonian (16.298), the alternative conventional Hamiltonian (16.306) accomplishes the goal of a balanced number, 12 each, of coordinates and momenta.

#### 16.15.5 ADM gauge condition

Section 16.15.1 rejected the possibility of treating the 18-component spatial connection  $\boldsymbol{\Gamma}$  as the gravitational coordinates, and the 18-component spatial area element  $d^2\mathbf{x}$  as their conjugate momenta, on the grounds that the area element contains excess degrees of freedom compared to the 12-component spatial line interval  $d\mathbf{x}$ . However, the idea of working with  $\boldsymbol{\Gamma}$  and  $d^2\mathbf{x}$ , as opposed to  $d\mathbf{x}$  and  $\boldsymbol{\pi}$ , is attractive, firstly because in Yang-Mills theories such as electromagnetism the coordinates are the connections, and  $\boldsymbol{\Gamma}$  are the (Lorentz) connections of gravity, and secondly because black hole thermodynamics points to area as the thing that is quantized in general relativity.

One way to reduce the excess degrees of freedom in the area element is to impose gauge conditions on the line interval  $\mathbf{dx}$ . A simple strategy is to eliminate the gauge freedom of the vierbein under Lorentz boosts by imposing the 3 ADM gauge conditions

$$e_{0\alpha} = 0 . \quad (16.307)$$

The gauge choice (16.307) is the starting point of the ADM formalism, Chapter 17, and is carried over into the BSSN formalism, §17.7. The gauge choice (16.307) is also a basic ingredient of Loop Quantum Gravity, §16.16. The ADM gauge condition (16.307) reduces the number of degrees of freedom of the spatial line interval  $\mathbf{dx}$  from 12 to  $3 \times 3 = 9$ , and of the spatial area element  $\mathbf{d}^2x$  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 (16.307) are invertibly related to each other.

To explore the consequences of the ADM gauge condition (16.307), it is convenient to extend the 3+1 form-splitting notation (16.284) to a double 3+1 split in which both coordinate indices and tetrad (Lorentz) indices are split out. Thus a multivector form  $\mathbf{a}$  splits into 4 components  $\mathbf{a}_{\bar{0}\bar{t}}$ ,  $\mathbf{a}_{\bar{0}}$ ,  $\mathbf{a}_{\bar{t}}$ , and  $\mathbf{a}$  that represent respectively the time-time, time-space, space-time, and space-space components of the multivector form,

$$\mathbf{a} \rightarrow \mathbf{a}_{\bar{0}\bar{t}} + \mathbf{a}_{\bar{0}} + \mathbf{a}_{\bar{t}} + \mathbf{a} \equiv a_{0A\bar{t}\Lambda} \boldsymbol{\gamma}^0 \wedge \boldsymbol{\gamma}^A d^p x^{\bar{t}\Lambda} + a_{\bar{0}A\Lambda} \boldsymbol{\gamma}^0 \wedge \boldsymbol{\gamma}^A d^p x^{\Lambda} + a_{A\bar{t}\Lambda} \boldsymbol{\gamma}^A d^p x^{\bar{t}\Lambda} + a_{A\Lambda} \boldsymbol{\gamma}^A d^p x^{\Lambda} , \quad (16.308)$$

implicitly summed over distinct antisymmetric sequences of tetrad and coordinate indices  $A$  and  $\Lambda$ . In this notation, the ADM gauge condition (16.307) is

$$\mathbf{d}\mathbf{x}_{\bar{0}} = 0 . \quad (16.309)$$

The spatial area element  $\mathbf{d}^2x$  is the 9-component bivector 2-form

$$\mathbf{d}^2x = -\frac{1}{2} \mathbf{dx} \wedge \mathbf{dx} = -2 e_{a\alpha} e_{b\beta} \boldsymbol{\gamma}^a \wedge \boldsymbol{\gamma}^b d^2x^{\alpha\beta} . \quad (16.310)$$

implicitly summed over distinct antisymmetric sets of spatial indices  $ab$  and  $\alpha\beta$ . The momenta conjugate to the spatial area element  $\mathbf{d}^2x$  are the  $3 \times 3 = 9$  components of the spatial Lorentz connections  $\boldsymbol{\Gamma}_{\bar{0}}$  with one Lorentz index the tetrad time index 0, also called the extrinsic curvature, §17.1.4,

$$\boldsymbol{\Gamma}_{\bar{0}} = \Gamma_{0a\alpha} \boldsymbol{\gamma}^0 \wedge \boldsymbol{\gamma}^a dx^{\alpha} . \quad (16.311)$$

Double-splitting the variation  $\delta S_g$  of the gravitational action, equation (16.231) into time and space parts gives

$$\begin{aligned} \delta S_g &= \frac{I}{8\pi} \oint_{t_i}^{t_f} (\mathbf{d}^2x \wedge \delta \boldsymbol{\Gamma})_{\bar{0}\bar{t}} + \left[ \oint (\mathbf{d}^2x \wedge \delta \boldsymbol{\Gamma})_{\bar{0}} \right]_{t_i}^{t_f} \\ &+ \frac{I}{8\pi} \int -(\mathbf{dx} \wedge \mathbf{S})_{\bar{0}\bar{t}} \wedge \delta \boldsymbol{\Gamma} - (\mathbf{dx} \wedge \mathbf{S})_{\bar{t}} \wedge \delta \boldsymbol{\Gamma}_{\bar{0}} - (\mathbf{dx} \wedge \mathbf{S})_{\bar{0}} \wedge \delta \boldsymbol{\Gamma}_{\bar{t}} - (\mathbf{dx} \wedge \mathbf{S}) \wedge \delta \boldsymbol{\Gamma}_{\bar{0}\bar{t}} \\ &\quad - \delta \mathbf{dx} \wedge (\mathbf{dx} \wedge \mathbf{R})_{\bar{0}\bar{t}} - \delta \mathbf{dx}_{\bar{0}} \wedge (\mathbf{dx} \wedge \mathbf{R})_{\bar{t}} - \delta \mathbf{dx}_{\bar{t}} \wedge (\mathbf{dx} \wedge \mathbf{R})_{\bar{0}} - \delta \mathbf{dx}_{\bar{0}\bar{t}} \wedge (\mathbf{dx} \wedge \mathbf{R}) . \end{aligned} \quad (16.312)$$

Variation of the combined gravitational and matter actions with respect to the variations  $\delta \boldsymbol{\Gamma}_{\bar{0}}$  and  $\delta \mathbf{dx}$  yields

$9 + 9 = 18$  equations of motion for the area element  $\mathbf{d}^2\mathbf{x}$  and their conjugate momenta  $\mathbf{\Gamma}_{\bar{0}}$ ,

$$9 \text{ equations: } (\mathbf{dx} \wedge \mathbf{S})_{\bar{t}} = -8\pi \mathbf{\Sigma}_{\bar{t}}^{**}, \quad (16.313a)$$

$$9 \text{ equations: } (\mathbf{dx} \wedge \mathbf{R}_{\bar{0}})_{\bar{t}} = -8\pi \mathbf{T}_{\bar{0}\bar{t}}^{**}, \quad (16.313b)$$

while variation with respect to  $\delta\mathbf{dx}_{\bar{0}}$  yields the 3 equations of motion

$$3 \text{ equations: } (\mathbf{dx} \wedge \mathbf{R})_{\bar{t}} = -8\pi \mathbf{T}_{\bar{t}}^{**}. \quad (16.314)$$

Note that the ADM gauge condition  $\mathbf{dx}_{\bar{0}} = 0$  is a gauge condition, fixed after equations of motion are derived, so it is correct to vary  $\mathbf{dx}_{\bar{0}}$  in the action, leading to equation (16.314). Explicitly, the equations of motion (16.313) and (16.314) are, similarly to equations (16.291),

$$9 \text{ equations: } d_{\bar{t}}\mathbf{d}^2\mathbf{x} + \frac{1}{2}[\mathbf{\Gamma}_{\bar{t}}, \mathbf{d}^2\mathbf{x}] + \mathbf{dx} \wedge (dd\mathbf{x}_{\bar{t}} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{dx}_{\bar{t}}]) + \mathbf{dx}_{\bar{t}} \wedge \mathbf{S} = -8\pi \mathbf{\Sigma}_{\bar{t}}^{**}, \quad (16.315a)$$

$$9 \text{ equations: } \mathbf{dx} \wedge (d_{\bar{t}}\mathbf{\Gamma}_{\bar{0}} + \frac{1}{2}[\mathbf{\Gamma}_{\bar{t}}, \mathbf{\Gamma}_{\bar{0}}] + d\mathbf{\Gamma}_{\bar{0}\bar{t}} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\Gamma}_{\bar{0}\bar{t}}]) + \mathbf{dx}_{\bar{t}} \wedge \mathbf{R}_{\bar{0}} = -8\pi \mathbf{T}_{\bar{0}\bar{t}}^{**}, \quad (16.315b)$$

$$3 \text{ equations: } \mathbf{dx} \wedge (d_{\bar{t}}\mathbf{\Gamma} + d\mathbf{\Gamma}_{\bar{t}} + \frac{1}{2}[\mathbf{\Gamma}, \mathbf{\Gamma}_{\bar{t}}]) + \mathbf{dx}_{\bar{t}} \wedge \mathbf{R} = -8\pi \mathbf{T}_{\bar{t}}^{**}. \quad (16.315c)$$

Equation (16.314) is an equation of motion in the sense that it involves a time derivative  $d_{\bar{t}}\mathbf{\Gamma}$ ; but  $\mathbf{\Gamma}$  is not one of the momenta  $\mathbf{\Gamma}_{\bar{0}}$  conjugate to the area element  $\mathbf{d}^2\mathbf{x}$ , so equation (16.314) has a different status from the  $9 + 9$  equations of motion (16.313). In the ADM formalism, §16.15.6, equation (16.314) is discarded as redundant with the 3 momentum constraints (16.316d), on the grounds that the energy-momentum tensor is symmetric (for vanishing torsion). The BSSN formalism on the other hand, §16.15.7, retains equation (16.314) as a distinct equation of motion.

The earlier equations (16.291) had the problem that the time derivative in the equation of motion for  $\mathbf{\Gamma}$  was  $\mathbf{dx} \wedge d_{\bar{t}}$  as opposed to just  $d_{\bar{t}}$ . Equation (16.315b) seems to have the same difficulty, but here it is no longer a problem, because the 9-component trivector 3-form  $\mathbf{dx} \wedge d_{\bar{t}}\mathbf{\Gamma}_{\bar{0}}$  is invertibly related to the 9-component bivector 2-form  $d_{\bar{t}}\mathbf{\Gamma}_{\bar{0}}$ , so equation (16.315b) can be rearranged as an equation for  $d_{\bar{t}}\mathbf{\Gamma}_{\bar{0}}$ .

Variation of the action with respect to  $\delta\mathbf{\Gamma}$ ,  $\delta\mathbf{\Gamma}_{\bar{t}}$ ,  $\delta\mathbf{\Gamma}_{\bar{0}\bar{t}}$ ,  $\delta\mathbf{dx}_{\bar{t}}$  and  $\delta\mathbf{dx}_{\bar{0}\bar{t}}$  yields 9 identities and 10 constraints involving only spatial derivatives

$$9 \text{ identities: } (\mathbf{dx} \wedge \mathbf{S}_{\bar{0}})_{\bar{t}} = -8\pi \mathbf{\Sigma}_{\bar{0}\bar{t}}^{**}, \quad (16.316a)$$

$$3 \text{ Gaussian constraints: } \mathbf{dx} \wedge \mathbf{S}_{\bar{0}} = -8\pi \mathbf{\Sigma}_{\bar{0}}^{**}, \quad (16.316b)$$

$$3 \text{ Gaussian constraints: } \mathbf{dx} \wedge \mathbf{S} = -8\pi \mathbf{\Sigma}^{**}, \quad (16.316c)$$

$$3 \text{ momentum constraints: } \mathbf{dx} \wedge \mathbf{R}_{\bar{0}} = -8\pi \mathbf{T}_{\bar{0}}^{**}, \quad (16.316d)$$

$$1 \text{ Hamiltonian constraint: } \mathbf{dx} \wedge \mathbf{R} = -8\pi \mathbf{T}^{**}. \quad (16.316e)$$

The 9 identities (16.316a) are not equations of motion (they involve no time derivatives), despite having a form index  $\bar{t}$ . Explicitly,

$$9 \text{ identities: } \frac{1}{2}[\mathbf{\Gamma}_{\bar{0}\bar{t}}, \mathbf{d}^2\mathbf{x}] + \frac{1}{2}[\mathbf{\Gamma}_{\bar{0}}, \mathbf{d}^2\mathbf{x}_{\bar{t}}] = -8\pi \mathbf{\Sigma}_{\bar{0}\bar{t}}^{**}. \quad (16.317)$$

### 16.15.6 ADM formalism

The ADM formalism is pursued at length in Chapter 17 in traditional (coordinate and tetrad) index notation. However, it is useful to offer a few comments on the ADM formalism in the present context of multivector forms notation.

In addition to invoking the ADM gauge condition (16.307), the ADM formalism assumes from the outset that torsion vanishes, as indeed general relativity assumes. One of the consequences of vanishing torsion is that the energy-momentum tensor is symmetric, equation (16.262). This motivates the ADM strategy of simply discarding the 6 antisymmetric components of the Einstein equations. These 6 components comprise the 3 antisymmetric components of the 9 equations of motion (16.313b), corresponding to the 3 antisymmetric space-space Einstein equations, and the 3 equations of motion (16.314), corresponding to the 3 antisymmetric time-space Einstein equations. Discarding the antisymmetric Einstein's equations seems innocent enough, until one realises that the antisymmetry of the energy-momentum is a consequence of the law of conservation of spin angular-momentum, equation (16.261), which law is responsible for the Gaussian constraints (16.316b) and (16.316c). 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. This should raise a red flag: it may not be a good numerical strategy to trade hyperbolic equations of motion for elliptic 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 (16.307) removes the gauge freedom of Lorentz boosts). In particular, the 9 spatial vierbein  $e_{\alpha\alpha}$  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  $\Gamma_{\alpha\beta\gamma}$  together with the  $3 \times 6 = 18$  torsion-free coordinate-frame spatial connections (Christoffel symbols)  $\Gamma_{\alpha\beta\gamma}$ . In summary, in the ADM formalism there are  $6 + 6 = 12$  equations of motion for  $g_{\alpha\beta}$  and  $\Gamma_{\alpha\beta\gamma}$ , 18 identities for  $\Gamma_{\alpha\beta\gamma}$ , and 4 Hamiltonian constraints.

### 16.15.7 BSSN formalism

The BSSN formalism, discussed further in Chapter 17, §17.7, has gained popularity because it proves to have better numerical stability when applied to problems such as the merger of two black holes (Shibata and Nakamura, 1995; Baumgarte and Shapiro, 1999).

The BSSN formalism follows ADM for the most part, in particular imposing the ADM gauge choice (16.307). However, BSSN retains the 3 momentum Einstein equations (16.314) as equations of motion, and restores the 3 Gaussian constraints (16.316c). Like the Hamiltonian constraints, the Gaussian constraints must be arranged to be satisfied in the initial conditions, but may be discarded thereafter.

### 16.15.8 An alternative strategy for solving the gravitational equations numerically

A drawback of the ADM and BSSN approaches is that expending 3 of the Lorentz gauge conditions on the vierbein, equation (16.307), preempts being able to use those Lorentz gauge conditions in potentially

advantageous ways on the Lorentz connections, as is done for example in proofs of singularity theorems (conditions (18.26), for example). Moreover, as discussed in §16.15.6, transferring Lorentz gauge conditions from the Lorentz connections to the vierbein has the side-effect of replacing “nice” hyperbolic equations of motion with “nasty” elliptic identities.

This suggests that better numerical behaviour could result from using the Hamiltonian system gravitational equations in the native form derived in §16.15.3, where they consist of 12 + 12 equations of motion, 6 + 4 constraints, and 6 identities, equations (16.300) and (16.301) (as opposed to the 6 + 6 equations of motion, 4 constraints, and 18 identities of the ADM formalism). As remarked previously, the system of gravitational equations derived in §16.15.3 resembles the Hamiltonian system of electromagnetic equations, which consist of 3 + 3 equations of motion, 1 constraint, and 3 identities, equations (16.76) and (16.77).

Consider briefly the problem of solving the system of gravitational equations (16.300) and (16.301) numerically. First, all 40 equations must be arranged to be satisfied on an initial hypersurface of constant time. Subsequently, the 10 constraint equations may be discarded (or kept as a check of numerical accuracy). In integrating the remaining system of 30 equations from one hypersurface of constant time to the next, the first step is to use the 12 + 12 equations of motion to find the 12 gravitational coordinates  $\mathbf{dx}$  and their 12 conjugate momenta  $\boldsymbol{\pi}$  on the next (upper) hypersurface. As in ADM, coordinate gauge freedom allows the time components  $\mathbf{dx}_{\bar{t}}$  of the vierbein (the lapse and shift) to be chosen arbitrarily. The 12 components of  $\mathbf{dx}$  and the 4 components of  $\mathbf{dx}_{\bar{t}}$  account for all 16 components of the vierbein. Lorentz gauge freedom allows the time components  $\boldsymbol{\Gamma}_{\bar{t}}$  of the Lorentz connection to be chosen arbitrarily. The 12 components of  $\boldsymbol{\pi}$  and the 6 components of  $\boldsymbol{\Gamma}_{\bar{t}}$  account for 18 of the 24 degrees of freedom of the Lorentz connection. The remaining 6 degrees of freedom of the Lorentz connection are determined by the 6 identities.

The 6 identities are packaged together with the 6 Gaussian constraints in the 12 equations (16.304a). Now the 12 equations (16.304a) comprise an expression for the 12 components of the spatial vector 2-form  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$  in terms of spatial exterior derivatives  $d\mathbf{dx}$  of the spatial vierbein and the spin angular-momentum (the latter vanishing if torsion is assumed to vanish). Since  $d\mathbf{dx}$  over the upper spatial hypersurface of constant time is already in hand from integrating the equations of motion, taking the spatial exterior derivative  $d\mathbf{dx}$  is a matter of differentiating numerically over the spatial hypersurface. The 12 components of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$  are

$$\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}] = \boldsymbol{\Gamma} \cdot \mathbf{dx} = -2 \Gamma_{k[\alpha\beta]} \boldsymbol{\gamma}^k d^2x^{\alpha\beta}. \quad (16.318)$$

The Gaussian constraints involve the 6 components of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$  comprising

$$\mathbf{dx} \wedge \frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}] = -12 e_{k\alpha} \Gamma_{l[\beta\gamma]} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l d^3x^{\alpha\beta\gamma}, \quad (16.319)$$

summed over distinct antisymmetric sequences of indices  $kl$  and  $\alpha\beta\gamma$ . The components of equation (16.319) may be regarded as a  $6 \times 12$  matrix acting on a 12-component vector consisting of the  $\Gamma_{k[\alpha\beta]}$ . The null space of the  $6 \times 12$  matrix is another  $6 \times 12$  matrix which, when acting on the 12-component vector  $\Gamma_{k[\alpha\beta]}$ , gives the 6 linear combinations of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$  whose values are determined by the 6 identities distinct from the 6 Gaussian constraints. Actually, it is not essential to choose the 6 identities to lie in the 6-dimensional null space; it suffices that the 6 identities span the null space. That is, the 6 identities and 6 Gaussian constraints should together span the full 12-dimensional space of  $\frac{1}{2}[\boldsymbol{\Gamma}, \mathbf{dx}]$ . Equivalently, no linear combination of the 6 identities should be a constraint.

## 16.16 Loop Quantum Gravity

### 16.16.1 Ashtekar variables

Loop quantum gravity (LQG) was instigated by Ashtekar (1987), who was inspired by the fact that in 4 dimensions bivectors have a natural complex structure. The complex structure allows bivectors to be split into distinct right- and left-handed chiral parts. Selecting one of the two chiral parts of each of the spatial Lorentz connection  $\Gamma$  and the spatial area element  $d^2x$ , both bivectors, halves the number of degrees of freedom of each from 18 to 9. The 9 components of the chiral spatial area element are invertibly related to the 9 components of the spatial line interval  $dx$  subject to the ADM condition (16.307). Thus the ADM analysis of §16.15.5 carries through, with the difference that the elimination of 9 degrees of freedom in the spatial connection eliminates the need for the 9 identities (16.316a). It was hoped that the resulting simplification of the gravitational equations would facilitate quantization. The hope was not realised, but Ashtekar's approach rekindled interest in the prospect of quantizing gravity by traditional methods of quantum field theory.

The complex structure of bivectors in 4 dimensions arises because, in 4 dimensions, the pseudoscalar  $I$  times a bivector is a bivector. The complex structure allows bivectors to be split into two distinct complex chiral parts. Thus the Lorentz connection  $\Gamma$  splits into right-handed  $\Gamma_+$  and left-handed  $\Gamma_-$  chiral parts,

$$\Gamma = \frac{1}{2}(\Gamma_+ + \Gamma_-), \quad \Gamma_\pm \equiv (1 \pm iI)\Gamma, \quad (16.320)$$

Here  $I$  is the pseudoscalar of the spacetime algebra, and  $i$  is the commuting imaginary of quantum mechanics. The chiral factors are, modulo a factor of  $\frac{1}{2}$ , projection operators  $P_\pm \equiv \frac{1}{2}(1 \pm iI)$  satisfying  $P_\pm^2 = P_\pm$ , and with zero overlap,  $P_+P_- = 0$ . The Riemann curvature bivector  $R$  has corresponding right- and left-handed components  $R_\pm$ , satisfying the Cartan structure equation

$$R_\pm \equiv (1 \pm iI)R = d\Gamma_\pm + \frac{1}{4}[\Gamma_\pm, \Gamma_\pm]. \quad (16.321)$$

The area-element bivector  $d^2x$  similarly has right-and left-handed components,

$$d^2x_\pm \equiv (1 \pm iI)d^2x. \quad (16.322)$$

Without loss of generality, restrict to the right-handed (+) case. Consider the complex chiral gravitational Lagrangian 4-form

$$L_+ \equiv \frac{I}{8\pi} d^2x \wedge R_+ = \frac{I}{16\pi} d^2x_+ \wedge R_+ = \frac{I}{16\pi} d^2x_+ \wedge (d\Gamma_+ + \frac{1}{4}[\Gamma_+, \Gamma_+]). \quad (16.323)$$

An equivalent expression for the chiral Lagrangian (16.323) is (compare equation (16.216); recall that in the notation adopted here, explained at the beginning of §16.14, the dot product signifies a dot product of multivectors, but an exterior product of forms)

$$L_+ \equiv \frac{1}{8\pi} (*d^2x - i d^2x) \cdot R. \quad (16.324)$$

In components, the imaginary part of the chiral Lagrangian 4-form is

$$\text{Im } L_+ = -\frac{1}{8\pi} d^2x \cdot R = -\frac{6}{8\pi} R_{0123} d^4x^{0123}, \quad (16.325)$$

which involves the purely antisymmetric part  $R_{0123}$  of the Riemann tensor. The purely antisymmetric Riemann tensor vanishes if torsion vanishes. Note that for the imaginary contribution (16.325) to be non-trivial, it is necessary to regard torsion as a dynamical field that vanishes (or not) only as a result of the equations of motion. Thus if torsion vanishes, the Lagrangian (16.325) vanishes only *after* equations of motion are imposed. An explicit expression for the imaginary part of the chiral action in terms of the torsion  $\mathbf{S}$  is

$$\begin{aligned} -\frac{1}{8\pi} \int d^2\mathbf{x} \cdot \mathbf{R} &= \frac{1}{16\pi} \int d\mathbf{x} \cdot (d\mathbf{x} \cdot \mathbf{R}) = -\frac{1}{16\pi} \int d\mathbf{x} \cdot (d\mathbf{S} + \frac{1}{2}[\Gamma, \mathbf{S}]) \\ &= \frac{1}{16\pi} \oint d\mathbf{x} \cdot \mathbf{S} - \frac{1}{16\pi} \int \mathbf{S} \cdot \mathbf{S} , \end{aligned} \quad (16.326)$$

the third expression of which follows from the Bianchi identity (16.376a), and the final expression from an integration by parts.

Ashtekar (1987) proposed replacing the standard Hilbert Lagrangian with the chiral Lagrangian (16.323). The gravitational coordinates are nominally the right-handed connections  $\Gamma_+$ , a complex bivector 1-form with  $(6 \times 4)/2 = 12$  degrees of freedom, and the right-handed area element  $d^2\mathbf{x}_+$ , a complex bivector 2-form with  $(6 \times 6)/2 = 18$  degrees of freedom. However, the 18-component complex right-handed area element  $d^2\mathbf{x}_+$  has excess degrees of freedom compared to the 16-component real line interval  $d\mathbf{x}$ . Notwithstanding LQG's focus on the area element, LQG follows the standard philosophy that the line interval  $d\mathbf{x}$  has (before gauge-fixing) 16 degrees of freedom. Thus in varying the chiral action (16.323), the degrees of freedom are taken to be the 12 components of  $\Gamma_+$  and the 16 components of  $d\mathbf{x}$ . Varying the action with chiral Lagrangian (16.323) with respect to  $\Gamma_+$  and  $d\mathbf{x}$  yields (compare equation (16.231))

$$\delta S_+ = \frac{I}{16\pi} \oint d^2\mathbf{x}_+ \wedge \delta\Gamma_+ + \frac{I}{16\pi} \int -(\mathbf{dx} \wedge \mathbf{S})_+ \wedge \delta\Gamma_+ - 2\delta\mathbf{dx} \wedge (\mathbf{dx} \wedge \mathbf{R}_+) . \quad (16.327)$$

With matter included, Hamilton's equations are then (compare equations (16.234))

$$(\mathbf{dx} \wedge \mathbf{S})_+ = -8\pi \overset{\star\star}{\Sigma}_+ , \quad (16.328a)$$

$$\mathbf{dx} \wedge \mathbf{R}_+ = -8\pi \overset{\star\star}{T} . \quad (16.328b)$$

The torsion equation (16.328a) is a bivector 3-form with  $(6 \times 4)/2 = 12$  complex degrees of freedom, while the Einstein equation (16.328b) is a trivector 3-form with  $4 \times 4 = 16$  complex degrees of freedom. The imaginary part of the Einstein equation (16.328b) is

$$\text{Im}(\mathbf{dx} \wedge \mathbf{R}_+) = \mathbf{dx} \wedge (I\mathbf{R}) = I(\mathbf{dx} \cdot \mathbf{R}) = R_{[\kappa\lambda\mu]n} I\boldsymbol{\gamma}^n d^3x^{\kappa\lambda\mu} , \quad (16.329)$$

which vanishes provided that torsion vanishes. Thus the chiral Einstein equation (16.328b) is real and reproduces the usual Einstein equation provided that torsion vanishes. The torsion equation (16.328a) has 12 complex components, half the usual (non-chiral) number 24 of torsion equations, but 12 is the correct number to determine the 12 complex components of the chiral connection  $\Gamma_+$ .

The next step is to carry out a 3+1 split similar to that in §16.15.5. After the 3+1 split, the gravitational coordinates are the 9 components of the spatial chiral connection  $\Gamma_+$ , and their conjugate momenta are the

9 components of the spatial chiral area element  $\mathbf{d}^2\mathbf{x}_+$ . These coordinates and momenta are called **Ashtekar variables**,

$$\boldsymbol{\Gamma}_+, \quad \mathbf{d}^2\mathbf{x}_+ \quad \text{canonically conjugate Ashtekar variables .} \quad (16.330)$$

The remaining  $28 - 18 = 10$  degrees of freedom in the connection and line interval are all gauge degrees of freedom. The 6 Lorentz gauge freedoms can be used to fix the 3 tetrad-time components  $\mathbf{dx}_{\bar{0}}$  of the line interval and the 3 time components  $\boldsymbol{\Gamma}_{+\bar{t}}$  of the Lorentz connection. The 4 coordinate gauge freedoms can be used to fix the 4 time components  $\mathbf{dx}_{\bar{t}}$  of the line interval. The 9 components of the spatial chiral area element  $\mathbf{d}^2\mathbf{x}_+$  are invertibly related to the 9 components  $\mathbf{dx}$  of the line interval subject to the ADM gauge condition (16.307). In all, variation of the action with respect to  $\delta\boldsymbol{\Gamma}_+$  and  $\delta\mathbf{dx}$  yields  $9 + 9 = 18$  equations of motion (compare equations (16.313)),

$$9 \text{ equations: } (\mathbf{dx} \wedge \mathbf{S})_{+\bar{t}} = -8\pi \overset{**}{\Sigma}_{+\bar{t}}, \quad (16.331a)$$

$$9 \text{ equations: } (\mathbf{dx} \wedge \mathbf{R}_+)_{\bar{0}\bar{t}} = -8\pi \overset{**}{T}_{\bar{0}\bar{t}}, \quad (16.331b)$$

while variation with respect to the gauge variables  $\delta\boldsymbol{\Gamma}_{+\bar{t}}$ ,  $\delta\mathbf{dx}_{\bar{0}}$ ,  $\delta\mathbf{dx}_{\bar{t}}$  and  $\delta\mathbf{dx}_{\bar{0}\bar{t}}$  yields 10 constraints (compare equations (16.316)),

$$3 \text{ Gaussian constraints: } (\mathbf{dx} \wedge \mathbf{S})_+ = -8\pi \overset{**}{\Sigma}_+, \quad (16.332a)$$

$$3 \text{ Gaussian constraints: } (\mathbf{dx} \wedge \mathbf{R}_+)_{\bar{t}} = -8\pi \overset{**}{T}_{\bar{t}}. \quad (16.332b)$$

$$3 \text{ momentum constraints: } (\mathbf{dx} \wedge \mathbf{R}_+)_{\bar{0}} = -8\pi \overset{**}{T}_{\bar{0}}, \quad (16.332c)$$

$$1 \text{ Hamiltonian constraint: } \mathbf{dx} \wedge \mathbf{R}_+ = -8\pi \overset{**}{T}. \quad (16.332d)$$

The novel feature is that there are no identities, only equations of motion and constraints.

### 16.16.2 Loop Quantum Gravity

Ashtekar's proposal simplified the system of gravitational equations, but at the price of complexifying the connection  $\boldsymbol{\Gamma}$ . Complexification introduced other difficulties (the need to introduce additional "reality constraints") that obstructed successful quantization.

Meanwhile, it began to be realised, starting with Jacobson and Smolin (1988), that if gravity is treated as a gauge theory of Lorentz transformations, then a gauge rotation (spatial Lorentz transformation) by  $2\pi$  around a loop will have a quantum mechanical effect analogous to the Aharonov-Bohm effect of electromagnetism. This was the origin of Loop Quantum Gravity.

Eventually it was realised that the loop structure of LQG could be retained without the problems of complexification, by using a real version of the chiral action, the **Holst Lagrangian** (Holst, 1996),

$$L_\beta \equiv \frac{I}{8\pi} \mathbf{d}^2\mathbf{x} \wedge \mathbf{R}_\beta, \quad (16.333)$$

where

$$\mathbf{R}_\beta \equiv \left(1 + \frac{I}{\beta}\right) \mathbf{R} , \quad (16.334)$$

and  $\beta$  is an arbitrary constant called the **Barbero-Immirzi parameter** (Barbero G., 1995; Immirzi, 1997). The choice  $\beta = \infty$  gives the classic Hilbert action, while  $\beta = \mp 1$  recovers Ashtekar's chiral Lagrangian (16.323). Most later studies in LQG choose  $\beta$  to be real. The canonically conjugate variables are

$$\mathbf{\Gamma}_\beta \equiv \left(1 + \frac{I}{\beta}\right) \mathbf{\Gamma} , \quad d^2x . \quad (16.335)$$

The reader is referred to the arXiv for recent reviews of LQG.

**Exercise 16.12 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 alternative Hilbert gravitational Lagrangian?
5. What is the variation of the alternative gravitational Lagrangian?
6. What is the space+time  $(N-1)+1$  split of the alternative gravitational equations of motion?
7. What is the space+time  $(N-1)+1$  split of the gravitational equations of motion when the ADM gauge condition (16.307) is imposed?

#### Solution.

1. The Hilbert gravitational Lagrangian in  $N$  spacetime dimensions is the scalar  $N$ -form, generalizing equation (16.218),

$$L_g = \frac{I_N}{2S_{N-2}} d^{N-2}x \wedge \mathbf{R} , \quad (16.336)$$

where  $I_N$  is the  $N$ -dimensional spacetime pseudoscalar, and  $S_N$  is the area of an  $N$ -dimensional unit sphere, equation (16.356). The Lagrangian (16.336) is in super-Hamiltonian form

$$L_g = \frac{I_N}{2S_{N-2}} d^{N-2}x \wedge d\mathbf{\Gamma} - H_g , \quad (16.337)$$

with super-Hamiltonian, generalizing equation (16.220),

$$H_g = -\frac{I_N}{8S_{N-2}} d^{N-2}x \wedge [\mathbf{\Gamma}, \mathbf{\Gamma}] . \quad (16.338)$$

2. The variation of the gravitational action in  $N$  spacetime dimensions is, generalizing equation (16.231),

$$\delta S_g = \frac{(-)^N I_N}{2S_N} \oint d^{N-2}x \wedge \delta \mathbf{\Gamma} + \frac{(-)^N I_N}{2S_N} \int -(\mathbf{d}^{N-3}x \wedge \mathbf{S}) \wedge \delta \mathbf{\Gamma} - \delta dx \wedge (\mathbf{d}^{N-3}x \wedge \mathbf{R}) . \quad (16.339)$$

The variation of the matter action is defined by equations (16.232) in any spacetime dimension. With matter, the equations of motion generalizing equations (16.234) are

$$\frac{1}{2}N^2(N-1) \text{ equations } \mathbf{d}^{N-3}\mathbf{x} \wedge \mathbf{S} = (-)^{N-1} 2S_N \tilde{\Sigma}^*, \quad (16.340\text{a})$$

$$N^2 \text{ equations } \mathbf{d}^{N-3}\mathbf{x} \wedge \mathbf{R} = (-)^{N-1} 2S_N \tilde{\mathbf{T}}^*. \quad (16.340\text{b})$$

3. The alternative Hilbert gravitational Lagrangian in  $N$  spacetime dimensions is, generalizing equation (16.237),

$$L'_g = \frac{I_N}{2S_{N-2}} d\mathbf{dx} \wedge \boldsymbol{\pi} - H_g, \quad (16.341)$$

where  $\boldsymbol{\pi}$  is momentum pseudovector ( $N-2$ )-form, generalizing equation (16.238),

$$\boldsymbol{\pi} \equiv \mathbf{d}^{N-3}\mathbf{x} \wedge \boldsymbol{\Gamma}, \quad (16.342)$$

and  $H_g$  is the same super-Hamiltonian (16.338) as before.

4. The variation of the alternative action (16.341) is, generalizing equation (16.241)

$$\delta S'_g = \frac{I_N}{2S_N} \oint \delta \mathbf{dx} \wedge \boldsymbol{\pi} + \frac{I_N}{2S_N} \int \mathbf{S} \wedge \delta \boldsymbol{\pi} + \delta \mathbf{dx} \wedge \boldsymbol{\Pi}, \quad (16.343)$$

where the curvature pseudovector ( $N-1$ )-form  $\boldsymbol{\Pi}$  is, generalizing equations (16.242),

$$\boldsymbol{\Pi} \equiv (-)^{N-1} \mathbf{d}^{N-3}\mathbf{x} \wedge \mathbf{R} + \mathbf{d}^{N-4}\mathbf{x} \wedge \mathbf{S} \wedge \boldsymbol{\Gamma} = d\boldsymbol{\pi} + \frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{\pi}] + (-)^N \frac{1}{4} \mathbf{d}^{N-3}\mathbf{x} \wedge [\boldsymbol{\Gamma}, \boldsymbol{\Gamma}]. \quad (16.344)$$

The variation of the matter action is defined by equations (16.243) in any spacetime dimension. With matter, the equations of motion generalizing equations (16.246) are

$$\frac{1}{2}N^2(N-1) \text{ equations } \mathbf{S} = 2S_N \tilde{\Sigma}, \quad (16.345\text{a})$$

$$N^2 \text{ equations } \boldsymbol{\Pi} = 2S_N \tilde{\mathbf{T}}. \quad (16.345\text{b})$$

5. Split into time and space parts, the alternative spacetime equations of motion (16.345) split into equations of motion that involve time derivatives  $d_{\bar{t}}$  of the  $N(N-1)$  spatial coordinates  $\mathbf{dx}$  and  $N(N-1)$  spatial momenta  $\boldsymbol{\pi}$ , generalizing equations (16.300),

$$N(N-1) \text{ equations } \mathbf{S}_{\bar{t}} = 2S_N \tilde{\Sigma}_{\bar{t}}, \quad (16.346\text{a})$$

$$N(N-1) \text{ equations } \boldsymbol{\Pi}_{\bar{t}} = 2S_N \tilde{\mathbf{T}}_{\bar{t}}, \quad (16.346\text{b})$$

and purely spatial equations involving no time derivatives  $d_{\bar{t}}$ , generalizing equations (16.301),

$$\frac{1}{2}N(N-1) \text{ Gaussian constraints and } \frac{1}{2}N(N-1)(N-3) \text{ identities } \mathbf{S} = 2S_N \tilde{\Sigma}, \quad (16.347\text{a})$$

$$N \text{ Hamiltonian constraints } \boldsymbol{\Pi} = 2S_N \tilde{\mathbf{T}}. \quad (16.347\text{b})$$

6. ADM imposes the  $N-1$  ADM gauge conditions  $\mathbf{dx}_{\bar{0}} = 0$ , equation (16.307), reducing the number of degrees of freedom of the spatial line element  $\mathbf{dx}$  to  $(N-1)^2$ , and likewise the number of degrees of freedom of the spatial area element  $\mathbf{d}^{N-2}\mathbf{x}$  to  $(N-1)^2$ . The momenta conjugate to the spatial area

element are  $\Gamma_{\bar{0}}$ , 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 (16.331),

$$(N - 1)^2 \text{ equations: } (\mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{S})_{\bar{t}} = (-)^{N-1} 2 S_N \overset{**}{\Sigma}_{\bar{t}}, \quad (16.348a)$$

$$(N - 1)^2 \text{ equations: } (\mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{R}_{\bar{0}})_{\bar{t}} = (-)^{N-1} 2 S_N \overset{**}{T}_{\bar{0}\bar{t}}. \quad (16.348b)$$

There are a further  $N - 1$  equations of motion that are discarded in the ADM formalism (incidentally demoting the Gaussian constraints (16.350c) from constraints to identities) but retained in the BSSN formalism, generalizing equation (16.332b),

$$N - 1 \text{ equations: } (\mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{R})_{\bar{t}} = (-)^{N-1} 2 S_N \overset{**}{\mathbf{T}}_{\bar{t}}. \quad (16.349)$$

The remaining equations, containing no time derivatives, comprise  $\frac{1}{2}(N - 1)^2(N - 2)$  identities and  $\frac{1}{2}N(N + 1)$  constraints, generalizing equations (16.332),

$$\frac{1}{2}(N - 1)^2(N - 2) \text{ identities: } (\mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{S}_{\bar{0}})_{\bar{t}} = (-)^{N-1} 2 S_N \overset{**}{\Sigma}_{\bar{0}\bar{t}}, \quad (16.350a)$$

$$\frac{1}{2}(N - 1)(N - 2) \text{ Gaussian constraints: } \mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{S}_{\bar{0}} = (-)^{N-1} 2 S_N \overset{**}{\Sigma}_{\bar{0}}, \quad (16.350b)$$

$$N - 1 \text{ Gaussian constraints: } \mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{S} = (-)^{N-1} 2 S_N \overset{**}{\Sigma}, \quad (16.350c)$$

$$N - 1 \text{ momentum constraints: } \mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{R}_{\bar{0}} = (-)^{N-1} 2 S_N \overset{**}{\mathbf{T}}_{\bar{0}}, \quad (16.350d)$$

$$1 \text{ Hamiltonian constraint: } \mathbf{d}^{N-3} \mathbf{x} \wedge \mathbf{R} = (-)^{N-1} 2 S_N \overset{**}{\mathbf{T}}. \quad (16.350e)$$

**Exercise 16.13 Volume of a ball and area of a sphere.** What is the volume  $V_N$  of a unit  $N$ -ball, and the area  $S_N$  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  $S_{N-1} R^{N-1}$  of an  $(N-1)$ -sphere of radius  $R$  integrated over  $R$  from 0 to 1,

$$V_N = S_{N-1} \int_0^1 R^{N-1} dR, \quad (16.351)$$

yielding

$$V_N = \frac{S_{N-1}}{N}. \quad (16.352)$$

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  $1$ ,

$$V_N = V_{N-1} \int_{-1}^1 r^{N-1} dz = V_{N-1} \int_0^\pi \sin^N \theta d\theta. \quad (16.353)$$

The integral  $\int_0^\pi \sin^N \theta d\theta$  can be expressed in terms of  $\Gamma$  functions. Iterated twice, equation (16.353) gives the recurrence relation

$$V_N = \frac{2\pi V_{N-2}}{N}. \quad (16.354)$$

Equations (16.352) and (16.354) imply

$$S_N = 2\pi V_{N-1} . \quad (16.355)$$

Initial values of the recurrence are  $V_1 = 2$  and  $V_2 = \pi$ . General expressions for the volume and area are

$$V_N = \frac{\pi^{N/2}}{\Gamma(\frac{N}{2} + 1)} , \quad S_N = \frac{2\pi^{(N+1)/2}}{\Gamma(\frac{N+1}{2})} . \quad (16.356)$$


---

## 16.17 Bianchi identities in multivector forms notation

The Bianchi identities, equations (16.376) 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  $dx$  and Lorentz connection  $\Gamma$  in accordance with Cartan's equations (16.207) and (16.203), then the Bianchi identities are satisfied automatically. The contracted Bianchi identities (16.379) enforce conservation laws on the total spin angular-momentum  $\Sigma$  and matter energy-momentum  $T$ .

### 16.17.1 Covariant exterior derivative of a multivector form

The exterior derivatives of the multivector forms  $\Gamma$  and  $dx$  in equations (16.201) and (16.205) 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. Covariance means, by definition, that  $D$  commutes with both coordinate and Lorentz transformations. There is a torsion-free covariant exterior derivative  $\overset{\circ}{D}$ , and a torsion-full covariant exterior derivative  $D$ .

If  $a$  is a multivector  $p$ -form, then its torsion-free covariant exterior derivative  $\overset{\circ}{D}a$  is a sum of the coordinate exterior derivative plus a torsion-free Lorentz connection term, equation (15.4),

$$\overset{\circ}{D}a \equiv da + \overset{\circ}{\Gamma}a , \quad (16.357)$$

where the torsion-free Lorentz connection operator  $\overset{\circ}{\Gamma}$  acting on the multivector form  $a$  is, equation (15.19),

$$\overset{\circ}{\Gamma}a \equiv \frac{1}{2}[\overset{\circ}{\Gamma}, a] , \quad (16.358)$$

with  $\overset{\circ}{\Gamma} \equiv \overset{\circ}{\Gamma}_{kl\kappa} \gamma^k \wedge \gamma^l dx^\kappa$  the torsion-free bivector 1-form, the torsion-free version of equation (16.199a).

The torsion-full covariant exterior derivative  $Da$  is a sum of the coordinate exterior derivative plus a torsion-full Lorentz connection term plus a torsion term,

$Da \equiv da + \overset{\circ}{\Gamma}a + \hat{S}a$

(16.359)

where the Lorentz connection operator  $\hat{\Gamma}$  acting on the multivector form  $\mathbf{a}$  is, equation (15.19),

$$\hat{\Gamma}\mathbf{a} \equiv \frac{1}{2}[\Gamma, \mathbf{a}] . \quad (16.360)$$

The torsion-full Lorentz connection bivector 1-form  $\Gamma$ , equation (16.199a), is as usual the sum of the torsion-free Lorentz connection  $\dot{\Gamma}$  and the contortion  $\mathbf{K}$ , equations (11.55),

$$\Gamma = \dot{\Gamma} + \mathbf{K} . \quad (16.361)$$

The coordinate exterior derivative  $d$  is a torsion-free covariant curl, so the torsion operator  $\hat{S}$  in equation (16.359) must be included when torsion does not vanish, as for example in equation (2.70). 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  $\mathbf{a} = a_{\lambda A} d^p x^{\lambda A}$  is a multivector  $p$ -form, with implicit summation over distinct antisymmetric sequences  $\lambda A$  of  $p$  indices, then the torsion term  $\hat{S}\mathbf{a}$  is the multivector  $(p+1)$ -form defined by

$$\hat{S}\mathbf{a} \equiv \frac{p(p+1)}{2} S_{\kappa\lambda}^{\mu} a_{\mu A} d^{p+1} x^{\kappa\lambda A} , \quad (16.362)$$

implicitly summed over distinct antisymmetric sequences  $\kappa\lambda A$  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 (16.359), of the multivector  $p$ -form  $\mathbf{a}$  is the  $(p+1)$ -form

$$D\mathbf{a} = (p+1) D_\kappa a_{\lambda A} d^{p+1} x^{\kappa\lambda A} = (p+1) (\partial_\kappa a_{\lambda A} + \frac{1}{2} [\Gamma_\kappa, a_{\lambda A}] + \frac{1}{2} p S_{\kappa\lambda}^{\mu} a_{\mu A}) d^{p+1} x^{\kappa\lambda A} , \quad (16.363)$$

with the implicit summation over distinct antisymmetric sequences  $\kappa\lambda A$  of  $p+1$  indices.

Equation (16.359) shows that the operator  $\hat{\Gamma}$  can be interpreted as the connection associated with local Lorentz transformations, while the torsion operator  $\hat{S}$  can be interpreted as the connection associated with displacements.

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 (15.66),

$$D(\mathbf{ab}) = (D\mathbf{a})\mathbf{b} + (-)^p \mathbf{a}(D\mathbf{b}) . \quad (16.364)$$

### 16.17.2 A third covariant exterior derivative

A third covariant exterior derivative for which it can be useful to have a special notation, besides the torsion-free  $\dot{D}$  and torsion-full  $D$  covariant exterior derivatives, is

$$D_\Gamma \equiv d + \hat{\Gamma} , \quad (16.365)$$

which is torsion-free acting on coordinate indices, and torsion-full acting on multivector indices. For example, the Lie derivative of a multivector form can be expressed in terms of  $D_\Gamma$ , equation (16.269).

### 16.17.3 Torsion from the covariant exterior derivative

By construction, the covariant exterior derivative  $D$  commutes with both coordinate and Lorentz transformations. Thus the covariant exterior derivative of the line element  $d\mathbf{x}$  defined by equation (16.199b) vanishes,  $Dd\mathbf{x} = 0$ . Applied to the line interval  $d\mathbf{x}$ , equation (16.359) recovers the definition (16.207) of the torsion vector 2-form  $\mathbf{S}$ ,

$$0 = Dd\mathbf{x} = d\mathbf{dx} + \frac{1}{2}[\mathbf{\Gamma}, d\mathbf{x}] - \mathbf{S} , \quad (16.366)$$

since  $\hat{\mathbf{S}}d\mathbf{x}$  is just (minus) the vector 2-form torsion  $\mathbf{S}$ ,

$$\hat{\mathbf{S}}d\mathbf{x} = S_{\kappa\lambda}^{\mu} e_{m\mu} \boldsymbol{\gamma}^m d^2x^{\kappa\lambda} = S_{m\kappa\lambda} \boldsymbol{\gamma}^m d^2x^{\kappa\lambda} = -\mathbf{S} . \quad (16.367)$$

Equation (16.366) is Cartan's first equation of structure. It defines the torsion  $\mathbf{S}$ .

The torsion-free version of Cartan's equation (16.366) is  $dd\mathbf{x} + \frac{1}{2}[\mathring{\mathbf{\Gamma}}, d\mathbf{x}] = 0$ . Subtracting this from the torsion-full Cartan's equation (16.366) yields the relation between the torsion  $\mathbf{S}$  and the contortion  $\mathbf{K}$ ,

$$\mathbf{S} = \frac{1}{2}[\mathbf{K}, d\mathbf{x}] . \quad (16.368)$$

### 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

$$DD \equiv [D_\kappa, D_\lambda] d^2x^{\kappa\lambda} , \quad (16.369)$$

implicitly summed over distinct antisymmetric pairs of indices  $\kappa\lambda$ . Acting on any multivector form  $\mathbf{a}$ , the square of the covariant exterior derivative gives (compare equation (15.21))

$$DD\mathbf{a} = \hat{\mathbf{R}}\mathbf{a} + \hat{\mathbf{S}}D\mathbf{a} . \quad (16.370)$$

If  $\mathbf{a} = \mathbf{a}_{\mu A} d^p x^{\mu A}$  is a multivector  $p$ -form, then the Riemann operator  $\hat{\mathbf{R}}$  acting on  $\mathbf{a}$  is the  $(p+2)$ -form

$$\hat{\mathbf{R}}\mathbf{a} = \frac{(p+1)(p+2)}{2} (\frac{1}{2}[\mathbf{R}_{\kappa\lambda}, \mathbf{a}_{\mu A}] + \frac{1}{3}p R_{\kappa\lambda\mu}{}^\nu \mathbf{a}_{\nu A}) d^{p+2}x^{\kappa\lambda\mu A} , \quad (16.371)$$

implicitly summed over distinct antisymmetric sequences  $\kappa\lambda\mu A$  of  $p+2$  indices. The components of the Riemann tensor are those of the Riemann bivector 2-form  $\mathbf{R}$ , equation (16.200). Equation (16.370) along with the definition (16.359) of the exterior covariant recovers the definition (16.203) of the Riemann curvature  $\mathbf{R}$  in terms of the Lorentz connection  $\mathbf{\Gamma}$ . This is Cartan's second equation of structure, equation (16.203). In equation (16.370), the scalar 2-form covariant derivative operator  $\hat{\mathbf{S}}D \equiv S_{\kappa\lambda}^\nu D_\nu$  acting on the multivector  $p$ -form  $\mathbf{a} = \mathbf{a}_A d^p x^A$  is the  $(p+2)$ -form

$$\hat{\mathbf{S}}D\mathbf{a} = \frac{(p+1)^2(p+2)}{2} S_{\kappa\lambda}^\nu D_\nu [\mathbf{a}_A] d^{p+2}x^{\kappa\lambda A} , \quad (16.372)$$

implicitly summed over distinct antisymmetric sequences  $\kappa\lambda A$  of  $p+2$  indices.

### 16.17.5 Bianchi identities

The Jacobi identity for the covariant exterior derivative is

$$D(DD) - (DD)D = 0 . \quad (16.373)$$

Applied to an arbitrary multivector form  $\mathbf{a}$ , the Jacobi identity (16.373) implies

$$0 = D(DD)\mathbf{a} - (DD)D\mathbf{a} = (D\hat{\mathbf{R}} - \hat{\mathbf{S}}\hat{\mathbf{R}})\mathbf{a} + (D\hat{\mathbf{S}} - \hat{\mathbf{S}}\hat{\mathbf{S}} - \hat{\mathbf{R}})\mathbf{D}\mathbf{a} . \quad (16.374)$$

Equation (16.374) holds for arbitrary  $\mathbf{a}$ , so the coefficients of  $\mathbf{a}$  and  $D\mathbf{a}$  must vanish, implying the Bianchi identities

$$D\mathbf{S} - \hat{\mathbf{S}}\mathbf{S} + \frac{1}{2}[\mathbf{dx}, \mathbf{R}] = 0 , \quad (16.375a)$$

$$D\mathbf{R} - \hat{\mathbf{S}}\mathbf{R} = 0 , \quad (16.375b)$$

where  $\mathbf{S}$  is the vector 2-form torsion (16.204) with  $4 \times 6 = 24$  components, and  $\mathbf{R}$  is the bivector 2-form Riemann curvature (16.200) with  $6 \times 6 = 36$  components. These equations (16.375) were given previously in component form by equations (11.69) and (11.91). Equation (16.375a) is a vector 3-form, with  $4 \times 4 = 16$  components, while equation (16.375b) is a bivector 3-form, with  $6 \times 4 = 24$  components. Equivalently, in terms of the exterior derivative  $d$  instead of the covariant exterior derivative  $D$ , the Bianchi identities (16.375) are

$$d\mathbf{S} + \frac{1}{2}[\mathbf{T}, \mathbf{S}] + \frac{1}{2}[\mathbf{dx}, \mathbf{R}] = 0 , \quad (16.376a)$$

$$d\mathbf{R} + \frac{1}{2}[\mathbf{T}, \mathbf{R}] = 0 . \quad (16.376b)$$

### 16.17.6 Interpretation of the Bianchi identities

The torsion Bianchi identity (16.376a) looks like a covariant conservation equation for torsion  $\mathbf{S}$ , except that there is a source term  $\frac{1}{2}[\mathbf{dx}, \mathbf{R}]$ . This term is a vector 3-form whose 16 components comprise the antisymmetric part (in components,  $R_{[[kl][mn]]}$ ) of the 36-component Riemann tensor (see Exercise 11.6 for the relation between  $R_{[klm]n}$  and  $R_{[[kl][mn]]}$ ),

$$\frac{1}{2}[\mathbf{dx}, \mathbf{R}] = \mathbf{dx} \cdot \mathbf{R} = R_{[\kappa\lambda\mu]n} \boldsymbol{\gamma}^n d^3x^{\kappa\lambda\mu} . \quad (16.377)$$

Since torsion  $\mathbf{S}$  is determined completely by its equation of motion (16.234a) or (16.246a) in terms of the spin angular-momentum  $\mathbf{\Sigma}$ , the torsion Bianchi identity (16.376a) can be interpreted as determining the 16-component antisymmetric part  $\mathbf{dx} \cdot \mathbf{R}$  of the Riemann tensor in terms of the torsion. If torsion  $\mathbf{S}$  vanishes, or more generally if it satisfies the covariant conservation equation  $d\mathbf{S} + \frac{1}{2}[\mathbf{T}, \mathbf{S}] = 0$ , then the Riemann tensor is symmetric,  $\mathbf{dx} \cdot \mathbf{R} = 0$ , but if torsion does not vanish, then generically the Riemann tensor is not symmetric.

The Riemann Bianchi identity (16.376b) looks like a covariant conservation equation for the Riemann tensor  $\mathbf{R}$ . In contrast to the torsion  $\mathbf{S}$ , the Riemann tensor  $\mathbf{R}$  is not determined completely by its equation of motion (16.234b) or (16.246b) in terms of the matter energy-momentum  $\mathbf{T}$ . Rather, the equation of motion

determines only the contracted part  $\mathbf{dx} \wedge \mathbf{R}$  of the Riemann tensor, which is the Einstein tensor (strictly, minus its double dual, equation (16.230)). The Riemann Bianchi identity (16.376b) has 24 components. Of these 24 components, 4 provide an equation for  $d(\mathbf{dx} \cdot \mathbf{R})$ , which is a differential constraint on the antisymmetric part of the Riemann tensor, a further 4 components provide an equation for  $d(\mathbf{dx} \wedge \mathbf{R})$ , which is a differential constraint on the Einstein tensor, and the remaining 16 components provide Maxwell-like differential equations on the symmetric part of the Riemann tensor, as discussed previously in §3.2. These Maxwell-like equations govern the behaviour of gravitational waves, which are encoded in the part of the Riemann tensor that is not determined by the equations of motion, namely the 10-component symmetric Weyl tensor. The 16 components are subject to a 6-component conservation law,

$$D_\Gamma(D_\Gamma \mathbf{R}) = (D_\Gamma D_\Gamma) \mathbf{R} = \frac{1}{2} [\mathbf{R}, \mathbf{R}] = 0 , \quad (16.378)$$

the last step of which follows from equation (16.198b). Equation (16.378) represents conservation of the Weyl current, equation (3.9).

### 16.17.7 Contracted Bianchi identities

The equations of motion (16.234) for torsion and curvature are sourced by the spin angular-momentum and matter energy-momentum. The Bianchi identities (16.376) on the other hand are independent of matter sources. The Bianchi identities impose differential constraints on the equations of motion that must be satisfied regardless of the form of the spin angular-momentum and matter energy-momentum.

The equations of motion (16.234) are equations for  $\mathbf{dx} \wedge \mathbf{S}$  and  $\mathbf{dx} \wedge \mathbf{R}$ . Differential constraints on these combinations are obtained by contracting the Bianchi identities (16.376) by pre-multiplying by  $\mathbf{dx} \wedge$ . The contracted Bianchi identities for torsion and Riemann curvature constitute respectively a bivector 4-form, with 6 components, and a pseudovector 4-form, with 4 components,

$$d(\mathbf{dx} \wedge \mathbf{S}) + \frac{1}{2} [\boldsymbol{\Gamma}, \mathbf{dx} \wedge \mathbf{S}] + \frac{1}{2} [d^2 \mathbf{x}, \mathbf{R}] = 0 , \quad (16.379a)$$

$$d(\mathbf{dx} \wedge \mathbf{R}) + \frac{1}{2} [\boldsymbol{\Gamma}, \mathbf{dx} \wedge \mathbf{R}] - \mathbf{S} \wedge \mathbf{R} = 0 . \quad (16.379b)$$

The final term in the contracted torsion Bianchi identity (16.379a) is a bivector 4-form whose 6 components constitute the antisymmetric part of the Ricci tensor (see eq. (16.214)),

$$\frac{1}{2} [d^2 \mathbf{x}, \mathbf{R}] = \mathbf{dx} \wedge (\mathbf{dx} \cdot \mathbf{R}) = \mathbf{dx} \cdot (\mathbf{dx} \wedge \mathbf{R}) = -e_{k\textcolor{brown}{\mu}} e_{l\textcolor{brown}{\lambda}} R_{[\mu\nu]} \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l d^4 x^{\textcolor{brown}{\lambda}\mu\nu} . \quad (16.380)$$

Combining the contracted Bianchi identity (16.379b) with the torsion Bianchi identity (16.376a) yields the pseudovector 4-form identity for the curvature  $\boldsymbol{\Pi}$  defined by equation (16.242),

$$d\boldsymbol{\Pi} + \frac{1}{2} [\boldsymbol{\Gamma}, \boldsymbol{\Pi}] + \frac{1}{2} [\mathbf{dx}, \mathbf{R}] \wedge \boldsymbol{\Gamma} = 0 . \quad (16.381)$$

### 16.17.8 Interpretation of the contracted Bianchi identities

The contracted torsion Bianchi identity (16.379a) is the 6-component conservation law associated with invariance of the gravitational Lagrangian under Lorentz transformations. The contracted Riemann iden-

tity (16.379b), or equivalently (16.381), is the 4-component conservation law associated with invariance of the gravitational Lagrangian under coordinate transformations.

The contracted torsion Bianchi identity (16.379a) enforces continued satisfaction of the Gaussian constraint (16.290a). The contracted Riemann Bianchi identity (16.379b) enforces continued satisfaction of the Hamiltonian constraint (16.290b).

# 17

---

## 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. In practical calculations however, it is usually more convenient to follow a 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, and Misner (1959) and Arnowitt, Deser, and Misner (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. The ADM formalism provides the backbone for modern codes that implement numerical general relativity.

The ADM formalism shows that 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* 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. Gauge choices other than ADM can be made, and are often preferred. In cosmology for example, the preferred gauge choice is conformal Newtonian (Copernican) gauge, §28.8, in which only 3 of the 6 physical perturbations are part of the spatial coordinate metric  $g_{\alpha\beta}$  (the scalar  $\Phi$  and the 2 components of the tensor  $h_{ab}$ ), while the remaining 3 physical perturbations are part of the time components  $g_{t\beta}$  of the metric (the scalar  $\Psi$  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, 1999).

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, §17.7, and the elegant 4-dimensional version of it proposed by Pretorius (2005b), §17.8.

In this chapter, torsion is assumed to vanish.

## 17.1 ADM formalism

The ADM formalism splits the spacetime coordinates  $x^\mu$  into a time coordinate  $t$  and spatial coordinates  $x^\alpha$ ,  $\alpha = 1, 2, 3$ ,

$$x^\mu \equiv \{t, x^\alpha\} . \quad (17.1)$$

At each point of spacetime, the spacelike hypersurface of constant time  $t$  has a unique future-pointing unit normal  $\gamma_0$ , defined to have unit length and to be orthogonal to the spatial tangent axes  $e_\alpha$ ,

$$\gamma_0 \cdot \gamma_0 = -1 , \quad \gamma_0 \cdot e_\alpha = 0 \quad \alpha = 1, 2, 3 . \quad (17.2)$$

The central idea of the ADM approach is to work in a tetrad frame  $\gamma_m$  consisting of this time axis  $\gamma_0$ , together with three spatial tetrad axes  $\gamma_a$ , also called the **triad**, that are orthogonal to the tetrad time axis  $\gamma_0$ , 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  $\gamma_{mn}$  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  $\gamma^{mn}$  is correspondingly

$$\gamma^{mn} = \begin{pmatrix} -1 & 0 \\ 0 & \gamma^{ab} \end{pmatrix}, \quad (17.5)$$

whose spatial part  $\gamma^{ab}$  is the inverse of  $\gamma_{ab}$ . Given the conditions (17.2) and (17.3), the vierbein  $e_m^{\mu}$  and inverse vierbein  $e^m_{\mu}$  take the form

$$e_m^{\mu} = \begin{pmatrix} 1/\alpha & \beta^{\alpha}/\alpha \\ 0 & e_a^{\alpha} \end{pmatrix}, \quad e^m_{\mu} = \begin{pmatrix} \alpha & 0 \\ -e^a_{\alpha} \beta^{\alpha} & e^a_{\alpha} \end{pmatrix}, \quad (17.6)$$

where  $\alpha$  and  $\beta^{\alpha}$  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,  $e_a^{\alpha} e^b_{\alpha} = \delta_a^b$ . 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),

$$e_a^t = e^0_{\alpha} = 0. \quad (17.7)$$

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  $\alpha$  is called the **lapse**, while  $\beta^{\alpha}$  is called the **shift**. Physically, the lapse  $\alpha$  is the rate at which the proper time  $\tau$  of the tetrad rest frame elapses per unit coordinate time  $t$ , while the shift  $\beta^{\alpha}$  is the velocity at which the tetrad rest frame moves through the spatial coordinates  $x^{\alpha}$  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  $\partial_0$  in the time direction  $\gamma_0$ ,

$$\frac{d}{d\tau} = u^{\mu} \frac{\partial}{\partial x^{\mu}} = \partial_0. \quad (17.12)$$

Tetrad and coordinate derivatives  $\partial_m$  and  $\partial/\partial x^{\mu}$  are related to each other as usual by the vierbein and its inverse,

$$\partial_0 = \frac{1}{\alpha} \left( \frac{\partial}{\partial t} + \beta^{\alpha} \frac{\partial}{\partial x^{\alpha}} \right), \quad \frac{\partial}{\partial t} = \alpha \partial_0 - \beta^a \partial_a, \quad (17.13a)$$

$$\partial_a = e_a{}^{\alpha} \frac{\partial}{\partial x^{\alpha}}, \quad \frac{\partial}{\partial x^{\alpha}} = e^a{}_{\alpha} \partial_a, \quad (17.13b)$$

where  $\beta^a \equiv e_a{}^{\alpha} \beta^{\alpha}$ . By construction, the only directed derivative involving a coordinate time derivative  $\partial/\partial t$  is the directed time derivative  $\partial_0$ , and conversely the only coordinate derivative involving the directed time derivative  $\partial_0$  is the coordinate time derivative  $\partial/\partial t$ .

**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. More precisely, the preferred cosmological frame is defined by conformal Newtonian (Copernican) gauge, §28.8, which is that gauge for which the retained gravitational perturbations are precisely the physical perturbations. 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. Interestingly, conformal Newtonian gauge does not conform to ADM gauge choices: in conformal Newtonian gauge, only 3 of the 6 physical perturbations ( $\Phi$  and  $h_{ab}$ ) are part of the spatial metric, while the remaining 3 physical perturbations ( $\Psi$  and  $W_a$ ) are part of the lapse and shift. Conformal Newtonian gauge holds as long as gravitational perturbations are weak, which is true even in highly non-linear collapsed systems such as galaxies and solar systems. Conformal Newtonian gauge breaks down in strongly gravitating systems such as black holes.

### 17.1.1 Traditional ADM approach

The traditional ADM approach sets the spatial tetrad axes  $\gamma_a$  equal to the spatial coordinate tangent axes  $e_{\alpha}$ ,

$$\gamma_a = \delta_a^{\alpha} e_{\alpha} \quad (\text{traditional ADM}), \quad (17.14)$$

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  $\gamma_a$  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  $\gamma_a$  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  $\gamma_{mn}$  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  $\gamma_0$  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  $e_\alpha$ , and tetrad transformations of the spatial tetrad axes  $\gamma_a$  (spatial rotations).

A covariant **spatial coordinate vector**  $A_\alpha$  is defined to be a vector that transforms like the spatial coordinate axes  $e_\alpha$ . Likewise a covariant **spatial tetrad vector**  $A_a$  is defined to be a vector that transforms like the spatial tetrad axes  $\gamma_a$ . 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  $\gamma_{ab}$  and their inverses, and spatial indices are transformed between coordinate and tetrad frames with the spatial vierbein  $e_a^\alpha$  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  $8\pi$  replaced by  $16\pi\alpha$  for convenience,

$$16\pi\alpha \frac{\delta L_g}{\delta(\partial\Gamma_{mn\lambda}/\partial t)} = \alpha(e^{m\textcolor{brown}{t}} e^{n\lambda} - e^{m\lambda} e^{n\textcolor{brown}{t}}) . \quad (17.15)$$

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  $\lambda$  must be a spatial coordinate index  $\alpha$ . 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,

$$8\pi\alpha \frac{\delta L'_g}{\partial(\partial e^{n\lambda}/\partial t)} = \alpha e^{mt} \pi_{nm\lambda} , \quad (17.17)$$

where  $\pi_{nm\lambda}$  are related to the Lorentz connections  $\Gamma_{nm\lambda}$  by equation (16.156). 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  $\lambda$  also be a spatial coordinate index  $\beta$ . 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 (16.156) 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 (16.158), 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  $\lambda$  are related to the Lorentz connections  $\Gamma_{mnl}$  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  $\Gamma_{a0b}$  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  $\Gamma_{a00}$ , which are called the **acceleration**  $K_a$ . The combined set of connections with middle index 0 is called the **generalized extrinsic curvature**  $K_{mol} \equiv \Gamma_{mol}$ . 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  $\gamma_{0a}$  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  $\gamma_{ab}$  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  $\boldsymbol{\gamma}_0$  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  $\pi_{ab}$  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  $\pi_{ab}$ .

### 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  $K_{mol} \equiv \Gamma_{mol}$  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  $\gamma_{mn}$  is constant, it is useful to allow a more general situation in which the spatial

tetrad metric  $\gamma_{ab}$  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  $\gamma_{ab}$  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  $\Gamma_{mnl}$  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  $\Gamma_{mnl}$  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  $\gamma_{lm}$  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  $e_m{}^\mu$ ,

$$0 = D_k e_m{}^\mu = \hat{D}_k e_m{}^\mu - K_{mk}^n e_n{}^\mu + K_{\nu k}^\mu e_m{}^\nu = \hat{D}_k e_m{}^\mu - K_{mk}^\mu + K_{mk}^\mu = \hat{D}_k e_m{}^\mu. \quad (17.31)$$

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 (15.47) 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  $\mathbf{K}_\kappa \equiv \frac{1}{2}K_{mnl\kappa} \gamma^m \wedge \gamma^n$  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 (11.61), 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), and (17.40d) are called respectively the **Ricci**, **Codazzi-Mainardi**, 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 11.6. 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, §17.7, 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 prove 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  $R_{km}$ ,

$$R_{00} = \hat{D}^m K_m - K^{ba} K_{ab} + K^a K_a , \quad (17.44a)$$

$$R_{a0} = -\hat{D}_a K + \hat{D}^b K_{ba} - K^b (K_{ab} - K_{ba}) \quad (17.44b)$$

$$= R_{0a} = \hat{R}_{0a} - K^b K_{ab} + K_a K_a , \quad (17.44c)$$

$$R_{ab} = \hat{R}_{ab} - \hat{D}_a K_b + \hat{D}_b K_{ba} + K_{ba} K - K_a K_b . \quad (17.44d)$$

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, §17.7, 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

$$L_{\text{ADM}} = \frac{1}{16\pi} (\hat{R} + K^{ba} K_{ab} - K^2) . \quad (17.47)$$

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 (16.159). 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 (11.32) (equations (11.53) allow for a non-constant spatial tetrad metric  $\gamma_{ab}$ , thus admitting the traditional ADM approach in which the spatial tetrad  $\gamma_a$  are set equal to the spatial coordinate tangent axes  $e_\alpha$ , 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  $\pi_{ab}$ , 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  $\Gamma_{abc}$ , like the spatial tetrad axes  $\gamma_a$ , transform under temporal coordinate transformations despite the absence of temporal indices. If the spatial tetrad metric  $\gamma_{ab}$  is taken to be constant, which is true if for example the spatial tetrad axes  $\gamma_a$  are taken to be orthonormal, then the tetrad connections  $\Gamma_{ab0}$  and  $\Gamma_{abc}$ , 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  $\Gamma_{ab0}$  and  $\Gamma_{abc}$  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  $\gamma_{ab}$ , 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  $\gamma_{ab}$  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 inverse 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

$$\boxed{\frac{\partial g_{\alpha\beta}}{\partial t} = -\hat{\mathcal{L}}_\beta g_{\alpha\beta} + 2\alpha K_{\alpha\beta}} , \quad (17.55)$$

where  $\hat{\mathcal{L}}_\beta g_{\alpha\beta}$  is the Lie derivative (7.132) of the spatial coordinate metric with respect to the shift  $\beta^\gamma$ , restricted to the hypersurface of constant time (hence the restricted  $\hat{\cdot}$  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 (7.130), 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 (7.132)),

$$\hat{\mathcal{L}}_\beta g_{\alpha\beta} = \hat{D}_\beta \beta_\alpha + \hat{D}_\alpha \beta_\beta . \quad (17.57)$$

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),

$$d_{0a}^a = -\frac{1}{\alpha} e^a_\alpha \partial_a \beta^\alpha = -\frac{1}{\alpha} \frac{\partial \beta^\alpha}{\partial x^\alpha} , \quad d_{a0}^a = e_a^\beta \partial_0 e^a_\beta = -\partial_0 \ln e , \quad (17.58)$$

where  $e \equiv |e_\alpha|$  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

$$\frac{\partial \ln g}{\partial t} = -\beta^\alpha \frac{\partial \ln g}{\partial x^\alpha} - 2 \frac{\partial \beta^\alpha}{\partial x^\alpha} + 2\alpha K , \quad (17.60)$$

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  $\Gamma_{abc}$ , 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  $\Gamma_{abc}$  through equations (17.48a) and (17.48d).

The final set of connections is  $\Gamma_{ab0}$ , 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,

$$R_{ab} = 8\pi (T_{ab} - \frac{1}{2}\gamma_{ab}T) . \quad (17.62)$$

Substituting the spatial Ricci tensor from equation (17.44d) transforms the spatial Einstein equations (17.62) into equations of motion for the extrinsic curvature  $K_{ab}$ ,

$$\boxed{\hat{D}_0 K_{ab} = \hat{D}_a K_b - K_{ab}K + K_a K_b - \hat{R}_{ab} + 8\pi (T_{ab} - \frac{1}{2}\gamma_{ab}T)} . \quad (17.63)$$

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  $d_{\beta 0}^c \equiv \partial_0 e_b^\alpha \partial_0 K_{\alpha\beta}$  of the spatial inverse 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

$$\hat{D}_0 K_{\alpha\beta} = \frac{1}{\alpha} \left( \frac{\partial K_{\alpha\beta}}{\partial t} + \hat{\mathcal{L}}_\beta K_{\alpha\beta} \right) - 2K_\alpha^\gamma K_{\gamma\beta} . \quad (17.66)$$

where  $\hat{\mathcal{L}}_\beta K_{\alpha\beta}$  is the Lie derivative of the extrinsic curvature in the direction  $\beta^\gamma$ , equation (7.129),

$$\hat{\mathcal{L}}_\beta K_{\alpha\beta} = \beta^\gamma \frac{\partial K_{\alpha\beta}}{\partial x^\gamma} + K_{\gamma\alpha} \frac{\partial \beta^\gamma}{\partial x^\beta} + K_{\gamma\beta} \frac{\partial \beta^\gamma}{\partial x^\alpha} . \quad (17.67)$$

As usual with a Lie derivative, equation (7.129), 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

$$\boxed{\frac{\partial K_{\alpha\beta}}{\partial t} = -\hat{\mathcal{L}}_\beta K_{\alpha\beta} + \alpha \left[ \hat{D}_\alpha K_\beta + 2K_\alpha^\gamma K_{\gamma\beta} - K_{\alpha\beta} K + K_\alpha K_\beta - \hat{R}_{\alpha\beta} + 8\pi (T_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}T) \right]} . \quad (17.68)$$

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  $\gamma_{ab}$  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  $\rho$  the proper energy density. In the coordinate frame, equation (17.70) becomes

$$\frac{\partial K}{\partial t} = -\beta^\alpha \frac{\partial K}{\partial x^\alpha} + \alpha \left[ \hat{D}_\alpha K^\alpha - K^2 + K^\alpha K_\alpha - \hat{R} + 12\pi(\rho - p) \right] . \quad (17.71)$$

### 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  $\alpha$  and shift  $\beta^\alpha$  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  $G_{a0} \equiv R_{a0}$ ,

$$G_{00} = \frac{1}{2}(\hat{R} - K^{ab}K_{ba} + K^2) = \frac{1}{2}(\hat{R} - \pi^{ab}K_{ba}) , \quad (17.73a)$$

$$G_{a0} = \hat{D}^bK_{ba} - \hat{D}_aK = \hat{D}^b\pi_{ba} . \quad (17.73b)$$

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**:

$$\frac{1}{2}(\hat{R} - K^{ab}K_{ba} + K^2) = 8\pi T_{00} \quad \text{Hamiltonian constraint ,} \quad (17.74a)$$

$$\hat{D}^bK_{ba} - \hat{D}_aK = 8\pi T_{a0} \quad \text{momentum constraints .} \quad (17.74b)$$

**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

$$\frac{\partial K}{\partial t} = -\beta^\alpha \frac{\partial K}{\partial x^\alpha} + \alpha \left[ \hat{D}_\alpha K^\alpha - K^{\alpha\beta} K_{\beta\alpha} + K^\alpha K_\alpha - 4\pi(\rho + 3p) \right] . \quad (17.76)$$

Equation (17.76) is the Raychaudhuri equation (18.22a) with vanishing vorticity and non-vanishing acceleration  $K_{\alpha}$ .

### 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  $\gamma_{ab}$  can still be allowed to be constant, as in the case of an orthonormal tetrad. If the inverse spatial vierbein  $e^a_{\alpha}$  is factored as a product of the conformal factor  $a$  and an inverse conformal vierbein  $\tilde{e}^a_{\alpha}$ , then the vierbein and inverse vierbein become

$$e_a^{\alpha} = \tilde{e}_a^{\alpha}/a, \quad e^a_{\alpha} = a \tilde{e}^a_{\alpha}. \quad (17.77)$$

The conformal vierbein and inverse conformal vierbein are inverse to each other,  $\tilde{e}_a^{\alpha} \tilde{e}^b_{\alpha} = \delta_a^b$ . The lapse  $\alpha$  and shift  $\beta^{\alpha}$  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  $\gamma_m$  (that is, transformations that leave the time axis  $\gamma_0$  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 (11.32). 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

$$d_{abn} \equiv -\gamma_{ac} e^c_{\alpha} \partial_n e_b^{\alpha} = \gamma_{ac} e^c_{\alpha} e_b^{\alpha} \partial_n \ln a - \gamma_{ac} \tilde{e}^c_{\alpha} \partial_n \tilde{e}_b^{\alpha} = \gamma_{ab} \partial_n \ln a + \tilde{d}_{abn}, \quad (17.79)$$

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  $\tilde{e}_b^{\alpha}$ . 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  $\Gamma_{abc}$  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  $e_m^{\mu}$  and the tetrad metric  $\gamma_{lm}$  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),

$$R_{00} = -\hat{D}_0 K + \hat{D}_a K_{00}^a - K^{ba} K_{ab} + K_{00}^a K_a , \quad (17.85a)$$

$$R_{a0} = -\hat{D}_a K + \hat{D}^b K_{ba} - K_{ab} K_{00}^b + K_{ba} K^b - K_{cab} K^{bc} , \quad (17.85b)$$

$$= R_{0a} = \hat{R}_{0a} - \hat{D}_0 K_{ab}^b - K_{00}^b K_{ab} + K_{a00} K - K^{bc} K_{cab} , \quad (17.85c)$$

$$R_{ab} = \hat{R}_{ab} - \hat{D}_a K_b + \hat{D}_b K_{ab} + \hat{D}^c K_{cba} + K_{ba} K - K_{a00} K_{b00} + K_{cba} K^c - K_{ad}^c K_{bc}^d . \quad (17.85d)$$

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

$$\hat{D}_0 K_{ab} = \hat{D}_a K_b - \hat{D}^c K_{cba} - K_{ab} K + K_{a00} K_{b00} - K_{cba} K^c + K_{ad}^c K_{bc}^d - \hat{R}_{ab} + 8\pi (T_{ab} - \frac{1}{2} \gamma_{ab} T) . \quad (17.86)$$

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

$$\begin{aligned} \frac{\partial K_{\alpha\beta}}{\partial t} &= \hat{\mathcal{L}}_\beta K_{\alpha\beta} + \alpha \left[ \hat{D}_\alpha K_\beta - \hat{D}^\gamma K_{\gamma\beta\alpha} + 2K_\alpha^\gamma K_{\gamma\beta} - K_{\alpha\beta} K + K_{\alpha 00} K_{\beta 00} - K_{\gamma\beta\alpha} K^\gamma + K_{\alpha\delta}^\gamma K_{\beta\gamma}^\delta - \hat{R}_{\alpha\beta} \right. \\ &\quad \left. + 8\pi (T_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} T) \right] . \end{aligned} \quad (17.87)$$

Again, this equation 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 3-dimensional Lie group is called a Bianchi space (Bianchi, 1898). A Lie group is a group of symmetry transformations that is also a differentiable manifold. Lie groups are generated by infinitesimal transformations called the generators of the group. A 3-dimensional Lie group has 3 linearly independent generators. The properties of a Lie group are determined by the commutators of its generators, or equivalently by its structure coefficients  $c_{ab}^c$ , equation (17.88), which for a Lie group are taken to be constant. A Bianchi space is consequently homogeneous. The assumption that a space is a Lie group is stronger than the assumption that the space is homogeneous, which requires merely that the tetrad-frame Riemann tensor be spatially constant. However, most homogeneous 3-dimensional spaces are Lie groups, hence Bianchi spaces, the notable exception being the closed cylindrical geometry, equation (17.127). Bianchi spaces are homogeneous but not necessarily isotropic.

Bianchi spacetimes, also known as Bianchi universes, are Bianchi spaces that evolve in time while preserving the posited Lie group 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

The assumption that a space is homogeneous requires that the space has a complete set of spacelike Killing vectors, thus 3 linearly independent spacelike Killing vectors in 3-dimensional space. The spatial components  $\gamma_a$  of the tetrad can be chosen to coincide with the 3 Killing vectors at each point. Equivalently, the 3 Killing

vectors can be identified with the directed derivatives  $\partial_a$  along the 3 spatial tetrad axes (see §7.32). The commutators of the directed derivatives define the structure coefficients  $c_{ab}^c$ ,

$$[\partial_a, \partial_b] = c_{ab}^c \partial_c , \quad (17.88)$$

which are necessarily antisymmetric in their last two indices  $ab$ . Homogeneity does not require that the structure coefficients be spatially constant; rather, homogeneity requires that the tetrad-frame Riemann tensor be spatially constant. However, Bianchi spaces are by assumption Lie groups, for which the structure coefficients are spatially constant. For Bianchi spaces, the Killing vectors  $\partial_a$  are the generators of the Lie group, whose properties are determined by the structure coefficients  $c_{ab}^c$ . The vector space of real linear combinations of the generators  $\partial_a$  defines a Lie algebra, with multiplication defined by equation (17.88).

The structure coefficients must be such that Jacobi identity  $[\partial_{[a}, [\partial_b, \partial_c]] = 0$  is satisfied. If the structure coefficients are spatially constant, then the Jacobi identity requires

$$c_{d[a}^e c_{bc]}^d = 0 . \quad (17.89)$$

#### 17.4.2 Bianchi line element

A Bianchi spacetime is a Bianchi space that evolves in time while preserving the Lie group spatial structure. The spatial Bianchi line element can be constructed out of 1-forms  $e^a_\alpha dx^\alpha$  which, being aligned with the Killing vectors  $\gamma_a$ , are by construction independent of the choice of spatial coordinates  $x^\alpha$ . The time coordinate  $t$  is chosen so that spatial surfaces of constant time are homogeneous. To preserve spatial homogeneity, the tetrad metric  $\gamma_{ab}$  must be independent of the spatial coordinates, but it may depend on time  $t$ . As usual in the ADM formalism, the tetrad time axis  $\gamma_0$  is chosen to be orthogonal to the spatial tetrad axes  $\gamma_a$ , which lie in the surfaces of constant time. The line element can thus be taken to be

$$ds^2 = -dt^2 + \gamma_{ab}(t) e^a_\alpha e^b_\beta dx^\alpha dx^\beta , \quad (17.90)$$

which is in ADM form with unit lapse and zero shift. The vierbein and its inverse are

$$e_m^\mu = \begin{pmatrix} 1 & 0 \\ 0 & e_a^\alpha \end{pmatrix} , \quad e^m_\mu = \begin{pmatrix} 1 & 0 \\ 0 & e^a_\alpha \end{pmatrix} . \quad (17.91)$$

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 Killing vectors do not depend on time,  $[\partial_0, \partial_a] = 0$ , so the vierbein, and the inverse vierbein, are independent of time. However, despite spatial homogeneity, the spatial vierbein coefficients  $e_a^\alpha$  may (and generically do) depend on the spatial coordinates, as they do for example in FLRW spacetimes. Likewise homogeneity allows that the structure coefficients  $c_{ab}^c$  defined by the commutators of the directed derivatives, equation (17.88), may be functions of the spatial coordinates. As emphasized above, Bianchi spaces are by assumption those for which the structure coefficients are spatially constant, but this is not required by homogeneity. Whether or not the structure coefficients are spatially constant, they satisfy  $c_{ab}^c \equiv 2d_{[ab]}^c$ , where  $d_{ab}^c$  are the spatial components of the vierbein derivatives, equation (11.32).

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 <sub>0</sub>	VI
0	+	+	VII <sub>0</sub>	VII
-	+	+	VIII	
+	+	+	IX	

### 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  $3 \times 3$  matrix  $n^{dc}$ , which can be decomposed into symmetric  $n^{(dc)}$  and antisymmetric  $n^{[dc]}$  parts,

$$c_{ab}^c = \varepsilon_{abd} n^{dc} = \varepsilon_{abd} (n^{(dc)} + n^{[dc]}) . \quad (17.92)$$

By an 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  $k_e$ ,

$$c_{ab}^c = \varepsilon_{abd} (\delta^{dc} n_c - \varepsilon^{dce} k_e) \quad (\text{no sum over } c) . \quad (17.93)$$

The Jacobi identity (17.89) 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

$$n^{fe} k_f = n_e k_e = 0 \quad (\text{each direction } e, \text{ no sum over } e) . \quad (17.94)$$

Thus in each direction, either  $n_e$  or  $k_e$  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 (17.88) of the directed derivatives  $\partial_a$  then reduce to (with at least one of  $n_1$  and  $k$  zero)

$$[\partial_3, \partial_2] = n_1 \partial_1 , \quad [\partial_1, \partial_3] = n_2 \partial_2 - k \partial_3 , \quad [\partial_2, \partial_1] = n_3 \partial_3 + k \partial_2 . \quad (17.95)$$

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 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  $\pm 1$ . 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.

#### 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),

$$\Gamma_{ab0} = \Gamma_{a0b} = -\Gamma_{0ab} = \frac{1}{2}\dot{\gamma}_{ab}, \quad \Gamma_{abc} = \frac{1}{2}(c_{cab} + c_{bac} - c_{abc}), \quad (17.96)$$

where the overdot represents the time derivative,  $\dot{\gamma}_{ab} \equiv d\gamma_{ab}/dt$  (an ordinary derivative because  $\gamma_{ab}$  varies only in time, not space), and  $c_{cab} \equiv \gamma_{cd}\gamma_{ab}^d$ . The connections with one time 0 index are symmetric in their spatial indices  $ab$ , while the purely spatial connections  $\Gamma_{abc}$  are antisymmetric in their first two indices  $ab$ . 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  $\Gamma_{abc}$ , but more generally the spatial connections can vary in space. For example, the spatial connections are spatially variable in all of the variants of the FLRW line element given in Chapter 10 (although FLRW spacetimes can be realized as Bianchi spacetimes with constant structure coefficients — see §17.5). The spatial connections  $\Gamma_{abc}$  also vary in time because, whereas  $c_{ab}^c$  with one index raised is constant in time, the coefficients  $c_{cab} \equiv \gamma_{cd}\gamma_{ab}^d$  with all indices lowered depend on time through the time-dependent metric  $\gamma_{cd}$ . Explicitly,

$$\Gamma_{abc} = \frac{1}{2}(\gamma_{cd}\gamma_{ab}^d + \gamma_{bd}\gamma_{ac}^d - \gamma_{ad}\gamma_{bc}^d). \quad (17.97)$$

The unique non-vanishing contraction of the spatial connections  $\Gamma_{abc}$  is

$$\Gamma_{ba}^a = c_{ab}^a = \epsilon_{abc}n^{ca} = -2k_b, \quad (17.98)$$

which is constant in time.

The extrinsic curvature  $K_{ab}$  is by definition

$$K_{ab} \equiv \Gamma_{a0b} = \frac{1}{2}\dot{\gamma}_{ab}, \quad (17.99)$$

with trace

$$K \equiv K_a^a = \frac{1}{2}\gamma^{ab}\dot{\gamma}_{ab} = \frac{d \ln \sqrt{\gamma}}{dt}, \quad (17.100)$$

where  $\gamma \equiv |\gamma_{ab}|$  is the determinant of the spatial tetrad metric. The last step of equations (17.100) is an application of equation (2.76). 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),

$$R_{a0b0} = -\dot{K}_{ab} + K_a^c K_{bc} , \quad (17.101a)$$

$$R_{abc0} = \Gamma_{cb}^d K_{da} - \Gamma_{ca}^d K_{db} + (\Gamma_{ab}^d - \Gamma_{ba}^d) K_{cd} , \quad (17.101b)$$

$$R_{abcd} = \hat{R}_{abcd} - K_{cb} K_{da} + K_{ca} K_{db} , \quad (17.101c)$$

where  $\hat{R}_{abcd}$  is the restricted Riemann tensor,

$$\hat{R}_{abcd} = \partial_a \Gamma_{cdb} - \partial_b \Gamma_{cda} + \Gamma_{cb}^e \Gamma_{eda} - \Gamma_{ca}^e \Gamma_{edb} + (\Gamma_{ab}^e - \Gamma_{ba}^e) \Gamma_{cde} . \quad (17.102)$$

If the structure coefficients are spatially constant, then the two derivative terms on the right hand side of equation (17.102) can be dropped. For spatially constant structure coefficients, equations (17.101) and (17.102) along with equations (17.96) give the Riemann tensor in terms of the structure coefficients  $c_{ab}^c$  and the tetrad metric  $\gamma_{ab}$ , without the need for an explicit form for the vierbein  $e_a^\alpha$ . 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 (17.94) 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 11.6.

Contracting the Riemann tensor yields the Ricci tensor  $R_{km}$ ,

$$R_{00} = -\dot{K} - K^{ab} K_{ab} , \quad (17.103a)$$

$$R_{a0} = \Gamma_{cb}^b K_a^c - \Gamma_{ab}^c K_c^b , \quad (17.103b)$$

$$R_{ab} = \hat{R}_{ab} + \dot{K}_{ab} - 2K_a^c K_{bc} + K_{ab} K , \quad (17.103c)$$

where  $\hat{R}_{ab}$  is the restricted Ricci tensor,

$$\hat{R}_{ab} = -\partial_a \Gamma_{bc}^c + \partial_c \Gamma_{ba}^c + \Gamma_{ba}^e \Gamma_{ed}^d - \Gamma_{ad}^e \Gamma_{be}^d . \quad (17.104)$$

Again, if the structure coefficients are spatially constant, then the two derivative terms on the right hand side of equation (17.104) can be dropped. And again, for spatially constant structure coefficients, the Jacobi identity (17.94) 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  $R$ ,

$$R = \hat{R} + 2\dot{K} + K^{ab} K_{ab} + K^2 , \quad (17.105)$$

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  $p_{ab}(t)$ ,

$$T_{00} = \rho, \quad T_{a0} = -f_a, \quad T_{ab} = p_{ab}. \quad (17.106)$$

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 (which is not being assumed here), the energy flux  $f_a$  would vanish, and the pressure tensor would be proportional to the metric tensor,  $p_{ab} = p\gamma_{ab}$ . The ADM equations of motion for a Bianchi spacetime are, equations (17.52) and (17.63),

$$\frac{d\gamma_{ab}}{dt} = 2K_{ab}, \quad (17.107a)$$

$$\frac{dK_{ab}}{dt} - 2K_a^c K_{bc} + K_{ab} K + \hat{R}_{ab} = 4\pi [2p_{ab} + \gamma_{ab}(\rho - 3p)]. \quad (17.107b)$$

The Hamiltonian constraint and the momentum constraints are

$$\frac{1}{2}(-K^{ab}K_{ab} + K^2 + \hat{R}) = 8\pi\rho, \quad (17.108a)$$

$$\Gamma_{cb}^b K_a^c - \Gamma_{ab}^c K_c^b = -8\pi f_a. \quad (17.108b)$$

Equations (17.107) 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 (17.107) is a tetrad ellipsoid that evolves in both size and rotation. Equation (17.103a) gives an equation for the evolution of the expansion rate  $K$  of the comoving volume element,

$$\dot{K} = -K^{ab}K_{ab} - 4\pi(\rho + 3p), \quad (17.109)$$

which is the same as the trace of the equation of motion (17.107b) minus twice the Hamiltonian constraint (17.108a). Equation (17.109) is the Raychaudhuri equation (17.76) in a Bianchi spacetime. Since the spatial metric  $\gamma_{ab}$  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 vierbein 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 VII<sub>0</sub>, yields a flat FLRW universe.

Bianchi spaces have spatially constant structure coefficients by assumption, but none of the various versions of the FLRW line element given in Chapter 10 have constant structure coefficients. The non-constancy of the structure coefficients poses no obstacle to casting the Friedmann equations into ADM form. For example, the isotropic (Poincaré) form (10.26) of the FLRW line element is

$$ds^2 = -dt^2 + \frac{4a^2}{[1 + \kappa(x^2 + y^2 + z^2)]^2}(dx^2 + dy^2 + dz^2), \quad (17.110)$$

which is in ADM form with zero lapse and shift. The line-element (17.110) takes ADM form (17.90) with spatial tetrad metric

$$\gamma_{ab} = a^2 \delta_{ab} , \quad (17.111)$$

and spatial vierbein

$$e_a{}^\alpha = \frac{1}{2} \delta_a^\alpha [1 + \kappa(x^2 + y^2 + z^2)] . \quad (17.112)$$

The structure coefficients, equation (17.93), have zero symmetric part, and non-constant antisymmetric part given by

$$k_e = \kappa x^e . \quad (17.113)$$

The extrinsic curvature  $K_{ab}$  is

$$K_{ab} = a \dot{a} \delta_{ab} , \quad (17.114)$$

and its trace  $K$  is 3 times the Hubble parameter,

$$K = \frac{3\dot{a}}{a} . \quad (17.115)$$

The restricted Ricci tensor  $\hat{R}_{ab}$  is

$$\hat{R}_{ab} = 2\kappa \delta_{ab} , \quad (17.116)$$

and the restricted Ricci scalar  $\hat{R}$  is

$$\hat{R} \equiv \frac{6\kappa}{a^2} . \quad (17.117)$$

The Hamiltonian constraint (17.126) is

$$\frac{3}{a^2} (\dot{a}^2 + \kappa) = 8\pi\rho , \quad (17.118)$$

which reproduces the first of the Friedmann equations (10.30). The equations of motion reduce to

$$\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + 2\frac{\kappa}{a^2} = 4\pi(\rho - p) . \quad (17.119)$$

With a factor of the Hamiltonian constraint (17.118) subtracted, the equation of motion (17.119) becomes

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p) , \quad (17.120)$$

which reproduces the second of the Friedmann equations (10.30). Equation (17.120) 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, so that there is 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 17.3. BKL call the phases between bounces Kasner epochs.

The simplest BKL model is where the axes of the tetrad ellipsoid  $\gamma_{ab}(t)$  of the Bianchi spacetime change in size, but they do not rotate. This is the model pursued in this section, and that you will explore in Exercise 17.6. 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 (17.107) for a Bianchi spacetime show that non-rotating solutions for the tetrad metric  $\gamma_{ab}$  exist if the restricted Ricci tensor  $\hat{R}_{ab}$  and the pressure tensor  $p_{ab}$  are diagonal in the frame where the tetrad metric is diagonal. For such solutions, the extrinsic curvature  $K_{ab}$  is diagonal in the frame where the tetrad metric is diagonal, and the momentum constraints (17.108b) then imply that the energy flux  $f_a$  vanishes. All Bianchi Types except IV include solutions for which the restricted Ricci tensor is diagonal.

The tetrad metric  $\gamma_{ab}$  in the non-rotating diagonal frame is conveniently written in terms of scale factors  $a_a$  along each of the three diagonal directions,

$$\gamma_{ab}(t) = a_a^2 \delta_{ab} . \quad (17.121)$$

The corresponding diagonal extrinsic curvature  $K_{ab}$  is then, from equation (17.107a),

$$K_{ab} = a_a \dot{a}_a \delta_{ab} . \quad (17.122)$$

The pressure is diagonal by assumption, with pressure  $p_a$  in the  $a$ 'th direction,

$$p_{ab} = p_a \delta_{ab} . \quad (17.123)$$

The equation of motion (17.107b) 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 (17.104), is given in terms of the structure coefficients  $c_{ab}^c$  and the tetrad metric  $\gamma_{ab}$  without the need for an explicit expression for the vierbein. In most (Type VI with  $k \neq 0$  is an exception) of the solutions for which  $\hat{R}_{ab}$  is diagonal in the frame where the metric is diagonal, including the BKL solutions, the symmetric part  $n^{(cd)}$  of the structure coefficients is diagonal in the same frame. In this case, the components of the restricted Ricci tensor  $\hat{R}_{ab}$  (17.104) are, in terms of the scale factors  $a_a$  and the parameters  $n_c$  and  $k_e$  of the structure coefficients, equation (17.92),

$$\hat{R}_{11} = a_1^2 \left( \frac{n_2 n_3 - 2k_1^2}{a_1^2} - \frac{2k_2^2}{a_2^2} - \frac{2k_3^2}{a_3^2} + \frac{n_1^2 a_1^2}{2a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{2a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{2a_1^2 a_2^2} \right), \quad (17.124a)$$

$$\hat{R}_{23} = k_1 \left( \frac{n_2 a_2^2 - n_3 a_3^2}{a_1^2} \right), \quad (17.124b)$$

and similarly with permuted indices for the other components. The off-diagonal components  $\hat{R}_{23}$  and company must vanish for the restricted Ricci tensor to be diagonal. Equation (17.124b) shows that one possibility, which covers the majority of cases (Type VII with  $k \equiv k_1 \neq 0$  and  $\sqrt{n_2} a_2 = \sqrt{n_3} a_3$  is an exception), is that the antisymmetric part  $k_e$  of the structure coefficients vanishes identically (that is,  $k \equiv k_1 = 0$ ). This is the solution pursued here, since it is the one that leads to the BKL solutions. In this case, where the symmetric part of the structure coefficients is diagonal in the frame where the metric is diagonal, and where the antisymmetric part vanishes identically, the ADM equations of motion (17.107) imply that the equation of motion for the scale factor  $a_1$  is

$$\frac{\ddot{a}_1}{a_1} + \frac{\dot{a}_1 \dot{a}_2}{a_1 a_2} + \frac{\dot{a}_1 \dot{a}_3}{a_1 a_3} + \frac{n_2 n_3}{a_1^2} + \frac{n_1^2 a_1^2}{2a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{2a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{2a_1^2 a_2^2} = \frac{R_{11}}{a_1^2} = 4\pi(2p_1 + \rho - 3p), \quad (17.125)$$

and like equations with permuted indices for  $a_2$  and  $a_3$ . The Hamiltonian constraint is

$$\frac{\dot{a}_2 \dot{a}_3}{a_2 a_3} + \frac{\dot{a}_3 \dot{a}_1}{a_3 a_1} + \frac{\dot{a}_1 \dot{a}_2}{a_1 a_2} + \frac{n_2 n_3}{2a_1^2} + \frac{n_3 n_1}{2a_2^2} + \frac{n_1 n_2}{2a_3^2} - \frac{n_1^2 a_1^2}{4a_2^2 a_3^2} - \frac{n_2^2 a_2^2}{4a_3^2 a_1^2} - \frac{n_3^2 a_3^2}{4a_1^2 a_2^2} = 8\pi\rho. \quad (17.126)$$

You will explore how these equations lead to BKL oscillatory collapse to a singularity in Exercise 17.6.

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 artifact of the assumption of spatial homogeneity, involves the dependence on time of the terms in the equations of motion (17.125) (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 (17.104). The form of the dependence of the restricted Ricci tensor on the scale factors follows from the fact that the restricted Ricci tensor (17.104) is proportional to two powers of the contravariant metric  $\gamma^{cd}$ , and two powers of the covariant metric  $\gamma_{cd}$ , 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.

Table 17.2 Bianchi spatial vierbein yielding a diagonal restricted Ricci tensor

Type	$e_a{}^\alpha$
I	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
V	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} & 0 \\ 0 & 0 & e^{kx} \end{pmatrix}$
II	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & x \\ 0 & 0 & 1 \end{pmatrix}$
VI <sub>0</sub>	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cosh x & \sinh x \\ 0 & \sinh x & \cosh x \end{pmatrix}$
III	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{2x} & e^{2x} \\ 0 & -1 & 1 \end{pmatrix}$
VI	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{(k+1)x} & e^{(k+1)x} \\ 0 & -e^{(k-1)x} & e^{(k-1)x} \end{pmatrix}$
VII <sub>0</sub>	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos x & \sin x \\ 0 & -\sin x & \cos x \end{pmatrix}$
VII (with $a_2 = a_3$ )	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{kx} \cos x & e^{kx} \sin x \\ 0 & -e^{kx} \sin x & e^{kx} \cos x \end{pmatrix}$
VIII	$\begin{pmatrix} 1 & 0 & 0 \\ -\sin x \tanh y & \cos x & \sin x \operatorname{sech} y \\ -\cos x \tanh y & -\sin x & \cos x \operatorname{sech} y \end{pmatrix}$
IX	$\begin{pmatrix} 1 & 0 & 0 \\ \sin x \tan y & \cos x & \sin x \sec y \\ \cos x \tan y & -\sin x & \cos x \sec y \end{pmatrix}$

Even though they are not needed in order to write down the Einstein equations, Table 17.2 lists explicit expressions for the spatial vierbein yielding a diagonal restricted Ricci tensor, which exist for all Bianchi Types except IV. The coordinates are scaled so that the eigenvalues of the structure coefficients are all  $n_a = 0$  or  $\pm 1$ . For the tabulated Types with  $k = 0$ , the time-space components  $R_{0a}$  of the Ricci tensor also vanish identically. For the tabulated Types with  $k \neq 0$ , the time-space components  $R_{0a}$  of the Ricci tensor do not all vanish identically, and their vanishing must be imposed as constraints on the initial conditions.

The notable exception mentioned at the beginning of §17.4 of a homogeneous space that cannot be realized 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

realized by the spatial vierbein

$$e_a{}^\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sec(\sqrt{\kappa}x) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (17.127)$$

An open ( $\kappa < 0$ ) cylindrical space on the other hand can be realized as a Bianchi space of Type III, with the spatial vierbein given in Table 17.2, with  $\kappa = -k/2 = -1/2$ .

**Exercise 17.3 Kasner spacetime.** The Kasner (1921) line element is

$$ds^2 = -dt^2 + a_x^2 dx^2 + a_y^2 dy^2 + a_z^2 dz^2, \quad (17.128)$$

where  $a_a(t)$  are functions only of time  $t$ . What Bianchi type is the Kasner line element (17.128)? Show that the Kasner line element (17.128) solves the vacuum Einstein equations if

$$a_a = |t|^{q_a} \quad (17.129)$$

with

$$q_x + q_y + q_z = 1, \quad q_x^2 + q_y^2 + q_z^2 = 1. \quad (17.130)$$

Show that a parametric solution of equation (17.130) is

$$q_x = \frac{-u}{1+u+u^2}, \quad q_y = \frac{1+u}{1+u+u^2}, \quad q_z = \frac{u(1+u)}{1+u+u^2}. \quad (17.131)$$

Plot the  $q_a$  versus  $u$ . Show that, if  $q_a$  are ordered such that  $q_1 \leq q_2 \leq q_3$ , then

$$-\frac{1}{3} \leq q_1 \leq 0 \leq q_2 \leq \frac{2}{3} \leq q_3 \leq 1. \quad (17.132)$$

**Solution.** Type I.

**Exercise 17.4 Schwarzschild interior as a Bianchi spacetime.** Inside the horizon of the Schwarzschild geometry, where the horizon function  $\Delta$  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

$$ds^2 = -dR^2 + |\Delta| dt^2 + r^2 (d\theta^2 + \sin^2\theta d\phi^2), \quad (17.133)$$

where  $dR \equiv dr/\sqrt{|\Delta|}$ . The line element (17.133) is in the form (17.90) with time coordinate  $R$ , spatial coordinates  $t, \theta, \phi$ , spatial tetrad metric

$$\gamma_{ab} = \text{diag}(|\Delta|, r^2, r^2), \quad (17.134)$$

and spatial vierbein and inverse vierbein

$$e_a{}^\alpha = \text{diag}(1, 1, 1/\sin\theta), \quad e^a{}_\alpha = \text{diag}(1, 1, \sin\theta). \quad (17.135)$$

Show that the Schwarzschild interior looks like a Kasner geometry near the singularity.

**Solution.** Type V. The interior near the singularity is Kasner (17.128) with  $t \propto r^{3/2}$ , and  $q_1 = -\frac{1}{3}$ ,  $q_2 = q_3 = \frac{2}{3}$ .

**Exercise 17.5 Kasner spacetime for a perfect fluid.** A generalization of the Kasner line element (17.128) is

$$ds^2 = -dt^2 + \sum_a a_a^2 dx_a^2 , \quad (17.136)$$

with scale factors

$$a_a = a T^{q_a - 1/3} , \quad (17.137)$$

where  $a(t)$  and  $T(t)$  are functions of time  $t$ , and the constants  $q_a$  are Kasner coefficients satisfying equation (17.130). The overall scale factor  $a$  satisfies

$$a \equiv (a_1 a_2 a_3)^{1/3} . \quad (17.138)$$

1. Show that the Einstein tensor corresponding to the Kasner line element (17.136) is diagonal.
2. Show that the energy-momentum is that of a perfect fluid (pressures  $p_a = p$  all equal) provided that  $a$  and  $T$  are related by

$$a = \left( 3K \frac{dt}{d \ln T} \right)^{1/3} , \quad (17.139)$$

where  $K$  is a constant.

3. Show that in this case of a perfect fluid the Einstein equations are

$$G_{00} = 3 \left( \frac{\dot{a}^2}{a^2} - \frac{K^2}{a^6} \right) = 8\pi\rho , \quad (17.140a)$$

$$G_{aa} = -\frac{2\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{3K^2}{a^6} = 8\pi p . \quad (17.140b)$$

The Einstein equations (17.140) resemble those (10.29) of the FLRW geometry except that the curvature terms  $\kappa/a^2$  in FLRW are replaced by terms proportional to  $-K^2/a^6$ .

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 (10.68) as for FLRW,

$$t = \int \frac{da}{aH} . \quad (17.141)$$

5. Show that the Einstein equations (17.140) enforce that the energy-momentum of the perfect fluid satisfies the first law of thermodynamics, similarly to FLRW, §10.9.2,

$$\frac{d\rho a^3}{dt} + p \frac{da^3}{dt} = 0 . \quad (17.142)$$

6. From the first law of thermodynamics, show that for a perfect fluid with equation of state  $p/\rho = w = \text{constant}$ , the density  $\rho$  is related to scale factor  $a$  by, as in FLRW,

$$\rho \propto a^{-3(1+w)} . \quad (17.143)$$

7. More generally, as in FLRW, the energy-momentum may comprise multiple perfect fluid components  $x$  satisfying the first law (17.142). The critical density  $\rho_{\text{crit}}$  is defined in terms of the Hubble parameter  $H$  in the usual way by equation (10.46). Argue that the Kasner Einstein equation (17.140a) implies that

$$\frac{3H^2}{8\pi} \equiv \rho_{\text{crit}} = \rho_K + \sum_{\text{species } x} \rho_x , \quad (17.144)$$

which differs from FLRW, equation (10.70), in that the FLRW curvature density  $\rho_k \propto a^{-2}$ , equation (10.48), is replaced by the Kasner curvature density  $\rho_K \propto a^{-6}$ ,

$$\rho_K \equiv \frac{3K^2}{8\pi a^6} . \quad (17.145)$$

The Kasner curvature density  $\rho_K$  behaves like a perfect fluid with an ultra-hard equation of state,  $w = 1$ .

8. Define  $a_K$  and  $H_K$  to be the cosmic scale factor and Hubble parameter at density-curvature equality, where  $\rho = \rho_K = \frac{1}{2}\rho_{\text{crit}}$ . Show that

$$K = \frac{a_K^3 H_K}{\sqrt{2}} . \quad (17.146)$$

9. From equation (17.139) conclude that

$$\ln T = 3K \int \frac{da}{a^4 H} . \quad (17.147)$$

Conclude that for a single perfect fluid with  $p/\rho = w = \text{constant}$ ,

$$T/T_K = \frac{(a/a_K)^3}{[1 + \sqrt{1 + (a/a_K)^{3(1-w)}}]^{2/(1-w)}} . \quad (17.148)$$

Conclude that the small and large  $a$  limits of  $T$  are, for  $w \leq 1$ ,

$$T/T_K \rightarrow \begin{cases} (a/a_K)^3 & a \ll a_K , \\ 1 & a \gg a_K . \end{cases} \quad (17.149)$$

Hence conclude that the perfect fluid Kasner solution goes over to vacuum Kasner for small  $a$  and to FLRW for large  $a$ .

10. For the particular case of a cosmological constant,  $w = -1$ , show that  $K = \sqrt{\Lambda/3}$ , and that

$$a/a_K = \sinh^{1/3}(\sqrt{3\Lambda} t) , \quad T/T_K = \tanh(\sqrt{3\Lambda} t/2) . \quad (17.150)$$

**Exercise 17.6 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

$$q_a \equiv \frac{d \ln a_a}{d \ln |t|} , \quad (17.151)$$

and let  $q$  be their sum,

$$q \equiv \sum q_a = \frac{d \ln(a_1 a_2 a_3)}{d \ln |t|} . \quad (17.152)$$

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

$$A_a \equiv n_a a_a^2 . \quad (17.153)$$

Show that, for vanishing energy-momentum, the equations of motion (17.125) are

$$\frac{dq_1}{d \ln |t|} + q_1(q-1) = \frac{1}{2} \left( \frac{t}{a_1 a_2 a_3} \right)^2 [(A_2 - A_3)^2 - A_1^2] , \quad (17.154a)$$

$$\frac{dq_2}{d \ln |t|} + q_2(q-1) = \frac{1}{2} \left( \frac{t}{a_1 a_2 a_3} \right)^2 [(A_1 - A_3)^2 - A_2^2] , \quad (17.154b)$$

$$\frac{dq_3}{d \ln |t|} + q_3(q-1) = \frac{1}{2} \left( \frac{t}{a_1 a_2 a_3} \right)^2 [(A_1 - A_2)^2 - A_3^2] , \quad (17.154c)$$

and that the Hamiltonian constraint (17.126) is

$$q^2 - \sum q_a^2 = \frac{1}{4} \left( \frac{t}{a_1 a_2 a_3} \right)^2 [2(A_1^2 + A_2^2 + A_3^2) - (A_1 + A_2 + A_3)^2] . \quad (17.155)$$

2. In gravitational collapse, the scale factors  $a_a$  might be expected to become small. Argue that if the right hand sides of equations (17.154) and (17.155) are neglected, then the solution is the Kasner solution, with  $q_a$  constant, satisfying equation (17.130).
3. In the Kasner solution, the  $q_a$  satisfy the inequalities (17.132). 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 (17.154) that the effect of 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. Choose either Type VIII or Type IX: they are equally fun. You will find better numerical behaviour by transforming to a time variable  $\tau$  defined by

$$\frac{d}{d\tau} \equiv a_1 a_2 a_3 \frac{d}{dt} = \frac{a_1 a_2 a_3}{t} \frac{d}{d \ln |t|} , \quad (17.156)$$

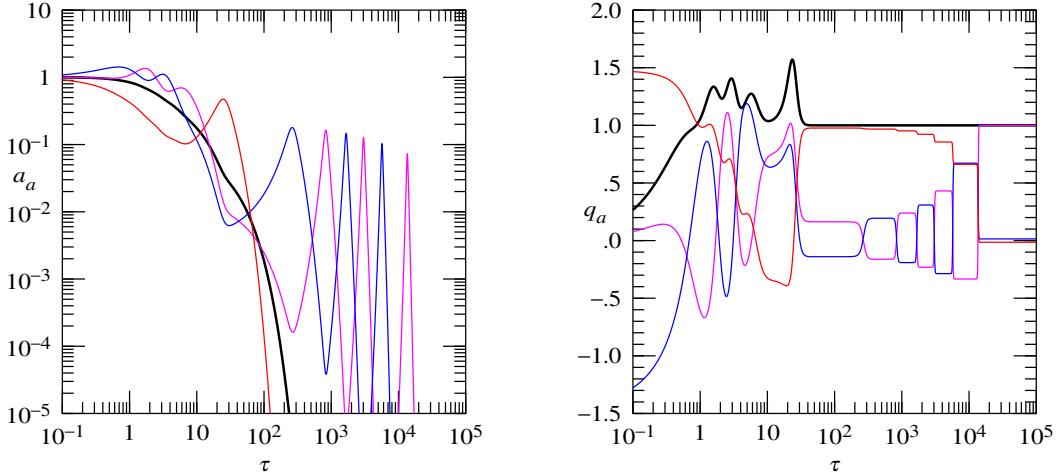


Figure 17.1 Left panel: Cosmic scale factors  $a_a$  in BKL collapse of a Bianchi Type IX spacetime (with eigenvalues normalized to  $n_a = 1$ ). The thick (black) line is the geometric average  $(a_1 a_2 a_3)^{1/3}$  of the scale factors, which is proportional to the cube root of the comoving volume element. Right panel: Logarithmic derivatives  $q_a$  of the scale factors, equation (17.151). 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 (17.159) 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  $a_1 a_2 a_3 \approx 10^{-230}$ .

which increases as  $t$  increases and  $\ln |t|$  decreases. Define

$$Q_a \equiv -\frac{1}{2} \frac{d \ln |A_a|}{d \tau} = \frac{a_1 a_2 a_3}{-t} q_a , \quad (17.157)$$

which has the same sign as  $q_a$ . Show that the equation of motion for  $A_1$  is

$$\frac{dQ_1}{d\tau} = \frac{1}{2} [A_1^2 - (A_2 - A_3)^2] , \quad (17.158)$$

and similarly for  $A_2$  and  $A_3$ . Show that the Hamiltonian constraint is

$$Q_2 Q_3 + Q_3 Q_1 + Q_1 Q_2 = \frac{1}{2} (A_1^2 + A_2^2 + A_3^2) - \frac{1}{4} (A_1 + A_2 + A_3)^2 . \quad (17.159)$$

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  $q = 1$ .

5. Show that for energy-momentum with equation of state  $p = w\rho$ , the proper energy density  $\rho$  varies as

$$\rho \propto (a_1 a_2 a_3)^{-(1+w)} . \quad (17.160)$$

Show that including energy-momentum in the equations of motion amounts to adding terms proportional

to  $(a_1 a_2 a_3)^2 \rho$  on the right hand sides of equations (17.158). By comparing these terms to the largest  $A_a$  terms on the right hand side, conclude that the influence of energy-momentum is sub-dominant 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 (17.131) for some  $u \geq 1$ . After the bounce, the exponents  $q_a$  satisfy the same equation with  $u$  flipped,

$$u \rightarrow -u . \quad (17.161)$$

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  $u \geq 1$ .

**Solution.** Figure 17.1 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.

**Exercise 17.7 Geodesics in Bianchi spacetimes.** Solve for the geodesics of particles in a Bianchi spacetime.

**Solution.** The effective Lagrangian of a particle can be taken to be

$$L = \frac{1}{2} \gamma_{mn} p^m p^n , \quad (17.162)$$

where  $p^m \equiv e^m_{\mu} dx^{\mu}/d\lambda$  is the tetrad-frame 4-momentum of the particle. There are 3 integrals of motion associated with the 3 Killing vectors, plus 1 integral of motion associated with conservation of rest mass  $m$ ,

$$p_a = \text{constant } (a = 1, 2, 3) , \quad p^n p_n = -m^2 . \quad (17.163)$$

The rest mass equation implies that the time component of the tetrad-frame 4-momentum is

$$p^0 = \sqrt{m^2 + \gamma^{ab} p_a p_b} . \quad (17.164)$$

The coordinate 4-momentum is

$$\frac{dx^{\mu}}{d\lambda} \equiv p^{\mu} = \{p^0, \gamma^{ab} e_a^{\alpha} p_b\} . \quad (17.165)$$


---

## 17.7 BSSN formalism

Numerical experiments during the 1990s established that the ADM equations, whether in the original form with momenta  $\pi_{ab}$ , 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, 1999), and variants thereof (Shinkai, 2009). The BSSN formalism differs from ADM in that it adjoins equations of motion (17.178) for a vector set of 3 BSSN momentum variables  $\hat{H}_{\alpha}$ , and treats the definition (17.177) 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 §??.

The superior numerical stability of the BSSN over the ADM formalism can be attributed to the fact that 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. 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).

### 17.7.1 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),

$$R_{\gamma t \alpha \beta} = \hat{R}_{\gamma t \alpha \beta} + K_{\beta t} K_{\alpha \gamma} - K_{\alpha t} K_{\beta \gamma}, \quad (17.166a)$$

$$\hat{R}_{\gamma t \alpha \beta} = \frac{\partial \hat{\Gamma}_{\alpha \beta t}}{\partial x^\gamma} - \frac{\partial \hat{\Gamma}_{\alpha \beta \gamma}}{\partial t} + \hat{\Gamma}_{\delta \alpha t} \hat{\Gamma}_{\beta \gamma}^\delta - \hat{\Gamma}_{\delta \beta t} \hat{\Gamma}_{\alpha \gamma}^\delta, \quad (17.166b)$$

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, for vanishing torsion. As shown below, equation (17.171), the index on the restricted connections in equation (17.166b) is raised with the spatial coordinate metric, not with the full metric,

$$\hat{\Gamma}_{\beta \mu}^\alpha \equiv g^{\alpha \gamma} \hat{\Gamma}_{\gamma \beta \mu}. \quad (17.167)$$

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  $\alpha$  and one time  $t$  index,

$$K_{\alpha t \nu} = \alpha K_{\alpha 0 \nu}, \quad (17.168)$$

which like the tetrad-frame generalized extrinsic curvature is antisymmetric in its first two indices  $\alpha t$ . 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  $K_\alpha \equiv K_{\alpha 00} \equiv e^a{}_\alpha K_{a00}$ . The tensor  $K_{\alpha t}$  in equation (17.166a) is

$$K_{\alpha t} \equiv K_{\alpha 0t} = e^m{}_t K_{\alpha 0m} = \alpha K_\alpha - \beta^\delta K_{\alpha\delta} . \quad (17.169)$$

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

$$\hat{\Gamma}_{\lambda\mu\nu} = e^l{}_\lambda e^m{}_\mu e^n{}_\nu (d_{lmn} + \hat{\Gamma}_{lmn}) . \quad (17.170)$$

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  $g^{\delta\epsilon}$ ,

$$\hat{\Gamma}_{\pi\alpha t} \hat{\Gamma}_{\beta\gamma}^\pi = \hat{\Gamma}_{p\alpha t} \hat{\Gamma}_{\beta\gamma}^p = \hat{\Gamma}_{d\alpha t} \hat{\Gamma}_{\beta\gamma}^d = \hat{\Gamma}_{\delta\alpha t} \hat{\Gamma}_{\beta\gamma}^\delta , \quad (17.171)$$

which justifies equation (17.167).

The coordinate connection  $\hat{\Gamma}_{[\alpha\beta]t}$  antisymmetrized over its spatial indices  $\alpha\beta$  is an antisymmetric spatial tensor, which can be denoted  $F_{\alpha\beta}$ ,

$$F_{\alpha\beta} \equiv \hat{\Gamma}_{[\alpha\beta]t} = \frac{1}{2} \left( \frac{\partial \beta_\beta}{\partial x^\alpha} - \frac{\partial \beta_\alpha}{\partial x^\beta} \right) . \quad (17.172)$$

The tensorial nature of  $F_{\alpha\beta}$  follows from the fact that the coordinate-frame curl of a vector is a tensor, Exercise 2.5. Expression (17.166b) for the restricted Riemann tensor can thus be written

$$\hat{R}_{\gamma t\alpha\beta} = \hat{D}_\gamma F_{\alpha\beta} - \frac{\partial \hat{\Gamma}_{[\alpha\beta]\gamma}}{\partial t} + \hat{\Gamma}_{(\delta\alpha)t} \hat{\Gamma}_{\beta\gamma}^\delta - \hat{\Gamma}_{(\delta\beta)t} \hat{\Gamma}_{\alpha\gamma}^\delta , \quad (17.173)$$

in which the only term containing mixed time-space second derivatives is  $\partial \hat{\Gamma}_{[\alpha\beta]\gamma} / \partial t$ . The coordinate connection  $\hat{\Gamma}_{(\alpha\beta)t}$  symmetrized over its spatial indices is

$$\hat{\Gamma}_{(\alpha\beta)t} = \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial t} . \quad (17.174)$$

Contracting the Riemann tensor  $R_{\gamma t\alpha\beta}$  yields the time-space components  $R_{t\alpha}$  of the Ricci tensor,

$$R_{t\alpha} = \hat{R}_{t\alpha} - K_t^\beta K_{\alpha\beta} + K_{\alpha t} K , \quad (17.175a)$$

$$\hat{R}_{t\alpha} = \hat{D}^\beta F_{\beta\alpha} + \frac{\partial \hat{\Gamma}_{[\alpha\beta]}^\beta}{\partial t} - \hat{\Gamma}_{(\delta\alpha)t} \hat{\Gamma}_{\beta\gamma}^\delta + \hat{\Gamma}^{(\delta\beta)}_t \hat{\Gamma}_{\alpha\delta\beta} , \quad (17.175b)$$

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

$$\hat{\Gamma}_{[\alpha\beta]}^{\beta} = \frac{1}{2}g^{\beta\gamma}\left(\frac{\partial g_{\alpha\gamma}}{\partial x^\beta} - \frac{\partial g_{\beta\gamma}}{\partial x^\alpha}\right) = -\frac{1}{2}\frac{g_{\alpha\gamma}}{g}\frac{\partial(gg^{\beta\gamma})}{\partial x^\beta}, \quad (17.176)$$

where  $g \equiv |g_{\alpha\beta}|$  is the determinant of the spatial metric. Equations (17.175) 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,

$$\hat{H}_\alpha \equiv \hat{\Gamma}_{\alpha\beta}^{\beta} + \frac{p}{2}\frac{\partial \ln g}{\partial x^\alpha} = 2\hat{\Gamma}_{[\alpha\beta]}^{\beta} + \frac{(1+p)}{2}\frac{\partial \ln g}{\partial x^\alpha} = -\frac{g_{\alpha\gamma}}{g^{(1-p)/2}}\frac{\partial(g^{(1-p)/2}g^{\beta\gamma})}{\partial x^\beta}, \quad (17.177)$$

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 (17.179) 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

$$\frac{1}{2}\frac{\partial \hat{H}_\alpha}{\partial t} = -\frac{(1+p)}{4}\frac{\partial^2 \ln g}{\partial t \partial x^\alpha} + \hat{\Gamma}_{(\delta\alpha)t}\hat{\Gamma}^{[\delta\beta]}_\beta - \hat{\Gamma}^{(\delta\beta)}_t\hat{\Gamma}_{\alpha\delta\beta} + \hat{D}^\beta F_{\alpha\beta} + K_t^\beta K_{\alpha\beta} - K_{\alpha t}K + 8\pi T_{ta}. \quad (17.178)$$

In the BSSN formalism, equation (17.177) is a constraint equation, which must be imposed in the initial conditions, but which is satisfied automatically thereafter.

### 17.7.2 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  $\hat{H}_\alpha$ ,

$$\hat{R}_{\alpha\beta} = -\frac{p}{2}\frac{\partial^2 \ln g}{\partial x^\alpha \partial x^\beta} - \frac{g^{\gamma\delta}}{2}\frac{\partial^2 g_{\alpha\beta}}{\partial x^\gamma \partial x^\delta} + \frac{1}{2}\frac{\partial \hat{H}_\beta}{\partial x^\alpha} + \frac{1}{2}\frac{\partial \hat{H}_\alpha}{\partial x^\beta} - \hat{\Gamma}_{\alpha\beta}^\delta \hat{\Gamma}_{\delta\gamma}^\gamma + \hat{\Gamma}^{\gamma\delta}{}_\alpha \hat{\Gamma}_{\beta\gamma\delta} + \hat{\Gamma}^{\gamma\delta}{}_\beta \hat{\Gamma}_{\alpha\gamma\delta} + \hat{\Gamma}^{\gamma\delta}{}_\alpha \hat{\Gamma}_{\gamma\delta\beta}. \quad (17.179)$$

The only explicit second spatial derivatives in the expression (17.179) 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 (17.179) 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

$$\frac{1}{2} \left[ \left( \frac{\partial}{\partial t} \right)^2 - g^{\gamma\delta} \frac{\partial^2}{\partial x^\gamma \partial x^\delta} \right] g_{\alpha\beta} + \frac{\partial^2 \ln(\alpha g^{-p/2})}{\partial x^\alpha \partial x^\beta} + \dots = 8\pi \left( T_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} T \right), \quad (17.180)$$

where ... signifies terms involving no higher than first time or space derivatives of the lapse  $\alpha$ , the shift  $\beta^\alpha$ , 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).

### 17.7.3 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 (17.178) 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 (17.177) 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.

## 17.8 Pretorius formalism

Pretorius (2005b) 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 (17.177) is (with  $p = 0$ )

$$H_\kappa \equiv \Gamma_{\kappa\lambda}^\lambda = 2\Gamma_{[\kappa\lambda]}^\lambda + \frac{\partial \ln \sqrt{-g}}{\partial x^\kappa} = -\frac{g_{\kappa\mu}}{\sqrt{-g}} \frac{\partial(\sqrt{-g} g^{\lambda\mu})}{\partial x^\lambda}, \quad (17.181)$$

in which  $g$  is the determinant of the full 4-dimensional metric. The contravariant components  $H^\kappa$  can be written as minus the d'Alembertian of the coordinates,

$$H^\kappa = -\frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} g^{\lambda\kappa})}{\partial x^\lambda} = -\frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^\lambda} \left( \sqrt{-g} g^{\lambda\mu} \frac{\partial x^\kappa}{\partial x^\mu} \right) = -\square x^\kappa , \quad (17.182)$$

which motivates calling  $H^\kappa$  the **harmonic function**. In the Pretorius formalism, the Ricci tensor takes the form

$$R_{\kappa\lambda} = -\frac{1}{2} g^{\mu\nu} \frac{\partial^2 g_{\kappa\lambda}}{\partial x^\mu \partial x^\nu} + \frac{1}{2} \frac{\partial H_\lambda}{\partial x^\kappa} + \frac{1}{2} \frac{\partial H_\kappa}{\partial x^\lambda} - \Gamma_{\kappa\lambda}^\nu H_\nu + \Gamma^{\mu\nu}{}_\kappa \Gamma_{\lambda\mu\nu} + \Gamma^{\mu\nu}{}_\lambda \Gamma_{\kappa\mu\nu} + \Gamma^{\mu\nu}{}_\kappa \Gamma_{\mu\nu\lambda} , \quad (17.183)$$

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

$$R_{\kappa\lambda} = 8\pi (T_{\kappa\lambda} - \frac{1}{2} g_{\kappa\lambda} T) . \quad (17.184)$$

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  $\gamma_0$  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  $H_\kappa$  defined by equation (17.181) is, in terms of the vierbein derivatives  $d_{klm}$  defined by equation (11.32), the tetrad-frame restricted connections  $\hat{\Gamma}_{klm}$ , and the generalized extrinsic curvature  $K_{lmn}$ ,

$$H_k \equiv e_k{}^\kappa H_\kappa = d_{km}{}^m + \Gamma_{km}{}^m = d_{km}{}^m + \hat{\Gamma}_{km}{}^m + K_{km}{}^m = d_{k0}{}^0 + \hat{H}_k - K_k , \quad (17.185)$$

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 (17.177) with  $p = 0$ . The tetrad-frame components  $H_k$  of the harmonic function are

$$H_0 = \frac{1}{\alpha} \partial_0 \alpha - K , \quad (17.186a)$$

$$H_a = \frac{1}{\alpha} e_{a\alpha} \partial_0 \beta^\alpha + \hat{H}_a - K_a . \quad (17.186b)$$

Pretorius (2005b) points out that the arbitrariness of the choice of coordinates  $x^\kappa$  translates into an arbitrariness in the choice of the 4 components  $H_\kappa$  of the harmonic function. Thus instead of treating the lapse and shift as arbitrarily adjustable functions, the harmonic functions  $H_\kappa$  can be adjusted arbitrarily. For example, the harmonic function can be chosen to vanish identically,  $H_\kappa = 0$ . Equations (17.186) can then be interpreted as evolution equations for the lapse  $\alpha$  and the shift  $\beta^\alpha$ . In this case the 4 Einstein equations with at least one temporal index are not used as evolution equations.

However, it is also possible to follow the BSSN strategy of choosing the lapse and shift arbitrarily, in which case the 4 Einstein equations with at least one temporal index provide evolution equations for the harmonic function  $H_\kappa$ , and equations (17.186) 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 (17.186), must be arranged to be satisfied on the initial hypersurface.

## 17.9 $M+N$ 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  $\gamma_v$  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  $\gamma_v$  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  $\gamma_v$  and  $\gamma_u$ , and a 2-dimensional space orthogonal to the null directions.

This section §17.9 considers the general case of an  $M+N$  split of an  $M+N$ -dimensional spacetime.

### 17.9.1 $M+N$ tetrad and extrinsic curvature

In an  $M+N$  split of spacetime, the tetrad-frame axes  $\gamma_m$  at each point are split into two orthogonal sets, of dimensions respectively  $N$  and  $M$ . Label the  $N$  tetrad axes  $\gamma_z$  of the first set with late letters  $z$ , and the  $M$  tetrad axes  $\gamma_a$  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

$$\gamma_a \cdot \gamma_z = 0 . \quad (17.187)$$

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  $\Gamma_{azm}$  with first two indices  $az$  from opposite subspaces form a tensor, the generalized extrinsic curvature  $K_{azm}$ ,

$$K_{azm} \equiv \Gamma_{azm} = \gamma_a \cdot \partial_m \gamma_z . \quad (17.188)$$

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

$$\gamma_a \cdot \gamma_y \partial_m L_z^y = 0 , \quad (17.189)$$

which vanishes because  $\gamma_a$  and  $\gamma_y$  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  $\Gamma_{mnl}$ , namely those with first two indices  $mn$  from the same subspace, constitute the restricted connections  $\hat{\Gamma}_{mnl}$ ,

$$\hat{\Gamma}_{mnl} \equiv \Gamma_{mnl} \quad \text{for } mn = yz \text{ or } mn = ab . \quad (17.190)$$

The vanishing of the mixed components  $\gamma_{az}$  of the tetrad metric implies that the generalized extrinsic curvature is antisymmetric in its first two indices,

$$K_{zal} = -K_{azl} . \quad (17.191)$$

The vanishing components of  $K_{mnl}$  and  $\hat{\Gamma}_{mnl}$  are

$$K_{abl} = K_{yzl} = 0 , \quad \hat{\Gamma}_{azl} = 0 . \quad (17.192)$$

### 17.9.2 $M+N$ Riemann and Ricci tensors

The extrinsic curvature  $K_{mnl}$  is a tensor under the restricted group of tetrad transformations. The restricted Riemannn curvature tensor  $\hat{R}_{klaz}$  with its last two indices from opposite subspaces vanishes since  $\hat{\Gamma}_{azk}$  vanishes,

$$\hat{R}_{klaz} = \partial_k \hat{\Gamma}_{azl} - \partial_l \hat{\Gamma}_{azk} + \hat{\Gamma}_{al}^p \hat{\Gamma}_{pzk} - \hat{\Gamma}_{ak}^p \hat{\Gamma}_{pzl} + (\Gamma_{kl}^p - \Gamma_{lk}^p) \hat{\Gamma}_{azp} = 0 . \quad (17.193)$$

If torsion vanishes, then the full Riemann curvature tensor  $R_{klmn}$  is symmetric in  $kl \leftrightarrow mn$ , but the restricted Riemann tensor  $\hat{R}_{klmn}$  is not symmetric. 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

$$R_{wxyz} = \hat{R}_{wxyz} + K_{yx}^a K_{azw} - K_{yw}^a K_{azz} , \quad (17.194a)$$

$$R_{xyaz} = \hat{D}_x K_{azy} - \hat{D}_y K_{azx} + (K_{xy}^c - K_{yx}^c) K_{azc} \quad (17.194b)$$

$$= R_{azxy} = \hat{R}_{azxy} + K_{xz}^c K_{cyx} - K_{xa}^c K_{cyz} , \quad (17.194c)$$

$$R_{byaz} = \hat{D}_b K_{azy} - \hat{D}_y K_{azb} + K_{by}^x K_{azz} - K_{yb}^c K_{azc} , \quad (17.194d)$$

$$R_{bcaz} = \hat{D}_b K_{azc} - \hat{D}_c K_{azb} + (K_{bc}^x - K_{cb}^x) K_{azz} \quad (17.194e)$$

$$= R_{azbc} = \hat{R}_{azbc} + K_{bz}^x K_{xca} - K_{ba}^x K_{xcz} , \quad (17.194f)$$

$$R_{abcd} = \hat{R}_{abcd} + K_{cb}^z K_{zda} - K_{ca}^z K_{zdb} . \quad (17.194g)$$

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 become identities. The Ricci tensor  $R_{km}$  is

$$R_{yz} = \hat{R}_{yz} + (\hat{D}_a + K_a) K_{zy}^a - \hat{D}_y K_z - K_{ya}^b K_{zb} , \quad (17.195a)$$

$$R_{za} = \hat{R}_{zba}^b + (\hat{D}_y + K_y) K_{az}^y - \hat{D}_z K_a - K_{ab}^y K_{zy}^b \quad (17.195b)$$

$$= R_{az} = \hat{R}_{ayz}^y + (\hat{D}_b + K_b) K_{za}^b - \hat{D}_a K_z - K_{ab}^y K_{zy}^b \quad (17.195c)$$

$$= \hat{D}_y K_{az}^y - \hat{D}_z K_a + \hat{D}_b K_{za}^b - \hat{D}_a K_z - 2K_{ab}^y K_{zy}^b + K_{ab}^y K_{yz}^b + K_{ba}^y K_{zy}^b \quad (17.195d)$$

$$R_{ab} = \hat{R}_{ab} + (\hat{D}_z + K_z) K_{ba}^z - \hat{D}_a K_b - K_{ay}^z K_{bz}^y . \quad (17.195e)$$

Contracting the Ricci tensor yields the Ricci scalar  $R$ ,

$$R = \hat{R} - 2\hat{D}_z K^z - 2\hat{D}_a K^a - K^{bza} K_{azb} - K^{zay} K_{yaz} - K^z K_z - K^a K_a . \quad (17.196)$$

**17.10 2+2 split**

For the particular case of a 2+2 split, equations (17.195) for the Ricci tensor  $R_{km}$  become

$$R_{vu} = \hat{R}_{vu} - \hat{D}_v K_u + (\hat{D}_a + K_a) K_{uv}^a - K_{va}^b K_{ub}^a , \quad (17.197a)$$

$$R_{vv} = -\hat{D}_v K_v + (\hat{D}_a + K_a) K_{vv}^a - K_{va}^b K_{vb}^a , \quad (17.197b)$$

$$R_{v+} = \hat{R}_{v++-} + \hat{D}_v K_{v+u} - \hat{D}_u K_{v+v} + K_y K_{+v}^y - K_{+b}^y K_{vy}^b \quad (17.197c)$$

$$= R_{+v} = -\hat{R}_{+vvu} - \hat{D}_+ K_{+v-} + \hat{D}_- K_{+v+} + K_b K_{v+}^b - K_{+b}^y K_{vy}^b \quad (17.197d)$$

$$= \hat{D}_v K_{v+u} - \hat{D}_u K_{v+v} - \hat{D}_+ K_{+v-} + \hat{D}_- K_{+v+} - 2K_{+b}^y K_{vy}^b + K_{+b}^y K_{yy}^b + K_{b+}^y K_{vy}^b , \quad (17.197e)$$

$$R_{++} = (\hat{D}_z + K_z) K_{++}^z - \hat{D}_+ K_+ - K_{+y}^z K_{+z}^y , \quad (17.197f)$$

$$R_{+-} = \hat{R}_{+-} + (\hat{D}_z + K_z) K_{-+}^z - \hat{D}_+ K_- - K_{+y}^z K_{-z}^y . \quad (17.197g)$$

# 18

---

## 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 (1997), 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, §§18.6 and 18.7. The Chapter concludes, §18.9, with a brief exposition of the original singularity theorem discovered by Penrose (1965).

### 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  $\lambda$  along the path. The derivative of the path defines the 4-velocity  $u^\mu$  along the path,

$$u^\mu \equiv \frac{dx^\mu}{d\lambda} . \quad (18.1)$$

If the congruence of paths is timelike, then the parameter  $\lambda$  may be taken equal to the proper time  $\tau$  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  $\boldsymbol{\gamma}_0$ ,

$$\boldsymbol{\gamma}_0 = \mathbf{u} . \quad (18.2)$$

The tetrad time vector  $\boldsymbol{\gamma}_0$  is the unique future-pointing vector that is tangent to the timelike path and normalized to  $\boldsymbol{\gamma}_0 \cdot \boldsymbol{\gamma}_0 = -1$ .

If the congruence of paths is null, then  $\lambda$  may be any arbitrary parameter, not necessarily an affine parameter. If the parameter  $\lambda$  is an affine parameter, then the path is said to be affinely parametrized. The 4-velocity  $u^\mu \equiv dx^\mu/d\lambda$  satisfies the normalization condition  $u_\mu u^\mu = 0$  regardless of whether the parameter  $\lambda$  is affine. The 4-velocity vector  $\mathbf{u} \equiv e_\mu u^\mu$  defines the tetrad null vector  $\boldsymbol{\gamma}_v$  (say),

$$\boldsymbol{\gamma}_v = \mathbf{u} . \quad (18.3)$$

Unlike the timelike case, the normalization condition  $\boldsymbol{\gamma}_v \cdot \boldsymbol{\gamma}_v = 0$  does not determine uniquely the null vector  $\boldsymbol{\gamma}_v$ .

For either a timelike or a null path, the 4-velocity  $\mathbf{u} = u^m \boldsymbol{\gamma}_m$  has tetrad-frame components

$$u^m = \{1, 0, 0, 0\} , \quad (18.4)$$

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

$$D_n u_m = \partial_n u_m - \Gamma_{mn}^k u_k = \Gamma_{mzn} . \quad (18.5)$$

The components for spatial  $m = a$  constitute by definition the generalized extrinsic curvature  $K_{azn}$ , equation (17.188),

$$D_n u_a = K_{azn} . \quad (18.6)$$

The 4-velocity along the path evolves as

$$\frac{Du^k}{D\lambda} = u^n \partial_n u^k + \Gamma_{mn}^k u^m u^n = \Gamma_{zz}^k , \quad (18.7)$$

whose spatial components constitute the acceleration  $K_{zz}^a$ ,

$$\frac{Du^a}{D\lambda} = K_{zz}^a . \quad (18.8)$$

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  $Du^v/D\lambda = \Gamma_{vv}^v = -\Gamma_{vvv}$  vanishes if the path is affinely parametrized, but not in general. If the null path is affinely parametrized, 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 parametrized amounts to choosing the null vector  $\gamma_v$  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 parametrized path.

The covariant divergence of the 4-velocity is

$$D_m u^m = \Gamma_{zz}^z + K_z . \quad (18.9)$$

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  $\Gamma_{vv}^v$  plus the trace  $K_v \equiv K_{va}^a$  of the extrinsic curvature. If the null path is affinely parametrized, 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 (18.10), 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  $\gamma_0$  tangent to the timelike paths, equation (18.2). 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  $\gamma_v$  tangent to the null paths. To cover both timelike and null cases at the same time, denote the tangent axis by  $\gamma_z$ , 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 (17.194) for the Riemann tensor in an  $M+N$  split of spacetime, §17.9. 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  $R_{bzaz}$  of the Riemann tensor, equation (17.194d),

$$\hat{D}_z K_{azb} - \hat{D}_b K_{azz} - K_{bz}^y K_{azy} + K_{zb}^c K_{azc} = -R_{bzaz} \quad (\text{no sum over } z) , \quad (18.10)$$

with  $z = 0$  for a timelike congruence, or  $z = v$  for a null congruence. Equation (18.10) 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  $R_{bzaz}$  encountered along the path.

The left hand side of equation (18.10) 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

$$K_{azz} = 0 . \quad (18.11)$$

For a geodesic congruence (not necessarily affinely parametrized), the Raychaudhuri equation (18.10) becomes

$$\hat{D}_z K_{azb} - K_{bz}^y K_{azy} + K_{zb}^c K_{azc} = -R_{bzaz} \quad (\text{no sum over } z). \quad (18.12)$$

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

$$\Gamma_{klz} = 0. \quad (18.13)$$

The conditions (18.13) 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 (18.12) reduces to the directed derivative  $\partial_z$ , and the equation becomes

$$\partial_z K_{azb} - K_{bz}^y K_{azy} + K_{zb}^c K_{azc} = -R_{bzaz} \quad (\text{no sum over } z). \quad (18.14)$$

### 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 (18.12) become

$$\hat{D}_0 K_{ab} + K^c_b K_{ac} = -R_{b0a0}. \quad (18.15)$$

In 4-dimensional spacetime, the 9 components of the extrinsic curvature  $K_{ab}$  are commonly resolved into an **expansion** scalar  $\vartheta$ , a 3-component antisymmetric **vorticity** tensor  $\varpi_{ab}$ , and a 5-component traceless symmetric **shear** tensor  $\sigma_{ab}$ ,

$$K_{ab} = \delta_{ab}\vartheta + \varpi_{ab} + \sigma_{ab}. \quad (18.16)$$

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 (18.5). The expansion  $\vartheta$  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 §18.6), 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  $\tau$  along with 3 spatial coordinates  $x^\alpha$  that remain constant along the geodesics of the congruence. The comoving coordinate 4-velocity along geodesics is  $u^\mu = \{1, 0, 0, 0\}$ . The 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 (18.9). The expansion  $\vartheta$  thus satisfies

$$3\vartheta = K = D_m u^m = D_{\mu} x^{\mu} = \frac{1}{\sqrt{g}} \frac{\partial(\sqrt{g} u^{\mu})}{\partial x^{\mu}} = u^{\mu} \frac{\partial \ln \sqrt{g}}{\partial x^{\mu}} = \frac{d \ln \sqrt{g}}{d\tau} , \quad (18.17)$$

where  $\sqrt{g} = e^{-1}$  is the square root of the determinant of the spatial metric of the comoving line element, which is the same as the determinant  $e^{-1}$  of the inverse vierbein.

In the ADM formalism, the tetrad time vector  $\gamma_0$  is chosen to be orthogonal to hypersurfaces of constant time  $t$ . If  $\gamma_0$  is so chosen, and if torsion vanishes as general relativity assumes, then vorticity  $\varpi_{ab}$  vanishes, as shown in §18.6, equation (18.38). 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  $\alpha$  is constant, equation (18.38). In the ADM formalism, the expansion satisfies equation (17.60), which reduces to equation (18.17) 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  $\tau$ .

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 Raychaudhuri equations (18.15) for the expansion, vorticity, and shear along a timelike geodesic congruence are

$$\hat{D}_0 \vartheta + \vartheta^2 + \frac{1}{3} \sigma^{ab} \sigma_{ab} - \frac{1}{3} \varpi^{ab} \varpi_{ab} = -\frac{1}{3} R_{00} , \quad (18.18a)$$

$$(\hat{D}_0 + 2\vartheta) \varpi_{ab} + \sigma^c{}_a \varpi_{cb} - \sigma^c{}_b \varpi_{ca} = 0 , \quad (18.18b)$$

$$(\hat{D}_0 + 2\vartheta) \sigma_{ab} + (\sigma^c{}_a \sigma_{cb} - \frac{1}{3} \delta_{ab} \sigma^{cd} \sigma_{cd}) - (\varpi^c{}_a \varpi_{cb} - \frac{1}{3} \delta_{ab} \varpi^{cd} \varpi_{cd}) = -C_{0a0b} , \quad (18.18c)$$

where  $C_{klmn}$  is the Weyl tensor, the traceless part of the Riemann tensor. The restricted derivatives in equations (18.18) are

$$\hat{D}_0 \vartheta = \partial_0 \vartheta , \quad (18.19a)$$

$$\hat{D}_0 \varpi_{ab} = \partial_0 \varpi_{ab} - \Gamma^c_{a0} \varpi_{cb} - \Gamma^c_{b0} \varpi_{ac} , \quad (18.19b)$$

$$\hat{D}_0 \sigma_{ab} = \partial_0 \sigma_{ab} - \Gamma^c_{a0} \sigma_{cb} - \Gamma^c_{b0} \sigma_{ac} . \quad (18.19c)$$

If the tetrad is chosen to be parallel-transported along the geodesic, then all 6 of the tetrad connections with final index 0 vanish,

$$\Gamma_{kl0} = 0 , \quad (18.20)$$

including not only the 3 components  $K_a \equiv K_{a00}$  of the acceleration, but also the 3 restricted connections  $\Gamma_{ab0}$ . 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,

$$\hat{D}_0 = \partial_0 = \frac{d}{d\tau} . \quad (18.21)$$

**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

$$\hat{D}_0\vartheta + \vartheta^2 + \frac{1}{3}\sigma^{ab}\sigma_{ab} - \frac{1}{3}\varpi^{ab}\varpi_{ab} - \frac{1}{3}\hat{D}^aK_a - \frac{1}{3}K^aK_a = -\frac{1}{3}R_{00}, \quad (18.22a)$$

$$(\hat{D}_0 + 2\vartheta)\varpi_{ab} + \sigma^c{}_a\varpi_{cb} - \sigma^c{}_b\varpi_{ca} + \frac{1}{2}(\hat{D}_aK_b - \hat{D}_bK_a) = 0, \quad (18.22b)$$

$$\begin{aligned} (\hat{D}_0 + 2\vartheta)\sigma_{ab} + \sigma^c{}_a\sigma_{cb} - \varpi^c{}_a\varpi_{cb} - \frac{1}{2}(\hat{D}_aK_b + \hat{D}_bK_a) - K_aK_b \\ - \frac{1}{3}\delta_{ab}(\sigma^{cd}\sigma_{cd} - \varpi^{cd}\varpi_{cd} - \hat{D}^cK_c - K^cK_c) = -C_{0a0b}, \end{aligned} \quad (18.22c)$$

with the restricted covariant derivatives given by equations (18.19). If the acceleration is the gradient of a potential,  $K_a = \partial_a \ln \alpha$ , and if torsion vanishes as general relativity assumes, then  $\hat{D}_aK_b - \hat{D}_bK_a = 0$ , 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  $\{\boldsymbol{\gamma}_v, \boldsymbol{\gamma}_u, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$ , with two null directions at each point, an “outgoing” direction  $\boldsymbol{\gamma}_v$ , and an “ingoing” direction  $\boldsymbol{\gamma}_u$ . The spin axes  $\boldsymbol{\gamma}_+$  and  $\boldsymbol{\gamma}_-$  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 (17.188). If the null congruence along the outgoing  $v$ -direction is geodesic, then the acceleration  $K_{avv}$  vanishes. Along outgoing null geodesics, the Raychaudhuri equations (18.12) are

$$\hat{D}_v K_{avb} + K_{vb}^c K_{avc} = -R_{bvav}. \quad (18.23)$$

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 (18.23) reduces to the directed derivative  $\partial_v$ . This requires that the null axis  $\boldsymbol{\gamma}_v$  and the 2 spatial axes  $\boldsymbol{\gamma}_\pm$  (but not the null axis  $\boldsymbol{\gamma}_u$ ) of the tetrad be parallel-transported along the null geodesic congruence. Parallel-transport of  $\boldsymbol{\gamma}_v$  and  $\boldsymbol{\gamma}_\pm$  amounts to imposing that 4 of the 6 tetrad connections vanish,

$$\Gamma_{uvv} = \Gamma_{+-v} = K_{+vv} = K_{-vv} = 0. \quad (18.24)$$

The condition  $\Gamma_{uvv} = 0$  is the condition that the geodesics along  $\boldsymbol{\gamma}_v$  be affinely parametrized, 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 (18.24), the restricted covariant derivative in the Raychaudhuri equation (18.23) equals a derivative with respect to an affine parameter  $\lambda$  along the null geodesic,

$$\hat{D}_v = \partial_v = \gamma_v \cdot \partial = \mathbf{u} \cdot \partial = \frac{dx^{\mu}}{d\lambda} \frac{\partial}{\partial x^{\mu}} = \frac{d}{d\lambda}. \quad (18.25)$$

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-Linquist 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

$$K_{+vv} = K_{-vv} = K_{+uu} = K_{-uu} = 0. \quad (18.26)$$

If the ingoing null direction is geodesic, then the Raychaudhuri equations along the ingoing null geodesic are the same as equations (18.23) 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 parametrized in either the  $\gamma_v$  or the  $\gamma_u$  direction (that is, either  $\Gamma_{uvv}$  or  $\Gamma_{vuu}$  vanishes), but in general it is not possible to arrange that both null directions are affinely parametrized. 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  $\gamma_v$  or the  $\gamma_u$  direction (that is, either  $\Gamma_{+-v}$  or  $\Gamma_{+-u}$  vanishes), but in general it is not possible to arrange that the spatial axes are parallel-transported along both null directions.

The Raychaudhuri equations (18.23) 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 (18.16), these 4 components are commonly decomposed into an **expansion** scalar  $\vartheta$ , an antisymmetric **vorticity** tensor  $\varpi_{ab} \equiv \varepsilon_{ab}\varpi$ , and a traceless symmetric **shear** tensor  $\sigma_{ab}$ ,

$$K_{avb} = \gamma_{ab}\vartheta + \varepsilon_{ab}\varpi + \sigma_{ab}. \quad (18.27)$$

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  $\gamma_+$  and  $\gamma_-$  are complex conjugates of each other. The totally antisymmetric tensor  $\varepsilon_{ab}$  in 2-dimensional spin space flips sign under complex conjugation, so is purely imaginary,  $\varepsilon_{+-} = i$ . The expansion and vorticity scalars  $\vartheta$  and  $\varpi$  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 (18.17), 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  $\gamma_u$  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  $\gamma_v$  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  $\lambda$  along with 2 spatial coordinates  $x^\alpha$  that remain constant along the geodesics of the congruence. The coordinate 4-velocity is  $u^\mu = \{1, 0, 0\}$ . Then analogously to equation (18.17) the null expansion satisfies, from equation (18.9) with  $\Gamma_{vv}^v = 0$  because the congruence is being taken to be affinely parametrized,

$$2\vartheta = K_v = D_m u^m = D_\mu x^\mu = \frac{1}{\sqrt{g}} \frac{\partial(\sqrt{g} u^\mu)}{\partial x^\mu} = u^\mu \frac{\partial \ln \sqrt{g}}{\partial x^\mu} = \frac{d \ln \sqrt{g}}{d\lambda} , \quad (18.28)$$

where  $g$  is the determinant of 2-dimensional spatial metric of the comoving line element. Thus the null expansion  $\vartheta$  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 (18.23) along the outgoing null geodesic direction  $v$  are

$$(\hat{D}_v + \vartheta)\vartheta - \varpi^2 + \sigma_{++}\sigma_{++}^* = -4\pi T_{vv} , \quad (18.29a)$$

$$(\hat{D}_v + 2\vartheta)\varpi = 0 , \quad (18.29b)$$

$$(\hat{D}_v + 2\vartheta)\sigma_{++} = -C_{v+v+} . \quad (18.29c)$$

The restricted covariant derivatives in equations (18.29) are

$$\hat{D}_v \vartheta = (\partial_v + \Gamma_{uvv})\vartheta , \quad (18.30a)$$

$$\hat{D}_v \varpi = (\partial_v + \Gamma_{uvv})\varpi , \quad (18.30b)$$

$$\hat{D}_v \sigma_{++} = (\partial_v + \Gamma_{uvv} + 2\Gamma_{+-v})\sigma_{++} . \quad (18.30c)$$

## 18.5 Sachs optical coefficients

If the null axis  $\gamma_v$  and the two spatial axes  $\gamma_\pm$  are taken to be parallel-transported along the null geodesic directions  $\gamma_v$  of the congruence, then the tetrad connections  $\Gamma_{uvv}$  and  $\Gamma_{+-v}$  in equations (18.29) vanish. In this case the expansion  $\vartheta$ , vorticity  $\varpi$ , 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 (18.29) simplify to

$$(\partial_v + \vartheta)\vartheta - \varpi^2 + \sigma\sigma^* = -4\pi T_{vv} , \quad (18.31a)$$

$$(\partial_v + 2\vartheta)\varpi = 0 , \quad (18.31b)$$

$$(\partial_v + 2\vartheta)\sigma = -C_{v+v+} . \quad (18.31c)$$

The directed derivative  $\partial_v$  equals a derivative  $d/d\lambda$  with respect to an affine parameter along the geodesic directions, equation (18.25).

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

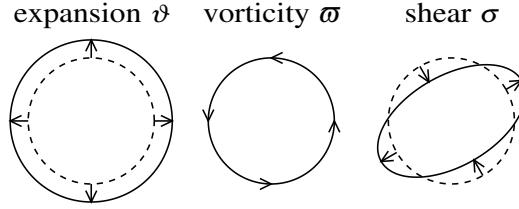


Figure 18.1 Illustrating how the Sachs optical coefficients, the expansion  $\vartheta$ , the vorticity  $\varpi$ , and the shear  $\sigma$ , characterize the rate at which a congruence of light rays changes shape as it propagates. The congruence of light rays is coming vertically upward out of the paper.

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.

## 18.6 Hypersurface-orthogonality for a timelike congruence

Singularity theorems consider special congruences that are both geodesic and vorticity-free. The Raychaudhuri equation (18.18b) guarantees that if the vorticity  $\varpi_{ab}$  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 (18.33), which proves to be a hypersurface of constant proper time, or equivalently of constant action. The next subsection, §18.6.1, 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

$$\mathbf{D} \wedge \mathbf{u} = \gamma^m \wedge \gamma^n (\partial_m u_n - \Gamma_{nm}^k u_k) = \gamma^m \wedge \gamma^n \Gamma_{n0m} = \gamma^0 \wedge \gamma^a K_a - \gamma^a \wedge \gamma^b \varpi_{ab}. \quad (18.32)$$

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  $\varpi_{ab}$ .

Equation (18.32) shows that the covariant curl  $\mathbf{D} \wedge \mathbf{u}$  vanishes if and only if both the acceleration  $K_a$  and the vorticity  $\varpi_{ab}$  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  $\tau$ ,

$$\mathbf{D} \wedge \mathbf{u} = 0 \Leftrightarrow \mathbf{u} = -\partial\tau. \quad (18.33)$$

The scalar  $\tau$  is just the proper time along the geodesics, as follows from

$$\mathbf{u} \cdot \mathbf{u} = -\mathbf{u} \cdot \partial\tau = -\frac{dx^\mu}{d\tau} \frac{\partial\tau}{\partial x^\mu} = -1. \quad (18.34)$$

Thus the 4-velocity  $\mathbf{u}$  is normal to 3-dimensional hypersurfaces of constant proper time  $\tau$ .

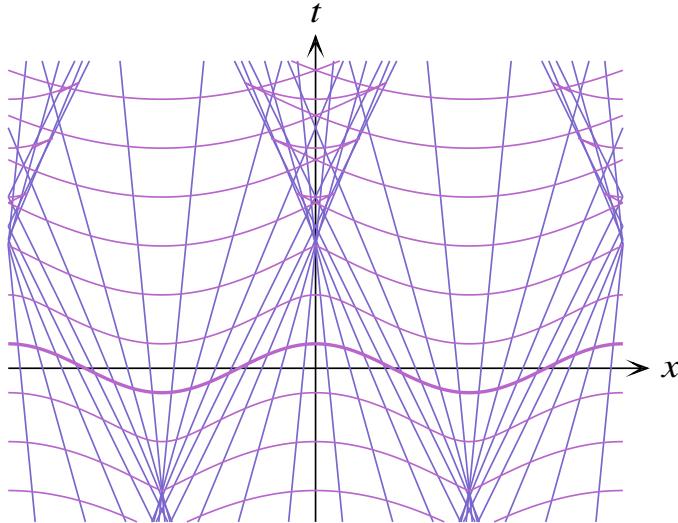


Figure 18.2 Spacetime diagram of Minkowski space illustrating a hypersurface-orthogonal congruence of timelike geodesics. The congruence is constructed by starting with an initial 3-dimensional spacelike hypersurface (thick like), here a cosine perturbation from the  $t = 0$  hypersurface, and projecting geodesics (blue lines) along its timelike normal direction. Hypersurfaces of constant proper time  $\tau$  (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 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),

$$S = -m\tau . \quad (18.35)$$

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),

$$p_{\mu} = \frac{\partial S}{\partial x^{\mu}} , \quad (18.36)$$

which reproduces the result  $u = -\partial\tau$ .

A weaker condition than the vanishing of  $D \wedge u$  is that the curl  $D \wedge (u/\alpha)$  of the 4-velocity scaled by some arbitrary factor  $\alpha$  vanishes. The covariant curl of the scaled 4-velocity  $u/\alpha$  is

$$\alpha D \wedge (u/\alpha) = D \wedge u + u \wedge \partial \ln \alpha = \gamma^0 \wedge \gamma^a (K_a - \partial_a \ln \alpha) - \gamma^a \wedge \gamma^b \varpi_{ab} . \quad (18.37)$$

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  $\varpi_{ab}$  vanishes,

$$K_a = \partial_a \ln \alpha , \quad \varpi_{ab} = 0 . \quad (18.38)$$



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.

The conditions (18.38) are precisely those established in the ADM formalism, with  $\alpha$  being the lapse. If conditions (18.38) hold, then  $D \wedge (\mathbf{u}/\alpha)$  vanishes, and if torsion also vanishes, then by Poincaré's lemma  $\mathbf{u}/\alpha$  is, at least locally, the gradient of a scalar  $t$ ,

$$D \wedge (\mathbf{u}/\alpha) = 0 \quad \Leftrightarrow \quad \mathbf{u} = -\alpha \partial t . \quad (18.39)$$

The scalar coordinate  $t$  is just the ADM time coordinate, as follows from

$$\mathbf{u} = \boldsymbol{\gamma}_0 = -\boldsymbol{\gamma}^0 = -e^0_{\mu} \mathbf{e}^{\mu} = -\alpha \mathbf{e}^t = -\alpha \mathbf{e}^{\mu} \frac{\partial t}{\partial x^{\mu}} = -\alpha \partial t . \quad (18.40)$$

### 18.6.1 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  $\tau$ , 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  $\tau$  even after they cross, but the proper time  $\tau$  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.

## 18.7 Hypersurface-orthogonality for a null congruence

For massive particles, the proper time  $\tau$ , or equivalently the action  $S = -m\tau = -m^2\lambda$ , where  $\lambda$  is the affine parameter, progresses along geodesics, and momenta along geodesics are orthogonal to hypersurfaces of constant action, equation (18.36). 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 (18.36) holds across all of the 4-dimensional spacetime. Rather, hypersurface-orthogonality for null congruences imposes that equation (18.36) holds in the massless limit along each 3-dimensional null hypersurface of constant action,

$$p_{\mu} = \lim_{m \rightarrow 0} m^2 \frac{\partial \lambda}{\partial x^{\mu}} . \quad (18.41)$$

To see why the definition of hypersurface-orthogonality for null congruences does not impose that the condition (18.36) hold over the entire 4-dimensional spacetime, suppose contrarily that it did. The 4-momentum along an outgoing null geodesic of the congruence satisfies  $p = p^v \gamma_v = p_u \gamma^u$  (no sum over  $v$  or  $u$ ). Poincaré’s lemma implies that equation (18.36) 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 (18.32),

$$\mathbf{D} \wedge \mathbf{p} = \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n (\partial_m p_n - \Gamma_{nm}^k p_k) = -p^v \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n \Gamma_{vm} , \quad (18.42)$$

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  $\Gamma_{uvv}$  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 (18.29) have no dependence on the 2 components  $\Gamma_{v[au]}$ , and the vorticity equation (18.30b) allows vorticity to vanish without requiring that  $\Gamma_{uvv}$  vanishes.

Thus hypersurface-orthogonality for null congruences is conventionally defined by the weaker condition that the limiting equation (18.41) 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

$$\mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) = -(p^v)^2 \boldsymbol{\gamma}^u \wedge (\boldsymbol{\gamma}^v \wedge \boldsymbol{\gamma}^a K_{avv} - \boldsymbol{\gamma}^a \wedge \boldsymbol{\gamma}^b \varpi_{ab}) . \quad (18.43)$$

The covariant curl (18.43) 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 (18.37), the hypersurface-orthogonality condition (18.43) for null congruences is unchanged by scaling the momentum  $\mathbf{p}$  by some arbitrary factor  $\alpha$ , since

$$\alpha \mathbf{p} \wedge (\mathbf{D} \wedge (\mathbf{p}/\alpha)) = \mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) + \mathbf{p} \wedge \mathbf{p} \wedge \partial \ln \alpha = \mathbf{p} \wedge (\mathbf{D} \wedge \mathbf{p}) , \quad (18.44)$$

because  $\mathbf{p} \wedge \mathbf{p} = 0$ .

The Raychaudhuri equation (18.30b) for the vorticity  $\varpi$  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 parametrized, that is, without imposing that  $\Gamma_{uvv}$  vanishes.

Hypersurface-orthogonality along the outgoing null congruence imposes only 3 conditions on the tetrad, 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.

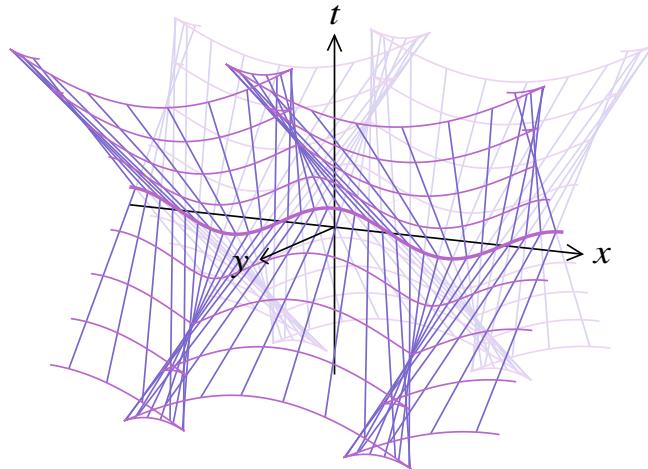


Figure 18.4 3D spacetime diagram of Minkowski space illustrating a pair of hypersurface-orthogonal congruences of null geodesics (blue lines) emerging from a 2-dimensional spacelike surface (thick line). The spacelike curves (purple lines) on the two null hypersurfaces are lines of constant affine parameter  $\lambda$ . 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  $t$ - $x$  plane.

### 18.7.1 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  $\gamma_v$  and  $\gamma_u$  of a Newman-Penrose tetrad. The spatial directions orthogonal to the two null directions define a plane whose tangent directions form the spatial directions  $\gamma_+$  and  $\gamma_-$  of the Newman-Penrose tetrad.

Again it should be emphasized that hypersurface-orthogonality for null congruences is defined not by condition (18.36) imposed over all spacetime, but rather by the limiting condition (18.41) imposed over each of the 3-dimensional null hypersurfaces of the congruence.

## 18.8 Focusing theorems

Focusing theorems exist for both timelike and null congruences. The focusing theorem follows from the Raychaudhuri equation for the expansion  $\vartheta$ , 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 §§18.6 and 18.7, 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.

### 18.8.1 Focusing theorem for a null geodesic congruence

If the vorticity  $\varpi$  vanishes, then in the frame parallel-transported along the null congruence, the Raychaudhuri equation (18.31a) for the expansion  $\vartheta$  along a null congruence simplifies to

$$\frac{d\vartheta}{d\lambda} + \vartheta^2 + \sigma\sigma^* + \frac{1}{2}G_{vv} = 0 . \quad (18.45)$$

The terms  $\vartheta^2$  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

$$G_{vv} = \frac{1}{2}(G_{00} + G_{33}) + \frac{1}{2}G_{33} . \quad (18.46)$$

The Einstein component  $G_{vv}$  has boost weight 2, and is therefore multiplied by  $e^{2\theta}$  under a boost by rapidity  $\theta$  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 (18.46) reduces to

$$G_{vv} = \frac{1}{2}(G_{00} + G_{33}) = 4\pi(\rho + p_3) , \quad (18.47)$$

where  $\rho$  is the energy density and  $p_3$  the pressure along the 3-direction. The Einstein component  $G_{vv}$  is therefore positive provided that

$$\boxed{\rho + p_3 \geq 0} , \quad (18.48)$$

which is called the **null energy condition**. If the null energy condition (18.48) holds, then the vorticity-free Raychaudhuri equation (18.45) shows that the expansion  $\vartheta$  must always decrease.

The Raychaudhuri equation (18.45) can be arranged as

$$\partial_v(1/\vartheta) = 1 + \frac{\sigma\sigma^* + \frac{1}{2}G_{vv}}{\vartheta^2}, \quad (18.49)$$

whose right hand side is greater than or equal to 1, given the null energy condition (18.48). If the expansion  $\vartheta$  is negative (meaning that light rays are converging), then equation (18.49) shows that  $1/\vartheta$  will reach 0 at a finite value of the affine parameter  $\lambda$ . In other words,  $\vartheta$  must become negative infinite at some finite value of  $\lambda$ .

A negative infinite value of the expansion means that the cross-sectional area of the null congruence has shrunk to zero. This 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.

### 18.8.2 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 (18.18a) along a timelike geodesic congruence is

$$\frac{d\vartheta}{d\tau} + \vartheta^2 + \frac{1}{3}\sigma^{ab}\sigma_{ab} + \frac{1}{3}R_{00} = 0 \quad (18.50)$$

in the orthonormal tetrad frame freely-falling along the geodesic. The component  $R_{00}$  of the Ricci tensor in the orthonormal tetrad is

$$R_{00} = 4\pi(\rho + 3p), \quad (18.51)$$

where  $\rho$  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

$$\boxed{\rho + 3p \geq 0}, \quad (18.52)$$

which is called the **strong energy condition**. Note that a cosmological constant violates the strong energy condition (18.52), but not the null energy condition (18.48).

## 18.9 Singularity theorems

This section gives an account of one version of the singularity theorems, the original null version proved by Penrose (1965). See Senovilla (1997) for a review of singularity theorems.

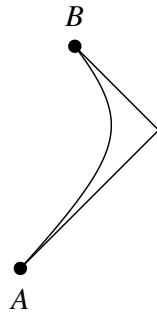


Figure 18.5 Spacetime diagram illustrating the dog-leg proposition. The dog-leg proposition asserts that any dog-leg path that joins 2 events  $A$  and  $B$  by a consecutive pair of null or timelike geodesics can be deformed into a strictly timelike path of longer proper time between  $A$  and  $B$ . The proposition is an assertion about the global causal structure of spacetime.

### 18.9.1 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 18.2.

The dog-leg proposition can be replaced by other plausible hypotheses. Much of the content of the book by Hawking and Ellis (1973) is concerned with exploring different plausible causality conditions. However, that will not be done here.

### 18.9.2 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 then 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

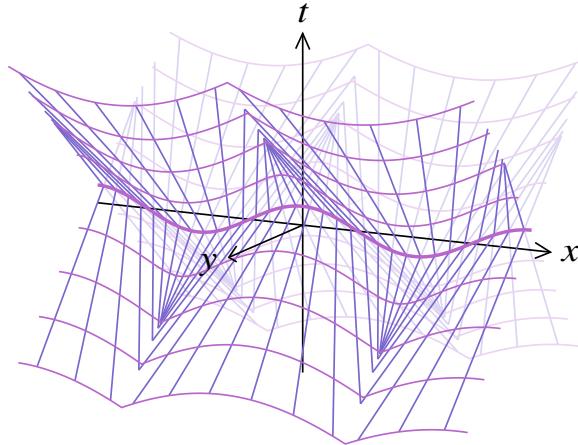


Figure 18.6 Null boundary of the future of a 2-dimensional spacelike surface. The null boundary is a pair of 3-dimensional null surfaces projecting orthogonally from the 2-surface (thick line), with the parts of the hypersurfaces excised after geodesic crossing, since the latter parts are connected by 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.

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  $\lambda$ , 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 18.2 How do singularity theorems apply to the Kerr geometry? Answer.** The Kerr geometry violates the deg-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 §23.22.1). As found in Exercise 23.1, 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 23.1.

**Concept question 18.3 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 23.1. 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](#)).

# 19

---

## 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 an inverse vierbein  $e^m{}_\mu$ , and consequently a vierbein  $e_m{}^\mu$ , and associated tetrad  $\gamma_m$ . 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  $\gamma_{mn}$  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  $u^\mu$  in the usual way, by applying the 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 vierbein,  $dx^\mu/d\tau = e_0{}^\mu$ . The coordinate 4-velocity  $u^\mu$  is related to the lapse  $\alpha$  and shift  $\beta^\alpha$  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  $\omega$ . Show that a clock attached to the wheel ticks slow by the Lorentz factor  $\gamma$  compared to a clock in the non-rotating frame, and that rulers

attached to the wheel measure the rim to be Lorentz-contracted by a factor  $\gamma$  compared to the non-rotating 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 inverse vierbein for the line-element (19.4) is  $e^m_\mu = \text{diag}(1, 1, r, 1)$ , and the corresponding vierbein is  $e_m^\mu = \text{diag}(1, 1, 1/r, 1)$ . Lorentz boost the vierbein into the tetrad frame of the wheel rotating at velocity  $v = r\omega$  in the azimuthal  $\phi$  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

$$ds^2 = -\gamma^2(dt - r^2\omega d\phi)^2 + dr^2 + r^2\gamma^2(d\phi - \omega dt)^2 + dz^2 . \quad (19.7)$$

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  $\gamma$  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

$$dl = r\gamma(d\phi - \omega dt) = r\gamma(1 - r^2\omega^2)d\phi = \frac{r d\phi}{\gamma} , \quad (19.9)$$

demonstrating that the rim is Lorentz-contracted by  $\gamma$  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 Alvar 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

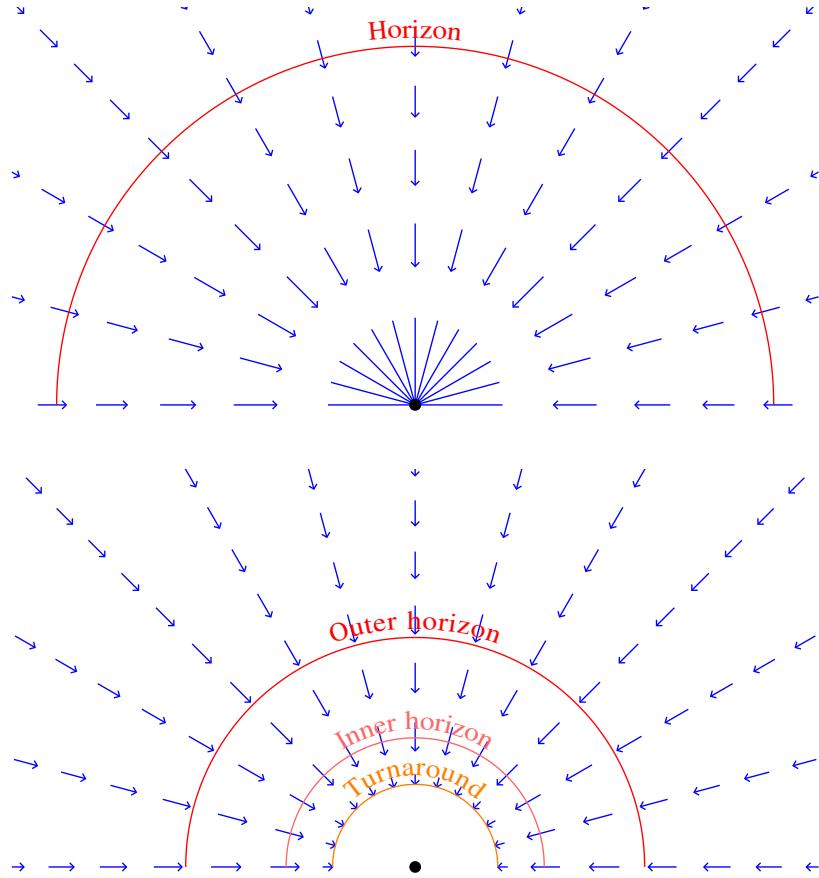


Figure 19.1 Radial velocity  $\beta$  in (upper panel) a Schwarzschild black hole, and (lower panel) a Reissner-Nordström black hole with electric charge  $Q = 0.96$ .

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  $\beta$  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  $t_{\text{ff}}$  is the free-fall time, the proper time experienced by a person who free-falls from rest at infinity. The radial velocity  $\beta$  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  $\beta$  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  $\beta$  given by equation (19.12) 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 an inverse 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 inverse vierbein  $e^m_\mu$  of the Gullstrand-Painlevé line-element (19.10), and the corresponding 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  $\beta$  represents a radial velocity, while  $t_{\text{ff}}$  coincides with the proper time in the tetrad frame.

The tetrad and coordinate axes  $\gamma_m$  and  $e_\mu$  are related to each other by the vierbein and inverse 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  $\gamma_m$  are thus related to the coordinate axes  $e_\mu$  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  $\gamma_a$  oriented in the polar directions  $r, \theta, \phi$ ) 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  $\gamma_m$  are parallel-transported, without precessing, along the worldlines of the radially free-falling observers can be confirmed by checking that the tetrad connections  $\Gamma_{nm0}$  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  $\partial_0$  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  $\beta^{\alpha}$  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 inverse vierbein  $e^m_{\mu}$  and 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  $\gamma_m$  of the Gullstrand-Painlevé-Cartesian line-element (19.22) are related to the coordinate tangent axes  $e_{\mu}$  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  $e_{\mu}$  are related to the tetrad axes  $\gamma_m$  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  $\beta^a$  of the radial velocity coincide with the coordinate-frame contravariant components  $\beta^{\alpha}$ ; for clarification of this point see the more general equation (19.54) 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  $r, \theta, \phi$ . Like the polar tetrad, the Cartesian tetrad axes  $\gamma_m$  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  $\Gamma_{nm0}$  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  $\beta$  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 fisheries

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  $\Gamma_{mn}^k$  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  $\delta\xi^n = e^n_\nu \delta x^\nu = \delta_\nu^n (\delta x^\nu - \beta^\nu \delta t_{\text{ff}}) = \delta x^n - \beta^n \delta\tau$  equals the distance  $\delta x^n$  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  $\delta p^k$  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 non-matching 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 (19.57). With the connections (19.29), the change  $\delta p^k$  (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  $\delta x^\nu$  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 (ingoing or outgoing) null congruences.

The Boyer-Lindquist orthonormal tetrad is encoded in the Boyer-Lindquist metric

$$ds^2 = -\frac{R^2\Delta}{\rho^2} (dt - a \sin^2\theta d\phi)^2 + \frac{\rho^2}{R^2\Delta} dr^2 + \rho^2 d\theta^2 + \frac{R^4 \sin^2\theta}{\rho^2} \left( d\phi - \frac{a}{R^2} dt \right)^2 , \quad (19.34)$$

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

$$e_m^\mu = \frac{1}{\rho} \begin{pmatrix} R/\sqrt{\Delta} & 0 & 0 & a/(R\sqrt{\Delta}) \\ 0 & R\sqrt{\Delta} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ a \sin\theta & 0 & 0 & 1/\sin\theta \end{pmatrix} , \quad (19.36)$$

with inverse vierbein  $e^m_{\mu}$

$$e^m_{\mu} = \begin{pmatrix} R\sqrt{\Delta}/\rho & 0 & 0 & -a\sin^2\theta R\sqrt{\Delta}/\rho \\ 0 & \rho/(R\sqrt{\Delta}) & 0 & 0 \\ 0 & 0 & \rho & 0 \\ -a\sin\theta/\rho & 0 & 0 & R^2\sin\theta/\rho \end{pmatrix}. \quad (19.37)$$

With respect to the Boyer-Lindquist tetrad, only the time component  $A^t$  of the electromagnetic potential  $A^m$  is non-vanishing,

$$A^m = \left\{ \frac{Qr}{\rho R\sqrt{\Delta}}, 0, 0, 0 \right\}. \quad (19.38)$$

Only the radial components  $E$  and  $B$  of the electric and magnetic fields are non-vanishing, 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 (??) and (22.54) with zero electric charge and current,  $j^n = 0$ , except at the singularity  $\rho = 0$ .

The non-vanishing 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 non-vanishing 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  $u^\mu$  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.

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  $\omega_{km}$ , equation (19.58) 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 (19.67).

Thanks to the twist, unlike the Gullstrand-Painlevé metric, the Doran metric is not spatially flat at constant free-fall time  $t_{\text{ff}}$ . Rather, the spatial metric is sheared in the azimuthal direction. Just as the velocity produces a Lorentz boost that makes the metric non-flat with respect to the time components, so also the twist produces a rotation that makes the metric non-flat with respect to the spatial components.

### 19.4.1 Doran-Cartesian coordinates

In place of the polar coordinates  $\{r, \theta, \phi_{\text{ff}}\}$  of the Doran metric, 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  $a$ )

$$x \equiv R \sin \theta \cos \phi_{\text{ff}}, \quad y \equiv R \sin \theta \sin \phi_{\text{ff}}, \quad z \equiv r \cos \theta. \quad (19.45)$$

The metric in Doran-Cartesian coordinates  $x^\mu \equiv \{t_{\text{ff}}, x^\alpha\} \equiv \{t_{\text{ff}}, x, y, z\}$ , is

$$ds^2 = -dt_{\text{ff}}^2 + \delta_{\alpha\beta} (dx^\alpha - \beta^\alpha \alpha_\kappa dx^\kappa) (dx^\beta - \beta^\beta \alpha_\lambda dx^\lambda) \quad (19.46)$$

where  $\alpha_\mu$  is the rotational velocity vector

$$\alpha_\mu = \left\{ 1, \frac{ay}{R^2}, -\frac{ax}{R^2}, 0 \right\}, \quad (19.47)$$

and  $\beta^\mu$  is the velocity vector

$$\beta^\mu = \frac{\beta R}{\rho} \left\{ 0, \frac{xr}{R\rho}, \frac{yr}{R\rho}, \frac{zR}{r\rho} \right\}. \quad (19.48)$$

The rotational velocity and radial velocity vectors are orthogonal

$$\alpha_\mu \beta^\mu = 0 . \quad (19.49)$$

For the Kerr-Newman metric, the radial velocity  $\beta$  is

$$\beta = \mp \frac{\sqrt{2Mr - Q^2}}{R} \quad (19.50)$$

with  $-$  for black hole (infalling),  $+$  for white hole (outfalling) solutions. Horizons occur where

$$\beta = \mp 1 , \quad (19.51)$$

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  $\beta^2$ , but rather differs from  $\beta^2$  by a factor of  $R^2/\rho^2$ :

$$\beta_\mu \beta^\mu = \beta_m \beta^m = \frac{\beta^2 R^2}{\rho^2} . \quad (19.52)$$

The point of the convention adopted here is that  $\beta(r)$  is any and only a function of  $r$ , rather than depending also on  $\theta$  through  $\rho$ . Moreover, with the convention here,  $\beta$  is  $\mp 1$  at horizons, equation (19.51). Finally, the 4-velocity  $\beta^\mu$  is simply related to  $\beta$  by  $\beta^\mu = (\beta/r) \partial r / \partial x^\mu$ .

The Doran-Cartesian metric (19.46) encodes a vierbein  $e_m^\mu$  and inverse vierbein  $e^m_\mu$

$$e_m^\mu = \delta_m^\mu + \alpha_m \beta^\mu , \quad e^m_\mu = \delta_\mu^m - \alpha_\mu \beta^m . \quad (19.53)$$

Here the tetrad-frame components  $\alpha_m$  of the rotational velocity vector and  $\beta^m$  of the radial velocity vector are

$$\alpha_m = e_m^\mu \alpha_\mu = \delta_m^\mu \alpha_\mu , \quad \beta^m = e^m_\mu \beta^\mu = \delta_\mu^m \beta^\mu , \quad (19.54)$$

which works thanks to the orthogonality (19.49) of  $\alpha_\mu$  and  $\beta^\mu$ . Equation (19.54) 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

$$\begin{aligned} \gamma_\parallel &\equiv \cos \phi_{\text{ff}} \gamma_1 + \sin \phi_{\text{ff}} \gamma_2 , & \gamma_\perp &\equiv -\sin \phi_{\text{ff}} \gamma_1 + \cos \phi_{\text{ff}} \gamma_2 , \\ e_\parallel &\equiv \cos \phi_{\text{ff}} e_x + \sin \phi_{\text{ff}} e_y , & e_\perp &\equiv -\sin \phi_{\text{ff}} e_x + \cos \phi_{\text{ff}} e_y . \end{aligned} \quad (19.55)$$

Then the relation between Doran-Cartesian tetrad axes  $\gamma_m$  and the tangent axes  $e_{\mu}$  of the Doran-Cartesian metric (19.46) is

$$\gamma_0 = e_{t_{ff}} + \beta^\alpha e_\alpha , \quad (19.56a)$$

$$\gamma_{\parallel} = e_{\parallel} , \quad (19.56b)$$

$$\gamma_{\perp} = e_{\perp} - \frac{a \sin \theta}{R} \beta^\alpha e_\alpha , \quad (19.56c)$$

$$\gamma_3 = e_z . \quad (19.56d)$$

The relations (19.56) 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 fishes

The tetrad-frame connections equal the ordinary coordinate partial derivatives in Doran-Cartesian coordinates of a bivector (antisymmetric tensor)  $\omega_{km}$

$$\boxed{\Gamma_{kmn} = -\delta_n^{\nu} \frac{\partial \omega_{km}}{\partial x^\nu}} , \quad (19.57)$$

which I call the river field because it encapsulates all the properties of the infalling river of space. The bivector river field  $\omega_{km}$  is

$$\boxed{\omega_{km} = \alpha_k \beta_m - \alpha_m \beta_k - \varepsilon_{0km} \zeta^a} , \quad (19.58)$$

where  $\beta_m = \eta_{mn} \beta^m$ , the totally antisymmetric tensor  $\varepsilon_{klmn}$  is normalized so that  $\varepsilon_{0123} = -1$ , and the vector  $\zeta^a$  points vertically upward along the rotation axis of the black hole

$$\zeta^a \equiv \{0, 0, 0, \zeta\} , \quad \zeta \equiv a \int_{\infty}^r \frac{\beta dr}{R^2} . \quad (19.59)$$

The electric part of  $\omega_{km}$ , where one of the indices is time 0, constitutes the velocity vector  $\beta^a$

$$\omega_{0a} = \beta^a \quad (19.60)$$

while the magnetic part of  $\omega_{km}$ , where both indices are spatial, constitutes the twist vector  $\mu^a$  defined by

$$\mu^a \equiv \frac{1}{2} \varepsilon^{0akm} \omega_{km} = \varepsilon^{0akm} \alpha_k \beta_m + \zeta^a . \quad (19.61)$$

The sense of the twist is that induces a right-handed rotation about an axis equal to the direction of  $\mu^a$  by an angle equal to the magnitude of  $\mu^a$ . In 3-vector notation, with  $\boldsymbol{\mu} \equiv \mu^a$ ,  $\boldsymbol{\alpha} \equiv \alpha_a$ ,  $\boldsymbol{\beta} \equiv \beta^a$ ,  $\boldsymbol{\zeta} \equiv \zeta^a$ ,

$$\boldsymbol{\mu} \equiv \boldsymbol{\alpha} \times \boldsymbol{\beta} + \boldsymbol{\zeta} . \quad (19.62)$$

In terms of the velocity and twist vectors, the river field  $\omega_{km}$  is

$$\omega_{km} = \begin{pmatrix} 0 & \beta^1 & \beta^2 & \beta^3 \\ -\beta^1 & 0 & \mu^3 & -\mu^2 \\ -\beta^2 & -\mu^3 & 0 & \mu^1 \\ -\beta^3 & \mu^2 & -\mu^1 & 0 \end{pmatrix}. \quad (19.63)$$

Note that the sign of the electric part  $\beta$  of  $\omega_{km}$  is opposite to the sign of the analogous electric field  $E$  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  $\beta^a$ , and right-handed rotations about the twist  $\mu^a$ . Like a static electric field, the velocity vector  $\beta^a$  is the gradient of a potential

$$\beta^a = \delta_{\alpha}^a \frac{\partial}{\partial x^{\alpha}} \int^r \beta dr, \quad (19.64)$$

but unlike a magnetic field the twist vector  $\mu^a$  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 (19.57), 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

$$\frac{dp^k}{d\tau} = \delta_{\nu}^k \frac{\partial \omega^k_m}{\partial x^{\nu}} u^n p^m. \quad (19.65)$$

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

$$\delta\omega^k_m = \delta\xi^n \frac{\partial \omega^k_m}{\partial x^n}. \quad (19.66)$$

Consequently the 4-vector  $p^k$  is changed by

$$p^k \rightarrow p^k + \delta\omega^k_m p^m. \quad (19.67)$$

But equation (19.67) 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  $\delta\omega^k_m$ , equation (19.66), 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

$$\delta\omega^k_m = (\delta x^{\nu} - \delta x^{\nu}_{\text{river}}) \frac{\partial \omega^k_m}{\partial x^{\nu}} = [\delta x^{\nu} - \beta^{\nu}(\delta t_{\text{ff}} - a \sin^2\theta \delta\phi_{\text{ff}})] \frac{\partial \omega^k_m}{\partial x^{\nu}}, \quad (19.68)$$

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 (19.57), 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 I suspect that that is not true.

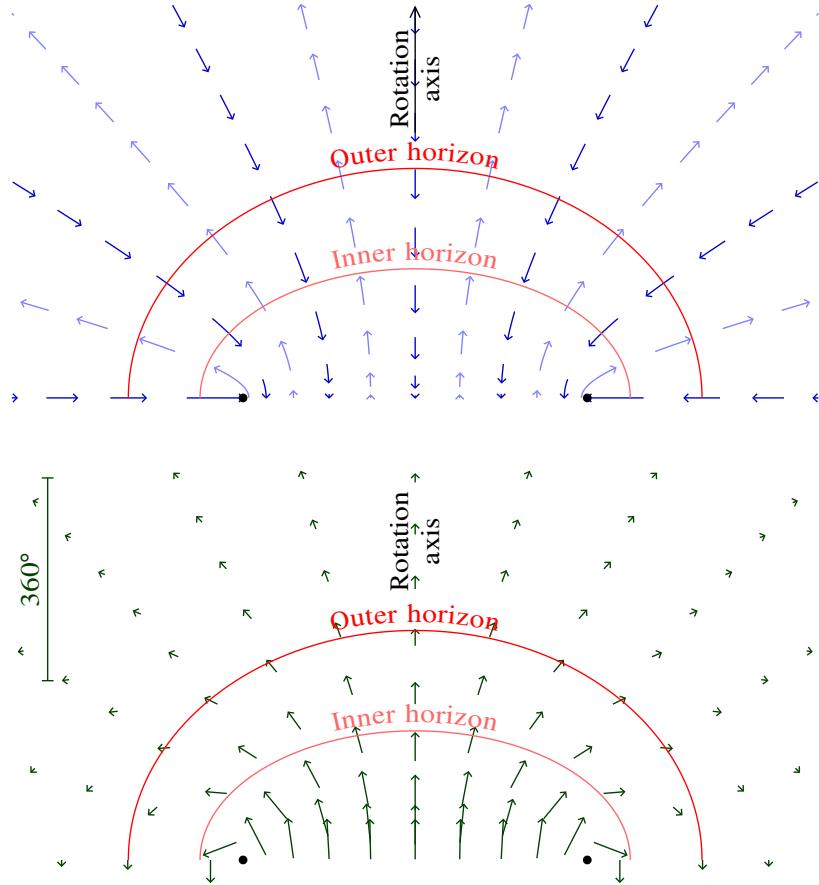


Figure 19.2 (Upper panel) velocity  $\beta^a$  and (lower panel) twist  $\mu^a$  vector fields for a Kerr black hole with spin parameter  $a = 0.96$ . Both vectors lie, as shown, in the plane of constant free-fall azimuthal angle  $\phi_{\text{ff}}$ . The vertical bar in the lower panel shows the length of a twist vector corresponding to a full rotation of  $360^\circ$ .

---

**Exercise 19.5 River model of the Friedmann-Lemaître-Robertson-Walker metric.** Show that the flat FLRW line-element

$$ds^2 = -dt^2 + a^2(dx^2 + x^2 do^2) \quad (19.69)$$

can be re-expressed as

$$ds^2 = -dt^2 + (dr - Hr dt)^2 + r^2 do^2 , \quad (19.70)$$

where  $r \equiv ax$  is the proper radial distance, and  $H \equiv \dot{a}/a$  is the Hubble parameter. Interpret the line-element (19.70). Is there a generalization to a non-flat FLRW universe?

**Exercise 19.6 Program geodesics in a rotating black hole.** Write a graphics (Java?) program that uses the prescription (19.66) to draw geodesics of test particles in an ideal (Kerr-Newman) black hole, expressed in Doran-Cartesian coordinates. Attach 3D bodies to your test particles, and use the same prescription (19.66) to rotate the bodies. Implement an option to translate to Boyer-Lindquist coordinates.

---

# 20

---

## General spherically symmetric spacetimes

### 20.1 Spherical spacetime

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 (20.37). 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  $\Phi$ . 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.

### 20.2 Spherical line-element

The spherical line-element adopted in this chapter is, in spherical polar coordinates  $x^\mu \equiv \{t, r, \theta, \phi\}$ ,

$$ds^2 = -\frac{dt^2}{\alpha^2} + \frac{1}{\beta_1^2} \left( dr - \beta_0 \frac{dt}{\alpha} \right)^2 + r^2 d\theta^2. \quad (20.1)$$

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  $1/\alpha$  and shift  $\beta_0/\alpha$ . The notation  $\beta_m$  is motivated by the fact that  $\{\beta_0, \beta_1, 0, 0\}$  forms a tetrad-frame 4-vector, equation (20.9). 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 inverse vierbein  $e^m_{\mu}$  can be read off from the line-element (20.1):

$$e^0_{\mu} dx^{\mu} = \frac{dt}{\alpha} , \quad (20.2a)$$

$$e^1_{\mu} dx^{\mu} = \frac{1}{\beta_1} \left( dr - \beta_0 \frac{dt}{\alpha} \right) , \quad (20.2b)$$

$$e^2_{\mu} dx^{\mu} = r d\theta , \quad (20.2c)$$

$$e^3_{\mu} dx^{\mu} = r \sin \theta d\theta . \quad (20.2d)$$

The vierbein  $e_m^{\mu}$  and inverse vierbein  $e^m_{\mu}$  corresponding to the spherical line-element (20.1) are

$$e_m^{\mu} = \begin{pmatrix} \alpha & \beta_0 & 0 & 0 \\ 0 & \beta_1 & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1/(r \sin \theta) \end{pmatrix} , \quad e^m_{\mu} = \begin{pmatrix} 1/\alpha & 0 & 0 & 0 \\ -\beta_0/(\alpha \beta_1) & 1/\beta_1 & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & 0 & 0 & r \sin \theta \end{pmatrix} . \quad (20.3)$$

As in the ADM formalism, §17.1, the tetrad time axis  $\gamma_0$  is chosen to be orthogonal to hypersurfaces of constant time  $t$ . However, the convention here for the vierbein coefficients differs from the ADM convention: here  $1/\alpha$  is the ADM lapse, while  $\beta_0/\alpha$  is the ADM shift. The directed derivatives  $\partial_0$  and  $\partial_1$  along the time and radial tetrad axes  $\gamma_0$  and  $\gamma_1$  are

$$\partial_0 = e_0^{\mu} \frac{\partial}{\partial x^{\mu}} = \alpha \frac{\partial}{\partial t} + \beta_0 \frac{\partial}{\partial r} , \quad \partial_1 = e_1^{\mu} \frac{\partial}{\partial x^{\mu}} = \beta_1 \frac{\partial}{\partial r} . \quad (20.4)$$

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  $u^{\mu}$  of such a person is

$$u^{\mu} = e_m^{\mu} u^m = e_0^{\mu} = \{\alpha, \beta_0, 0, 0\} . \quad (20.5)$$

A person instantaneously at rest in the tetrad frame satisfies  $dr/dt = \beta_0/\alpha$  according to equation (20.5), so it follows from the line-element (20.1) that the proper time  $\tau$  of a person at rest in the tetrad frame is related to the coordinate time  $t$  by

$$d\tau = \frac{dt}{\alpha} \quad \text{in tetrad rest frame} . \quad (20.6)$$

The directed time derivative  $\partial_0$  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  $\partial_m$  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  $\beta_0$  and  $\beta_1$  introduced above. Since  $\beta_m$  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

$$\boxed{1 - \frac{2M}{r} \equiv \beta_m \beta^m = -\beta_0^2 + \beta_1^2} \quad \text{a coordinate and tetrad scalar .} \quad (20.10)$$

The interpretation of  $M$  as the interior mass will become evident below, §20.9.

**Exercise 20.1 Apparent horizon.** Show that radial null geodesics in a spherical geometry satisfy

$$\frac{dr}{dt} = \frac{\beta_0 \pm \beta_1}{\alpha} . \quad (20.11)$$

An apparent horizon occurs where outgoing radial null geodesics are not moving radially,  $dr/dt = 0$ . Conclude that an apparent horizon occurs where (choosing  $\alpha$  and  $\beta_1$  positive without loss of generality)

$$\beta_0 = -\beta_1 , \quad (20.12)$$

or equivalently where

$$r = 2M . \quad (20.13)$$

### 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

$$g_{tt} dt^2 + 2 g_{tr} dt dr + g_{rr} dr^2 = \frac{1}{g_{tt}} [(g_{tt} dt + g_{tr} dr)^2 + (g_{tt} g_{rr} - g_{tr}^2) dr^2] . \quad (20.14)$$

This can be diagonalized by choosing the time coordinate  $t_\times$  such that

$$f dt_\times = g_{tt} dt + g_{tr} dr \quad (20.15)$$

for some integrating factor  $f(t, r)$ . Equation (20.15) can be solved by choosing  $t_\times$  to be constant along integral curves

$$\frac{dr}{dt} = -\frac{g_{tt}}{g_{tr}} . \quad (20.16)$$

The resulting diagonal **rest line-element** is

$$ds^2 = -\frac{dt_{\times}^2}{\alpha_{\times}^2} + \frac{dr^2}{1 - 2M/r} + r^2 do^2 . \quad (20.17)$$

The line-element (20.17) 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  $dr/dt = \beta_0/\alpha$  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 (20.15), the tetrad transformation being precisely that needed to make the line-element (20.17) diagonal. The metric coefficient  $g_{rr}$  in the line-element (20.17) follows from the fact that  $\beta_1^2 = 1 - 2M/r$  when  $\beta_0 = 0$ , equation (20.10). 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 (20.15),  $r_{\times}$  is chosen to satisfy

$$f dr_{\times} = g_{tr} dt + g_{rr} dr \equiv \frac{1}{\beta_1} \left( dr - \beta_0 \frac{dt}{\alpha} \right) \quad (20.18)$$

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  $dr/dt = \beta_0/\alpha$ , 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

$$ds^2 = -\frac{dt^2}{\alpha^2} + \frac{dr_{\times}^2}{\lambda^2} + r^2 do^2 , \quad (20.19)$$

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 (20.17) the tetrad was changed from one that was moving at  $dr/dt = \beta_0/\alpha$  to one that was at rest, here the transformation keeps the tetrad unchanged. In both the rest and comoving diagonal line-elements (20.17) and (20.19) the tetrad is at rest relative to the respective radial coordinate  $r$  or  $r_{\times}$ ; but whereas in the rest line-element (20.17) the radial coordinate was fixed to be the circumferential radius  $r$ , in the comoving line-element (20.19) 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  $\partial_0$  and  $\partial_1$  are unchanged:

$$\partial_0 = \alpha \frac{\partial}{\partial t} \Big|_{r_{\times}} = \alpha \frac{\partial}{\partial t} \Big|_r + \beta_0 \frac{\partial}{\partial r} \Big|_t , \quad \partial_1 = \lambda \frac{\partial}{\partial r_{\times}} \Big|_t = \beta_1 \frac{\partial}{\partial r} \Big|_t . \quad (20.20)$$

## 20.5 Tetrad connections

Now turn the handle to proceed towards the Einstein equations. The non-vanishing tetrad connections coefficients  $\Gamma_{kmn}$  corresponding to the spherical line-element (20.1) are

$$\Gamma_{100} = h_0 , \quad (20.21a)$$

$$\Gamma_{101} = h_1 , \quad (20.21b)$$

$$\Gamma_{202} = \Gamma_{303} = \frac{\beta_0}{r} , \quad (20.21c)$$

$$\Gamma_{212} = \Gamma_{313} = \frac{\beta_1}{r} , \quad (20.21d)$$

$$\Gamma_{323} = \frac{\cot\theta}{r} , \quad (20.21e)$$

where  $h_0$  is the proper radial acceleration (minus the gravitational force) experienced by a person at rest in the tetrad frame

$$h_0 \equiv -\partial_1 \ln \alpha = -\beta_1 \frac{\partial \ln \alpha}{\partial r} , \quad (20.22)$$

and  $h_1$  is the “Hubble parameter” of the radial flow, as measured in the tetrad rest frame, defined by

$$h_1 \equiv -\beta_0 \frac{\partial \ln \alpha}{\partial r} + \frac{\partial \beta_0}{\partial r} - \partial_0 \ln \beta_1 . \quad (20.23)$$

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

$$\frac{Du^m}{D\tau} = D_0 u^m = \partial_0 u^m + \Gamma_{00}^m u^0 = \Gamma_{00}^m = \{0, \Gamma_{00}^1, 0, 0\} = \{0, h_0, 0, 0\} , \quad (20.24)$$

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

$$D_1 u^m = \partial_1 u^m + \Gamma_{01}^m u^0 = \Gamma_{01}^m = \{0, \Gamma_{01}^1, 0, 0\} = \{0, h_1, 0, 0\} . \quad (20.25)$$

Confined to the  $\gamma_0$ - $\gamma_1$ -plane (that is, considering only Lorentz transformations in the  $t$ - $r$ -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  $h_n = \{h_0, h_1\}$ :

$$h_n = \Gamma_{10n} . \quad (20.26)$$

The Riemann tensor, equations (20.28) 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.

Since  $h_1$  is a kind of radial Hubble parameter, it can be useful to define a corresponding radial scale factor  $\lambda$  by

$$h_1 \equiv -\partial_0 \ln \lambda . \quad (20.27)$$

The scale factor  $\lambda$  is the same as the  $\lambda$  in the comoving line-element of equation (20.19). This is true because  $h_1$  is a tetrad connection and therefore coordinate gauge-invariant, and the line-element (20.19) 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 non-vanishing components of the tetrad-frame Riemann tensor  $R_{klmn}$  corresponding to the spherical line-element (20.1) are

$$R_{0101} = D_1 h_0 - D_0 h_1 , \quad (20.28a)$$

$$R_{0202} = R_{0303} = -\frac{1}{r} D_0 \beta_0 , \quad (20.28b)$$

$$R_{1212} = R_{1313} = -\frac{1}{r} D_1 \beta_1 , \quad (20.28c)$$

$$R_{0212} = R_{0313} = -\frac{1}{r} D_0 \beta_1 = -\frac{1}{r} D_1 \beta_0 , \quad (20.28d)$$

$$R_{2323} = \frac{2M}{r^3} , \quad (20.28e)$$

where  $D_m$  denotes the covariant derivative as usual. The non-vanishing components of the tetrad-frame Ricci tensor  $R_{km}$  are

$$R_{00} = R_{0101} + 2R_{0202} , \quad (20.29a)$$

$$R_{11} = -R_{0101} + 2R_{1212} , \quad (20.29b)$$

$$R_{01} = 2R_{0212} , \quad (20.29c)$$

$$R_{22} = R_{33} = -R_{0202} + R_{1212} + R_{2323} , \quad (20.29d)$$

whence

$$R_{00} = D_1 h_0 - D_0 h_1 - \frac{2}{r} D_0 \beta_0 , \quad (20.30a)$$

$$R_{11} = -D_1 h_0 + D_0 h_1 - \frac{2}{r} D_1 \beta_1 , \quad (20.30b)$$

$$R_{01} = -\frac{2}{r} D_0 \beta_1 = -\frac{2}{r} D_1 \beta_0 , \quad (20.30c)$$

$$R_{22} = R_{33} = \frac{1}{r} D_0 \beta_0 - \frac{1}{r} D_1 \beta_1 + \frac{2M}{r^3} . \quad (20.30d)$$

The Ricci scalar is

$$R = -2D_1 h_0 + 2D_0 h_1 + \frac{4}{r} D_0 \beta_0 - \frac{4}{r} D_1 \beta_1 + \frac{4M}{r^3} . \quad (20.31)$$

The non-vanishing components of the tetrad-frame Einstein tensor  $G^{km}$  are

$$G^{00} = 2R_{1212} + R_{2323} , \quad (20.32a)$$

$$G^{11} = 2R_{0202} - R_{2323} , \quad (20.32b)$$

$$G^{01} = -2R_{0212} , \quad (20.32c)$$

$$G^{22} = G^{33} = R_{0101} + R_{0202} - R_{1212} , \quad (20.32d)$$

whence

$$G^{00} = \frac{2}{r} \left( -D_1 \beta_1 + \frac{M}{r^2} \right) , \quad (20.33a)$$

$$G^{11} = \frac{2}{r} \left( -D_0 \beta_0 - \frac{M}{r^2} \right) , \quad (20.33b)$$

$$G^{01} = \frac{2}{r} D_0 \beta_1 = \frac{2}{r} D_1 \beta_0 , \quad (20.33c)$$

$$G^{22} = G^{33} = D_1 h_0 - D_0 h_1 + \frac{1}{r} (D_1 \beta_1 - D_0 \beta_0) . \quad (20.33d)$$

The non-vanishing components of the tetrad-frame Weyl tensor  $C_{klmn}$  are

$$\frac{1}{2} C_{0101} = -C_{0202} = -C_{0303} = C_{1212} = C_{1313} = -\frac{1}{2} C_{2323} = C , \quad (20.34)$$

where  $C$  is the Weyl scalar (the spin-0 component of the Weyl tensor),

$$C \equiv \frac{1}{6} (R_{0101} - R_{0202} + R_{1212} - R_{2323}) = \frac{1}{6} (G^{00} - G^{11} + G^{22}) - \frac{M}{r^3} . \quad (20.35)$$

## 20.7 Einstein equations

The tetrad-frame Einstein equations

$$G^{km} = 8\pi T^{km} \quad (20.36)$$

imply that

$$\begin{pmatrix} G^{00} & G^{01} & 0 & 0 \\ G^{01} & G^{11} & 0 & 0 \\ 0 & 0 & G^{22} & 0 \\ 0 & 0 & 0 & G^{33} \end{pmatrix} = 8\pi T^{km} = 8\pi \begin{pmatrix} \rho & f & 0 & 0 \\ f & p & 0 & 0 \\ 0 & 0 & p_{\perp} & 0 \\ 0 & 0 & 0 & p_{\perp} \end{pmatrix} \quad (20.37)$$

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,

$$\beta_0 = 0 , \quad (20.38)$$

as in the Schwarzschild or Reissner-Nordström line-elements. Alternatively, the tetrad frame could be chosen to be in free-fall,

$$h_0 = 0 , \quad (20.39)$$

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

$$G^{01} = 8\pi f = 0 . \quad (20.40)$$

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 (20.33b) with the middle expression of (20.33c), and (20.33a) with the final expression of (20.33c), respectively, along with the definition (20.10) of the interior mass  $M$ , and the Einstein equations (20.37), imply (note that  $D_m M = \partial_m M$  since  $M$  is a scalar)

$$p = \frac{1}{\beta_0} \left( -\frac{1}{4\pi r^2} \partial_0 M - \beta_1 f \right) , \quad (20.41a)$$

$$\rho = \frac{1}{\beta_1} \left( \frac{1}{4\pi r^2} \partial_1 M - \beta_0 f \right) . \quad (20.41b)$$

In the centre-of-mass frame,  $f = 0$ , these equations reduce to

$$\partial_0 M = -4\pi r^2 \beta_0 p , \quad (20.42a)$$

$$\partial_1 M = 4\pi r^2 \beta_1 \rho . \quad (20.42b)$$

Equations (20.42) amply justify the interpretation of  $M$  as the interior mass. The first equation (20.42a) can be written

$$\partial_0 M + p 4\pi r^2 \partial_0 r = 0 , \quad (20.43)$$

which can be recognized as an expression of the first law of thermodynamics,

$$dE + p dV = 0 , \quad (20.44)$$

with mass-energy  $E$  equal to  $M$ . The second equation (20.42b) can be written, since  $\partial_1 = \beta_1 \partial/\partial r$ , equation (20.4),

$$\frac{\partial M}{\partial r} = 4\pi r^2 \rho , \quad (20.45)$$

which looks exactly like the Newtonian relation between interior mass  $M$  and density  $\rho$ . Actually, this apparently Newtonian equation (20.45) is deceiving. The proper 3-volume element  $d^3r$  in the centre-of-mass tetrad frame is given by, equation (15.84),

$$d^3r = (1/e) d^3x^{r\theta\phi} = \frac{r^2 \sin \theta dr d\theta d\phi}{\beta_1} , \quad (20.46)$$

where  $e$ , equation (15.85), is the determinant of the  $3 \times 4$  vierbein submatrix  $e_a^{\mu}$  with tetrad indices  $a$  running over the 3 spatial indices 1, 2, 3, and coordinate indices  $\mu$  running over all 4 coordinates (the time column  $e_m^t$  of the vierbein is identically zero, equation (20.3), so the determinant of the  $3 \times 4$  submatrix  $e_a^{\mu}$  here coincides with the determinant of the  $3 \times 3$  submatrix  $e_a^{\alpha}$  with only spatial tetrad and coordinate indices  $a$  and  $\alpha$ ). Thus the proper 3-volume element  $dV \equiv d^3r$  of a radial shell of width  $dr$  is

$$dV = \frac{4\pi r^2 dr}{\beta_1} . \quad (20.47)$$

Consequently the “true” mass-energy  $dM_m$  associated with the proper density  $\rho$  in a proper radial volume element  $dV$  might be expected to be

$$dM_m = \rho dV = \frac{4\pi r^2 dr}{\beta_1} , \quad (20.48)$$

whereas equation (20.45) indicates that the actual mass-energy is

$$dM = \rho 4\pi r^2 dr = \beta_1 \rho dV . \quad (20.49)$$

A person in the centre-of-mass frame might perhaps, although there is really no formal justification for doing so, interpret the balance of the mass-energy as gravitational mass-energy  $M_g$

$$dM_g = (\beta_1 - 1) \rho dV . \quad (20.50)$$

Whatever the case, the moral of this is that you should beware of interpreting the interior mass  $M$  too literally as palpable mass-energy.

## 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 two non-vanishing equations represent conservation of energy and of radial momentum, and are

$$D_m T^{m0} = \partial_0 \rho + \frac{2\beta_0}{r} (\rho + p_\perp) + h_1 (\rho + p) + \left( \partial_1 + \frac{2\beta_1}{r} + 2h_0 \right) f = 0 , \quad (20.51a)$$

$$D_m T^{m1} = \partial_1 p + \frac{2\beta_1}{r} (p - p_\perp) + h_0 (\rho + p) + \left( \partial_0 + \frac{2\beta_0}{r} + 2h_1 \right) f = 0 . \quad (20.51b)$$

In the centre-of-mass frame,  $f = 0$ , these energy-momentum conservation equations reduce to

$$\partial_0 \rho + \frac{2\beta_0}{r} (\rho + p_\perp) + h_1 (\rho + p) = 0 , \quad (20.52a)$$

$$\partial_1 p + \frac{2\beta_1}{r} (p - p_\perp) + h_0 (\rho + p) = 0 . \quad (20.52b)$$

In a general situation where the mass-energy is a sum over several individual components  $x$ ,

$$T^{mn} = \sum_{\text{species } x} T_x^{mn} , \quad (20.53)$$

the individual mass-energy components  $x$  of the system each satisfy an energy-momentum conservation equation of the form

$$D_m T_x^{mn} = F_x^n , \quad (20.54)$$

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

$$\sum_{\text{species } x} F_x^n = 0 . \quad (20.55)$$

### 20.10.1 First law of thermodynamics

For an individual species  $x$ , the energy conservation equation (20.51a) in the centre-of-mass frame of the species,  $f_x = 0$ , can be written

$$D_m T_x^{m0} = \partial_0 \rho_x + (\rho_x + p_{\perp x}) \partial_0 \ln r^2 - (\rho_x + p_x) \partial_0 \ln \lambda_x = F_x^0 , \quad (20.56)$$

where  $\lambda_x$  is the radial “scale factor,” equation (20.27), in the centre-of-mass frame of the species (the scale factor is different in different frames). Equation (20.56) can be recognized as an expression of the first law of thermodynamics for a volume element  $V$  of species  $x$ , in the form

$$V^{-1} \left[ \partial_0(\rho_x V) + p_{\perp x} V_r \partial_0 V_{\perp} + p_x V_{\perp} \partial_0 V_r \right] = F_x^0 , \quad (20.57)$$

with transverse volume (area)  $V_{\perp} \propto r^2$ , radial volume (width)  $V_r \propto 1/\lambda_x$ , 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 (20.57) simplifies to

$$V^{-1} \left[ \partial_0(\rho_x V) + p_x \partial_0 V \right] = F_x^0 , \quad (20.58)$$

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,  $\alpha$ ,  $\beta_0$ , and  $\beta_1$ . 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 (20.33) give 4 equations for the 4 non-vanishing 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 (20.51). 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 (20.33c) and (20.33b) for  $\beta_0$  and  $\beta_1$ ,

$$\boxed{D_0 \beta_0 = -\frac{M}{r^2} - 4\pi r p} , \quad (20.59a)$$

$$\boxed{D_0 \beta_1 = 4\pi r f} , \quad (20.59b)$$

which are valid for any choice of tetrad frame, not just the centre-of-mass frame.

Equations (20.59) 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  $\beta_0$  equals the coordinate radial 4-velocity  $dr/d\tau = \partial_0 r = \beta_0$  of the tetrad frame, equation (20.5), and thus equation (20.59a) can be regarded as giving the proper radial acceleration

$D^2r/D\tau^2 = D\beta_0/D\tau = D_0\beta_0$  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  $\partial_0\beta_0 = D_0\beta_0 + \beta_1 h_0$ . The presence of the extra term  $\beta_1 h_0$ , 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 (20.59a) 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 (20.59a), 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 ingoing and outgoing 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 (20.59b) also plays an indispensable role. The quantity  $\beta_1 \equiv \partial_1 r$  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  $\beta_1$  determines which way an observer at rest in the tetrad frame thinks is “outwards,” the direction of larger circumferential radius  $r$ . A positive  $\beta_1$  means that the observer thinks the outward direction points away from the black hole, while a negative  $\beta_1$  means that the observer thinks the outward direction points towards from the black hole. Outside the outer horizon  $\beta_1$  is necessarily positive, because  $\beta_m$  must be spacelike there. But inside the horizon  $\beta_1$  may be either positive or negative. A tetrad frame can be defined as “ingoing” if the proper radial gradient  $\beta_1$  is positive, and “outgoing” if  $\beta_1$  is negative. In the Reissner-Nordström geometry, ingoing geodesics have positive energy, and outgoing geodesics have negative energy. However, the definition of ingoing or outgoing based on the sign of  $\beta_1$  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 (20.59b) shows that the proper rate of change  $D_0\beta_1$  in the radial gradient  $\beta_1$  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 ( $\beta_1$  positive) tend to see energy flux pointing away from the black hole ( $f$  positive), while outgoing observers ( $\beta_1$  negative) tend to see energy flux pointing towards the black hole ( $f$  negative). Thus the change in  $\beta_1$  tends to be in the same direction as  $\beta_1$ , amplifying  $\beta_1$  whatever its sign.

**Exercise 20.2 Birkhoff’s theorem.** Prove Birkhoff’s theorem from equations (20.59). Birkhoff’s theorem is the statement that any spherically symmetric spacetime that is devoid of energy-momentum outside some radius is Schwarzschild outside that radius.

**Concept question 20.3 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  $dr/dt = (\beta_0 \pm \beta_1)/\alpha$ , equation (20.11). 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 (20.59a), which is valid in any frame arbitrarily Lorentz-boosted in the radial direction, shows that  $\beta_0$ , 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  $\beta_0$  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 vierbein coefficient $\alpha$

Whereas the Einstein equations (20.59) give evolution equations for the vierbein coefficients  $\beta_0$  and  $\beta_1$ , there is no evolution equation for the vierbein coefficient  $\alpha$ . Indeed, the Einstein equations involve the vierbein coefficient  $\alpha$  only through the connections  $h_m$ , equations (20.21a) and (20.21b), and thus only as the radial derivative  $\partial \ln \alpha / \partial r$ , equations (20.22) and (20.23). 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,  $\alpha$  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  $\alpha$  itself.

Since  $\alpha$  is needed to propagate the equations from one coordinate time to the next (because  $\partial_0 = \alpha \partial / \partial t + \beta_0 \partial / \partial r$ ), it is necessary to construct  $\alpha$  by integrating  $h_0 \equiv -\beta_1 \partial \ln \alpha / \partial r$  along the radial direction  $r$  at each time step. The arbitrary normalization of  $\alpha$  at each step might be fixed by choosing  $\alpha$  to be unity at infinity, which corresponds to fixing the time coordinate  $t$  to equal the proper time at infinity.

In the particular case that the tetrad frame is taken to be in free-fall everywhere,  $h_0 = 0$ , as in the Gullstrand-Painlevé line-element, then  $\alpha$  is constant at fixed  $t$ , and without loss of generality it can be fixed equal to unity everywhere,  $\alpha = 1$ . I like to think of a free-fall frame as being realized 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  $1/\alpha$ , shift  $\beta_0/\alpha$ , and spatial metric  $g_{\alpha\beta} = \text{diag}(1/\beta_1^2, r^2, r^2 \sin^2\theta)$ . The non-vanishing components of the gravity  $K_a \equiv \Gamma_{a00}$  and of the extrinsic curvature  $K_{ab} \equiv \Gamma_{a0b}$  are

$$K_1 = h_0 , \quad (20.60a)$$

$$K_{11} = h_1 , \quad K_{22} = K_{33} = \frac{\beta_0}{r} . \quad (20.60b)$$

## 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 non-vanishing components of the electromagnetic field  $F_{mn}$  are then

$$-F_{01} = F_{10} = E = \frac{Q}{r^2} , \quad (20.61)$$

where  $E$  is the radial electric field, and  $Q(t, r)$  is the interior electric charge. Equation (20.61) 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 two of Maxwell's equations, the source-free ones (??). For the radial electric field (20.61), the other two Maxwell's equations, the sourced ones (22.54), are

$$\partial_1 Q = 4\pi r^2 q , \quad (20.62a)$$

$$\partial_0 Q = -4\pi r^2 j , \quad (20.62b)$$

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 (20.61), the electromagnetic energy-momentum tensor (16.144) in the tetrad frame is the diagonal tensor

$$T_e^{mn} = \frac{Q^2}{8\pi r^4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (20.63)$$

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  $\rho_e$  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

$$\rho_e = -p_e = p_{\perp e} = \frac{Q^2}{8\pi r^4} = \frac{E^2}{8\pi}. \quad (20.64)$$

The non-vanishing components of the covariant derivative  $D_m T_e^{mn}$  of the electromagnetic energy-momentum (20.63) are

$$D_m T_e^{m0} = \partial_0 \rho_e + \frac{4\beta_0}{r} \rho_e = -\frac{Q}{4\pi r^4} \partial_0 Q = -\frac{jQ}{r^2} = -jE, \quad (20.65a)$$

$$D_m T_e^{m1} = \partial_1 p_e + \frac{4\beta_1}{r} p_e = -\frac{Q}{4\pi r^4} \partial_1 Q = -\frac{qQ}{r^2} = -qE. \quad (20.65b)$$

The first expression (20.65a), 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 (20.65b), 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 (20.59b) then vanishes identically, while the Einstein equation (20.59a) becomes

$$\beta_1 h_0 = \frac{M}{r^2} + 4\pi r p, \quad (20.66)$$

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  $p$ , if positive as is the usual case for a star, enhances the inward gravitational force, helping to destabilize the star. Because  $\beta_0$  is zero, the interior mass  $M$  given by equation (20.10) reduces to

$$1 - 2M/r = \beta_1^2 . \quad (20.67)$$

When equations (20.66) and (20.67) are substituted into the momentum equation (20.149b), 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

$$\frac{\partial p}{\partial r} = - \frac{(\rho + p)(M + 4\pi r^3 p)}{r^2(1 - 2M/r)} .$$

(20.68)

In the Newtonian limit  $p \ll \rho$  and  $M \ll r$  this goes over to (with units restored)

$$\frac{\partial p}{\partial r} = - \rho \frac{GM}{r^2} , \quad (20.69)$$

which is the usual Newtonian equation of spherically symmetric hydrostatic equilibrium.

**Exercise 20.4 Constant density star.** Shortly after communicating to Einstein his celebrated solution, Schwarzschild (1916a) 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  $r$

$$\frac{dM}{dr} = 4\pi r^2 \rho , \quad (20.70)$$

and to the Volkov-Oppenheimer equation of hydrostatic equilibrium (20.68).

1. **Interior mass.** Suppose that the density  $\rho$  is constant. From equation (20.70) obtain an expression for the interior mass  $M$  as a function of radius  $r$  and the density  $\rho$ . [Hint: This is easy.]
2. **Hydrostatic equilibrium.** Given your expression for  $M$ , show that the Volkov-Oppenheimer equation rearranges to

$$\int_{p_c} \frac{dp}{(\rho + p)(\rho + 3p)} = - \int_0 \frac{4\pi r dr}{3 - 8\pi r^2 \rho} \quad (20.71)$$

where  $p_c$  is the central pressure, where the radius is zero,  $r = 0$ .

3. **Solve.** Integrate equation (20.71). 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

$$\frac{\rho + 3p_c}{\rho + p_c} = \sqrt{\frac{1}{1 - 2M_\star/R_\star}} \quad (20.72)$$

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

$$0 < \frac{2M_\star}{R_\star} < \frac{8}{9}. \quad (20.73)$$

5. **Comment.** Comment on what equation (20.73) implies physically. [Hint: What is the Schwarzschild radius?]
- 

## 20.15 Freely-falling dust

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).

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,  $f = p_1 = p_\perp = 0$ . The geodesic equation for the freely-falling dust implies that the proper acceleration vanishes,  $h_0 = 0$ , equation (20.24).

The equations admit two integrals of motion. The first integral of motion is the interior mass  $M$ , which equation (20.42a) shows is constant,  $\partial_0 M = 0$ , along the path of the freely-falling dust.

The second integral of motion is  $\beta_1$ , as follows from the second of the 2 Einstein equations (20.59). Since the acceleration vanishes, the covariant time derivative coincides with the directed time derivative,  $D_0 = \partial_0$ . The 2 Einstein equations (20.59) are then

$$\partial_0 \beta_0 = -\frac{M}{r^2}, \quad (20.74a)$$

$$\partial_0 \beta_1 = 0. \quad (20.74b)$$

The second equation (20.74b) shows that  $\beta_1$  is constant as claimed, an integral of motion along the path of the freely-falling dust. The first equation (20.74a), in combination with the definition (20.10) of the interior mass  $M$  and the constancy of  $\beta_1$ , recovers the constancy of  $M$ .

As discussed in §20.11.1, in a free-fall tetrad the vierbein coefficient  $\alpha$  can be set equal to unity everywhere,  $\alpha = 1$ . This corresponds to setting the time coordinate  $t$  equal to the proper time  $\tau$  attached to the freely-falling dust. The relation between (proper) time  $t$  and radius  $r$  along the path of the dust is obtained by integrating the equation for  $\beta_0 \equiv dr/dt$ ,

$$t = \tau = \int \frac{dr}{\sqrt{\beta_1^2 - 1 + 2M/r}}. \quad (20.75)$$

The relation between energy density  $\rho$  and the interior mass  $M$  is determined by equation (20.45). The initial conditions must be set up to satisfy this equation, but the evolution equations guarantee that equation (20.45) holds thereafter. The equation is a constraint equation: it is the Hamiltonian constraint.

**Exercise 20.5 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, 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  $\beta_0$  of dust shells is (with the index on  $\beta_0$  dropped for brevity)

$$\beta \equiv \beta_0 = -\sqrt{\frac{2M}{r}} . \quad (20.76)$$

Integrating equation (20.75) with  $\beta_1 = 1$  gives the relation between the radius  $r$  and time  $t$  along the trajectory of a shell with interior mass  $M$ ,

$$\frac{2}{3}r^{3/2} = \sqrt{2M}(t_c - t) , \quad (20.77)$$

where  $t_c(M)$  is the time at which the shell collapses to zero radius,  $r = 0$ .

A second simplifying assumption is self-similarity (see §20.17). In the present case, self-similar solutions occur when the collapse time  $t_c$  is proportional to the interior mass  $M$ ,

$$t_c = aM , \quad (20.78)$$

with  $a$  some positive dimensionless constant. Given the self-similar assumption (20.78), the relation (20.77) between the radius  $r$  and time  $t$  reduces to a cubic in the infall velocity  $\beta$ ,

$$a\beta^3 - \frac{2t}{r}\beta + \frac{4}{3} = 0 . \quad (20.79)$$

Equation (20.79) shows that the infall velocity  $\beta$  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  $\beta$  varies from 0 at  $t/r = -\infty$ , to  $-\infty$  at  $t/r = +\infty$ . The line-element is Gullstrand-Painlevé, equation (7.27), with  $\beta$  the real negative solution of the cubic (20.79).

Radial outgoing (+) and ingoing (−) null rays passing through the infalling dust follow

$$\frac{dr}{dt} = \beta \pm 1 . \quad (20.80)$$

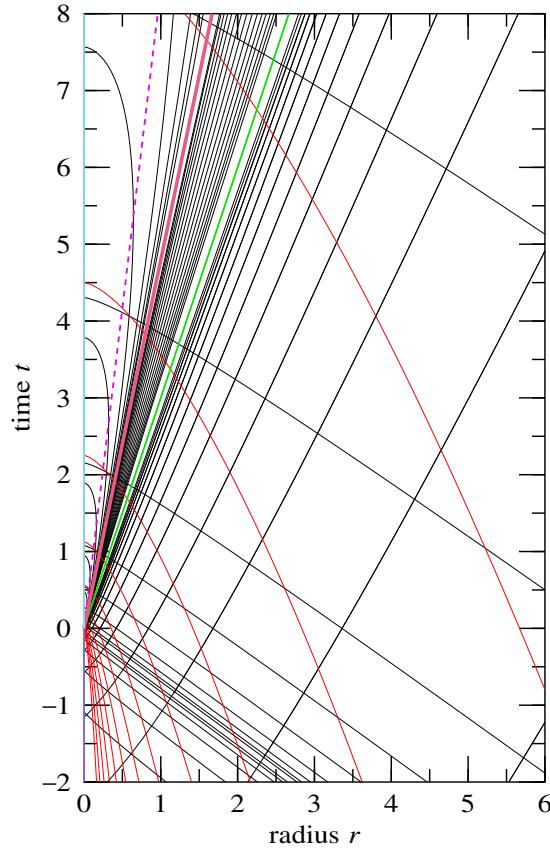


Figure 20.1 Spacetime diagram illustrating the formation of a naked singularity in self-similar collapse of dust, for  $a = 18$ , equation (20.78). Infalling (red) lines show trajectories of infalling dust, which are also contours of constant interior mass  $M$ . Approximately diagonal (black) lines 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. Radial null rays emanate from the naked singularity and extend to infinity in the region of spacetime between the Cauchy horizon (thick green line) and the true horizon (thick pink line). The apparent horizon (dashed pink line) is the locus of points where outgoing null rays turn around. The Cauchy, true, and apparent horizons are all straight lines emanating from the naked singularity at the origin.

Equation (20.80) for null geodesics can be recast as a differential equation between  $r$  and  $\beta$ , which integrates to

$$r \propto \exp \left[ \int \frac{2(\beta \pm 1)(2 - 3a\beta^3) d\beta}{\beta(\pm 4 - 2\beta \pm 3a\beta^3 + 3a\beta^4)} \right]. \quad (20.81)$$

The integrand in equation (20.81) is a rational function of  $\beta$ , 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 (20.81)), 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

$$a \geq \frac{26}{3} + 5\sqrt{3} \approx 17.3 . \quad (20.82)$$

For values of  $a$  exceeding the threshold (20.82), there is a naked singularity at the origin. The more negative of the two real roots marks the location of the absolute horizon, while the less negative marks the so-called Cauchy horizon. Radial outgoing null rays between the absolute 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 (20.82). The roots of the quartic in this case are

$$\begin{aligned} \beta &= -0.791475 , & t/r &= 4.79558 && \text{absolute horizon ,} \\ \beta &= -\frac{2}{3} , & t/r &= 3 && \text{Cauchy horizon .} \end{aligned} \quad (20.83)$$

All outgoing null geodesics inside the absolute horizon,  $\beta < -0.791475$ , in due course turn around and fall to zero radius,  $r = 0$ . Outgoing null geodesics between the absolute 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 Self-similar spherically symmetric spacetime

A third 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, 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. 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.

### 20.17.1 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  $\eta$  such that the geometry at any one time is conformally related to the geometry at any other time,

$$ds^2 = a(\eta)^2 \left[ g_{\eta\eta}^{(c)}(x) d\eta^2 + 2 g_{\eta x}^{(c)}(x) d\eta dx + g_{xx}^{(c)}(x) dx^2 + e^{2x} do^2 \right]. \quad (20.84)$$

Here the conformal metric coefficients  $g_{\mu\nu}^{(c)}(x)$  are functions only of conformal radius  $x$ , not of conformal time  $\eta$ . The choice  $e^{2x}$  of coefficient of  $do^2$  is a gauge choice of the conformal radius  $x$ , chosen here so as to bring the self-similar line-element into a form (20.88) below that resembles as far as possible the spherical line-element (20.1). In place of the conformal factor  $a(\eta)$  it is convenient to work with the circumferential radius  $r$

$$r \equiv a(\eta)e^x \quad (20.85)$$

which is to be considered as a function  $r(\eta, x)$  of the conformal coordinates  $\eta$  and  $x$ . The circumferential radius  $r$  has a gauge-invariant meaning, whereas neither  $a(\eta)$  nor  $x$  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  $x$ , independent of the circumferential radius  $r$ , are called dimensionless.

The fact that dimensionless quantities such as the conformal metric coefficients  $g_{\mu\nu}^{(c)}(x)$  are independent of

conformal time  $\eta$  implies that the tangent vector  $e_{\eta}$ , which by definition satisfies

$$\frac{\partial}{\partial \eta} = e_{\eta} \cdot \partial , \quad (20.86)$$

is a **conformal Killing vector**, also known as the homothetic vector. The tetrad-frame components of the conformal Killing vector  $e_{\eta}$  defines the tetrad-frame conformal Killing 4-vector  $\xi^m$

$$\frac{\partial}{\partial \eta} \equiv r \xi^m \partial_m , \quad (20.87)$$

in which the factor  $r$  is introduced so as to make  $\xi^m$  dimensionless. The conformal Killing vector  $e_{\eta}$  is the generator of the conformal time translation symmetry, and as such it is gauge-invariant (up to a global rescaling of conformal time,  $\eta \rightarrow b\eta$  for some constant  $b$ ). It follows that its dimensionless tetrad-frame components  $\xi^m$  constitute a tetrad 4-vector (again, up to global rescaling of conformal time).

### 20.17.2 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  $\xi^m$  made manifest:

$$ds^2 = r^2 \left[ -(\xi^0 d\eta)^2 + \frac{1}{\beta_1^2} (dx + \beta_1 \xi^1 d\eta)^2 + do^2 \right] . \quad (20.88)$$

The vierbein  $e_m{}^\mu$  and inverse vierbein  $e^m{}_\mu$  corresponding to the self-similar line-element (20.88) are

$$e_m{}^\mu = \frac{1}{r} \begin{pmatrix} 1/\xi^0 & -\beta_1 \xi^1/\xi^0 & 0 & 0 \\ 0 & \beta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/\sin \theta \end{pmatrix} , \quad e^m{}_\mu = r \begin{pmatrix} \xi^0 & 0 & 0 & 0 \\ \xi^1 & 1/\beta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sin \theta \end{pmatrix} . \quad (20.89)$$

It is straightforward to see that the coordinate time components of the inverse vierbein must be  $e^m{}_\eta = r \xi^m$ , since  $\partial/\partial\eta = e^m{}_\eta \partial_m$  equals  $r \xi^m \partial_m$ , equation (20.87).

### 20.17.3 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  $\beta_m$  (which unlike  $\xi^m$  is independent of any global rescaling of conformal time)

$$\beta_m \equiv \partial_m r . \quad (20.90)$$

It is straightforward to check that  $\beta_1$  defined by equation (20.90) is consistent with its appearance in the vierbein (20.89) provided that  $r \propto e^x$  as earlier assumed, equation (20.85).

With two distinct dimensionless tetrad 4-vectors in hand,  $\beta_m$  and the conformal Killing vector  $\xi^m$ , three gauge-invariant dimensionless scalars can be constructed,  $\beta^m \beta_m$ ,  $\xi^m \beta_m$ , and  $\xi^m \xi_m$ ,

$$1 - \frac{2M}{r} = -\beta^m \beta_m = -\beta_0^2 + \beta_1^2 , \quad (20.91)$$

$$v \equiv \xi^m \beta_m = \frac{1}{r} \frac{\partial r}{\partial \eta} = \frac{1}{a} \frac{\partial a}{\partial \eta}, \quad (20.92)$$

$$\Delta \equiv -\xi^m \xi_m = (\xi^0)^2 - (\xi^1)^2. \quad (20.93)$$

Equation (20.91) is essentially the same as equation (20.10). The dimensionless quantity  $v$ , equation (20.92), may be interpreted as a measure of the expansion velocity of the self-similar spacetime. Equation (20.92) shows that  $v$  is a function only of  $\eta$  (since  $a(\eta)$  is a function only of  $\eta$ ), and it therefore follows that  $v$  must be constant (since being dimensionless means that  $v$  must be a function only of  $x$ ). Equation (20.92) then also implies that the conformal factor  $a(\eta)$  must take the form

$$a(\eta) = e^{v\eta}. \quad (20.94)$$

Because of the freedom of a global rescaling of conformal time, it is possible to set  $v = 1$  without loss of generality, but in practice it is convenient to keep  $v$ , because it is then transparent how to take the static limit  $v \rightarrow 0$ . FIX  $v$  BY MATCHING CONFORMAL TIME TO DISTANT TIME. Equation (20.94) along with (20.85) shows that the circumferential radius  $r$  is related to the conformal coordinates  $\eta$  and  $x$  by

$$r = e^{v\eta+x}. \quad (20.95)$$

The dimensionless quantity  $\Delta$ , equation (20.93), is the dimensionless horizon function: horizons occur where the horizon function vanishes

$$\Delta = 0 \quad \text{at horizons}. \quad (20.96)$$

#### 20.17.4 Self-similar diagonal line-element

The self-similar line-element (20.88) can be brought to diagonal form by a coordinate transformation to diagonal conformal coordinates  $\eta_\times, x_\times$  (subscripted  $\times$  for diagonal)

$$\eta \rightarrow \eta_\times = \eta + f(x), \quad x \rightarrow x_\times = x - vf(x), \quad (20.97)$$

which leaves unchanged the conformal factor  $r$ , equation (20.95). The resulting diagonal metric is

$$ds^2 = r^2 \left( -\Delta d\eta_\times^2 + \frac{dx_\times^2}{1 - 2M/r + v^2/\Delta} + do^2 \right). \quad (20.98)$$

The diagonal line-element (20.98) 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 (20.88). The frame can be called the **similarity frame**. The form of the metric coefficients follows from the line-element (20.88) and the gauge-invariant scalars (20.91)–(20.93).

The conformal Killing vector in the similarity frame is  $\xi^m = \{\Delta^{1/2}, 0, 0, 0\}$ , 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

$$\xi^m = u^m \Delta^{1/2} \quad (20.99)$$

where  $u^m$  is the 4-velocity of the similarity frame with respect to the general frame. This shows that the conformal Killing vector  $\xi^m$  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

$$\text{proper 3-velocity of similarity frame through tetrad frame} = \frac{\xi^1}{\xi^0}. \quad (20.100)$$

### 20.17.5 Ray-tracing line-element

It proves useful to introduce a “ray-tracing” conformal radial coordinate  $X$  related to the coordinate  $x_\times$  of the diagonal line-element (20.98) by

$$dX \equiv \frac{\Delta dx_\times}{[(1 - 2M/r)\Delta + v^2]^{1/2}}. \quad (20.101)$$

In terms of the ray-tracing coordinate  $X$ , the diagonal metric is

$$ds^2 = r^2 \left( -\Delta d\eta_\times^2 + \frac{dX^2}{\Delta} + do^2 \right). \quad (20.102)$$

For the Reissner-Nordström geometry,  $\Delta = (1 - 2M/r)/r^2$ ,  $\eta_\times = t$ , and  $X = -1/r$ .

FIX: CHANGE SIGN OF  $X$ .

### 20.17.6 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

$$\frac{d}{d\tau} \frac{\partial L}{\partial u^\eta} = \frac{\partial L}{\partial \eta}, \quad (20.103)$$

with effective Lagrangian  $L = \frac{1}{2}g_{\mu\nu}u^\mu u^\nu$  for a particle with 4-velocity  $u^\mu$ . The self-similar metric depends on the conformal time  $\eta$  only through the overall conformal factor  $g_{\mu\nu} \propto a(\eta)^2$ . The derivative of the conformal factor is given by  $\partial \ln a / \partial \eta = v$ , equation (20.92), so it follows that  $\partial L / \partial \eta = 2vL$ . For a massive particle, for which conservation of rest mass implies  $g_{\mu\nu}u^\mu u^\nu = -1$ , Lagrange's equations (20.103) thus yield

$$\frac{du_\eta}{d\tau} = -v. \quad (20.104)$$

In the limit of zero accretion rate,  $v \rightarrow 0$ , equation (20.104) would integrate to give  $u_\eta$  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 (20.104) integrates to

$$u_\eta = -v\tau, \quad (20.105)$$

in which an arbitrary constant of integration has been absorbed into a shift in the zero point of the proper time  $\tau$ . Although the above derivation was for a massive particle, it holds also for a massless particle, with the understanding that the proper time  $\tau$  is constant along a null geodesic. The quantity  $u_{\eta}$  in equation (20.105) is the covariant time component of the coordinate-frame 4-velocity  $u^{\mu}$  of the particle; it is related to the covariant components  $u_m$  of the tetrad-frame 4-velocity of the particle by

$$u_{\eta} = e^m \eta u_m = r \xi^m u_m . \quad (20.106)$$

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

$$u_{\eta} = -v\tau , \quad u_{\phi} = L , \quad u_{\mu} u^{\mu} = -1 , \quad (20.107)$$

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 (20.107) takes its simplest form in the conformal coordinates  $\{\eta_x, X, \theta, \phi\}$  of the ray-tracing metric (20.102)

$$u^{\eta} = \frac{v\tau}{r^2\Delta} , \quad u^X = \pm \frac{1}{r^2} [v^2\tau^2 - (r^2 + L^2)\Delta]^{1/2} , \quad u^{\phi} = \frac{L}{r^2} . \quad (20.108)$$

### 20.17.7 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  $L \rightarrow \infty$ . To obtain finite results, define an affine parameter  $\lambda$  by CHECK  $d\lambda \equiv v\tau d\tau$ , and a 4-velocity in terms of it by  $v^{\mu} \equiv dx^{\mu}/d\lambda$ . The integrals of motion (20.107) then become, for a null geodesic,

$$v_{\eta_x} = -1 , \quad v_{\phi} = J , \quad v_{\mu} v^{\mu} = 0 , \quad (20.109)$$

where  $J \equiv L/(v\tau)$  is the (dimensionless) conformal angular momentum of the particle. The 4-velocity  $v^{\mu}$  along the null geodesic is then, in terms of the coordinates of the ray-tracing metric (20.102),

$$v^{\eta} = \frac{1}{r^2\Delta} , \quad v^X = \pm \frac{1}{r^2} (1 - J^2\Delta)^{1/2} , \quad v^{\phi} = \frac{J}{r^2} . \quad (20.110)$$

Equations (20.110) yield the shape of a null geodesic by quadrature

$$\phi = \int \frac{J dX}{(1 - J^2\Delta)^{1/2}} . \quad (20.111)$$

Equation (20.111) 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 (20.111) 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 non-stationary self-similar solutions there is no conformal radius where photons can orbit for ever

(to remain at fixed conformal radius, 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 (20.111) not only vanishes,  $v^X = 0$ , but is an extremum, which happens where the horizon function  $\Delta$  is an extremum,

$$\frac{d\Delta}{dX} = 0 \quad \text{at photon sphere equivalent .} \quad (20.112)$$

### 20.17.8 Dimensional analysis

Dimensional analysis shows that the conformal coordinates  $x^\mu \equiv \{\eta, x, \theta, \phi\}$  and the tetrad metric  $\gamma_{mn}$  are dimensionless, while the coordinate metric  $g_{\mu\nu}$  scales as  $r^2$ ,

$$x^\mu \propto r^0 , \quad \gamma_{mn} \propto r^0 , \quad g_{\mu\nu} \propto r^2 . \quad (20.113)$$

The vierbein  $e_m{}^\mu$  and inverse vierbein  $e^m{}_\mu$ , equations (20.89), scale as

$$e_m{}^\mu \propto r^{-1} , \quad e^m{}_\mu \propto r . \quad (20.114)$$

Coordinate derivatives  $\partial/\partial x^\mu$  are dimensionless, while directed derivatives  $\partial_m$  scale as  $1/r$ ,

$$\frac{\partial}{\partial x^\mu} \propto r^0 , \quad \partial_m \propto r^{-1} . \quad (20.115)$$

The tetrad connections  $\Gamma_{kmn}$  and the tetrad-frame Riemann tensor  $R_{klmn}$  scale as

$$\Gamma_{kmn} \propto r^{-1} , \quad R_{klmn} \propto r^{-2} . \quad (20.116)$$

### 20.17.9 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 and pressure of any species must scale as

$$\rho \propto p \propto p_\perp \propto r^{-2} . \quad (20.117)$$

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

$$w \equiv p/\rho , \quad w_\perp \equiv p_\perp/\rho , \quad (20.118)$$

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 (20.54), must scale as

$$F^n \propto r^{-3} . \quad (20.119)$$

Define dimensionless variables  $z$  and  $s^n$

$$\rho \equiv 4\pi r^2 z , \quad F^n \equiv 4\pi r^3 s^n . \quad (20.120)$$

### 20.17.10 Electrical conductivity

As an example, if any of the fluids is charged, then that charged fluid will experience a Lorentz force from the electric field, and will therefore exchange momentum with the electric field. If the fluid is non-conducting, 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  $q$ , the radial electric current  $j$ , and the radial electric field  $E \equiv Q/r^2$  scale as

$$q \propto j \propto r^{-2} , \quad E \propto r^{-1} , \quad (20.121)$$

consistent with the requirement that the flux of energy and momentum on the right hand sides of equations (20.65) scale as  $F^n \propto r^{-3}$ . In diffusive electrical conduction in a fluid of conductivity  $\sigma$ , an electric field  $E$  gives rise to a current

$$j = \sigma E , \quad (20.122)$$

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  $\rho$  of the conducting fluid

$$\sigma = \kappa \rho^{1/2} , \quad (20.123)$$

where  $\kappa$  is a dimensionless conductivity constant. But current saturates

$$j = \frac{\kappa \lambda \rho_e^{1/2} \rho}{\lambda \rho_e^{1/2} + \kappa \rho^{1/2}} \rightarrow \begin{cases} \kappa \rho_e^{1/2} \rho^{1/2} & \lambda \rho_e^{1/2} \gg \kappa \rho^{1/2} , \\ \lambda \rho & \lambda \rho_e^{1/2} \ll \kappa \rho^{1/2} . \end{cases} \quad (20.124)$$

The form (20.123) is required by self-similarity, and is not necessarily realistic (although it is realistic that the conductivity increases with density). However, the conductivity (20.123) 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)

$$\sigma = \frac{C}{e^2 \ln e^{-1}} \frac{kT}{\hbar} \quad (20.125)$$

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  $\sigma$  times the characteristic timescale  $t_{\text{BH}} \equiv GM/c^3$  of the black hole, which is of order

$$\sigma t_{\text{BH}} \sim \frac{T}{T_{\text{BH}}} \quad (20.126)$$

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  $T_{\text{BH}}$  is  $8\pi$  times the Hawking temperature). In the astronomical situation considered here the temperature  $T$  of the plasma is huge compared to the characteristic temperature  $T_{\text{BH}}$  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.

Fluids might also interact with each other by collisions between particles of the fluids.

$$F \propto r^{-3} \propto \rho^{3/2}. \quad (20.127)$$

Assume

$$(20.128)$$

### 20.17.11 Tetrad connections

The expressions for the tetrad connections for the self-similar spacetime are the same as those (20.21) for a general spherically symmetric spacetime. Expressions (20.22) and (20.23) for the proper radial acceleration  $h_0$  and the radial Hubble parameter  $h_1$  translate in the self-similar spacetime to

$$h_0 \equiv \partial_1 \ln(r\xi^0), \quad h_1 \equiv \partial_0 \ln(r\xi^1). \quad (20.129)$$

Comparing equations (20.129) to equations (20.22) and (20.27) shows that the vierbein coefficient  $\alpha$  and scale factor  $\lambda$  translate in the self-similar spacetime to

$$\alpha = \frac{1}{r\xi^0}, \quad \lambda = \frac{1}{r\xi^1}. \quad (20.130)$$

### 20.17.12 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 (20.28)–(20.33).

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.

### 20.17.13 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  $A(x)$  is independent of conformal time  $\eta$ . It follows that the partial derivative of any dimensionless quantity  $A(x)$  with respect to conformal time  $\eta$  vanishes

$$0 = \frac{\partial A(x)}{\partial \eta} = \xi^m \partial_m A(x) = (\xi^0 \partial_0 + \xi^1 \partial_1) A(x). \quad (20.131)$$

Consequently the directed radial derivative  $\partial_1 F$  of a dimensionless quantity  $A(x)$  is related to its directed time derivative  $\partial_0 F$  by

$$\partial_1 A(x) = -\frac{\xi^0}{\xi^1} \partial_0 A(x). \quad (20.132)$$

Equation (20.132) allows radial derivatives to be converted to time derivatives.

### 20.17.14 Integration variable

It is desirable to choose an integration variable that varies monotonically. A natural choice is the proper time  $\tau$  in the tetrad frame, since this is guaranteed to increase monotonically. Since the 4-velocity at rest in the tetrad frame is by definition  $u^m = \{1, 0, 0, 0\}$ , 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 (20.101), 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  $A(x)$  in the tetrad frame is related to its derivative  $dA/dX$  with respect to the ray-tracing variable  $X$  by

$$\partial_0 A = u^m \partial_m A = u^X \partial_X A = -\frac{\xi^1}{r} \frac{dA}{dX}. \quad (20.133)$$

In the third expression,  $u^X \partial_X A$  is  $u^m \partial_m A$  expressed in the similarity frame of §20.17.4, the time contribution  $u^{\eta_X} \partial_{\eta_X} A$  vanishing in the similarity frame because it is proportional to the conformal time derivative  $\partial A / \partial \eta_X = 0$ . In the last expression of (20.156),  $u^X$  has been replaced by  $-u^1 = -\xi^1 / \Delta^{1/2}$  in view of equation (20.99), the minus sign coming from the fact that  $u^X$  is the radial component of the tetrad 4-velocity of the tetrad frame relative to the similarity frame, while  $u^1$  in equation (20.99) is the radial component of the tetrad 4-velocity of the similarity frame relative to the tetrad frame. Also in the last expression of (20.156), the directed derivative  $\partial_X$  with respect to the ray-tracing variable  $X$  has been translated into its coordinate partial derivative,  $\partial_X = (\Delta^{1/2}/r) \partial/\partial X$ , which follows from the metric (20.102).

In summary, the chosen integration variable is the dimensionless ray-tracing variable  $-X$  (with a minus because  $-X$  is monotonically increasing), the derivative with respect to which, acting on any dimensionless function, is related to the proper time derivative in any tetrad frame (not just the baryonic frame) by

$$-\frac{d}{dX} = \frac{r}{\xi^1} \partial_0. \quad (20.134)$$

Equation (20.134) involves  $\xi^1$ , which is proportional to the proper velocity of the tetrad frame through the similarity frame, equation (20.100), 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.

### 20.17.15 Entropy

Substituting the self-similar expression (20.130) for  $\lambda$  into the energy conservation equation (20.56) for a species in its own centre-of-mass frame gives

$$\frac{d \ln [\rho r^{2(1+w_\perp)} (r \xi^1)^{1+w}]}{dx} = \frac{r F^0}{\rho} . \quad (20.135)$$

For a fluid with isotropic equation of state  $w = w_\perp$ ,

$$\frac{d \ln S}{dx} = \frac{r F^0}{\rho(1+w)} , \quad (20.136)$$

where in which the  $S$  is (up to an arbitrary constant) the entropy of a comoving volume element  $V \propto r^3 \xi^1$  of the fluid

$$S \equiv r^3 \xi^1 \rho^{1/(1+w)} . \quad (20.137)$$

That equation (20.163)

Two more equations complete the suite. The first, which represents energy conservation for the fluid, can be written as an equation governing the entropy  $S$  of the fluid

$$-\frac{d \ln S}{dX} = \frac{\sigma Q^2}{r(1+w)\rho\xi^1} , \quad (20.138)$$

really is an entropy equation can be confirmed by rewriting it as

$$\frac{1}{V} \left( \frac{d\rho V}{d\tau} + p \frac{dV}{d\tau} \right) = jE = \frac{\sigma Q^2}{r^4} , \quad (20.139)$$

in which  $jE$  is recognized as the Ohmic dissipation, the rate per unit volume at which the volume element  $V$  is being heated.

The final equation represents electromagnetic energy conservation, equation (20.65a), which can be written

$$-\frac{d \ln Q}{dX} = -\frac{4\pi r\sigma}{\xi^1} . \quad (20.140)$$

The (heat) energy going into the fluid is balanced by the (free) energy coming out of the electromagnetic field.

### 20.17.16 Integrals of motion

As remarked above, equation (20.131), in self-similar solutions  $\xi^m \partial_m A(x) = 0$  for any dimensionless function  $A(x)$ . If both the directed derivatives  $\partial_0 A(x)$  and  $\partial_1 A(x)$  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

$$0 = r \xi^m \partial_m \beta_0 = r \beta_1 (\xi^0 h_0 + \xi^1 h_1) - \xi^0 \left( \frac{M}{r} + 4\pi r^2 p \right) + \xi^1 4\pi r^2 f , \quad (20.141a)$$

$$0 = r \xi^m \partial_m \beta_1 = r \beta_0 (\xi^0 h_0 + \xi^1 h_1) + \xi^1 \left( \frac{M}{r} - 4\pi r^2 \rho \right) + \xi^0 4\pi r^2 f . \quad (20.141b)$$

Taking  $\xi^1$  times (20.141a) plus  $\xi^0$  times (20.141b), and then  $\beta_0$  times (20.141a) minus  $\beta_1$  times (20.141b), gives

$$0 = r v (\xi^0 h_0 + \xi^1 h_1) - 4\pi r^2 [\xi^0 \xi^1 (\rho + p) - ((\xi^0)^2 + (\xi^1)^2) f] , \quad (20.142a)$$

$$0 = r \xi^m \partial_m \frac{M}{r} = -v \frac{M}{r} + 4\pi r^2 [\beta_1 \xi^1 \rho - \beta_0 \xi^0 p + (\beta_0 \xi^1 - \beta_1 \xi^0) f] . \quad (20.142b)$$

The quantities in square brackets on the right hand sides of equations (20.142) are scalars for each species  $a$ , so equations (20.142) can also be written

$$r v (\xi^0 h_0 + \xi^1 h_1) = \varepsilon , \quad \varepsilon \equiv 4\pi r^2 \sum_{\text{species } a} \xi_a^0 \xi_a^1 (\rho_a + p_a) , \quad (20.143a)$$

$$v \frac{M}{r} = 4\pi r^2 \sum_{\text{species } a} (\beta_{a,1} \xi_a^1 \rho_a - \beta_{a,0} \xi_a^0 p_a) , \quad (20.143b)$$

where the sum is over all species  $a$ , and  $\beta_{a,m}$  and  $\xi_a^m$  are the 4-vectors  $\beta_m$  and  $\xi^m$  expressed in the rest frame of species  $a$ . Equations (20.143) are scalar equations, valid in any frame of reference.

For any fluid with equation of state  $p = w\rho$ , a further integral comes from considering

$$0 = r \xi^m \partial_m (r^2 p) = r [w \xi^0 \partial_0 (r^2 \rho) + \xi^1 \partial_1 (r^2 p)] , \quad (20.144)$$

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 (20.144) reduces to

$$0 = r \xi^m \partial_m \frac{Q}{r} = -v \frac{Q}{r} + 4\pi r^2 (\xi^1 q - \xi^0 j) , \quad (20.145)$$

which is valid in any radial tetrad frame, not just the centre-of-mass frame.

### 20.17.17 Equations

Define dimensionless parameters

$$z_a \equiv 4\pi r^2 \rho_a . \quad (20.146)$$

Flux of stuff into species  $a$  from species  $b$

$$F_{ab}^m \quad (20.147)$$

Into charged species from electromagnetic field, equations (20.65),

$$F^0 = jE, \quad F^1 = qE. \quad (20.148)$$

From (20.52a),

$$\partial_0 \rho + \frac{2\beta_0}{r}(\rho + p_\perp) + h_1(\rho + p) = F^t, \quad (20.149a)$$

$$\partial_1 p + \frac{2\beta_1}{r}(p - p_\perp) + h_0(\rho + p) = F^r, \quad (20.149b)$$

In the centre-of-mass frame of a given species,

$$(1+w)r(\xi^1 h_0 + w\xi^0 h_1) - 2w_\perp(\xi^1 \beta_1 - w\xi^0 \beta_0) = \frac{r}{\rho}(\xi^1 F^1 + w\xi^0 F^0) \quad (20.150)$$

Rearranges to

$$rh_0 = \frac{2w_\perp \xi^1 (\xi^1 \beta_1 - w\xi^0 \beta_0) - w(1+w)\xi^0(\varepsilon/v) + (r/\rho)\xi^1(\xi^1 F^1 + w\xi^0 F^0)}{(1+w)[(\xi^1)^2 - w(\xi^0)^2]} \quad (20.151)$$

where

$$\varepsilon \equiv \sum_{\text{species } a} \xi_a^0 \xi_a^1 (1+w_a) z_a \quad (20.152)$$

summed over all species  $a$  (including the one under consideration), where  $\xi_a^m$  is in the rest-frame of species  $a$ .

### 20.17.18 Summary of equations for a single charged fluid

GENERALISE TO ARBITRARY NUMBER OF INTERACTING FLUIDS.

For reference, it is helpful to collect here the full set of equations governing self-similar spherically symmetric evolution in the case of a single charged fluid with isotropic equation of state

$$p = p_\perp = w\rho, \quad (20.153)$$

and conductivity

$$\sigma = \kappa \rho^{1/2}. \quad (20.154)$$

In accordance with the arguments in §20.17.9, equations (20.118) and (20.123), self-similarity requires that the pressure-to-density ratio  $w$  and the conductivity coefficient  $\kappa$  both be (dimensionless) constants.

It is natural to work in the centre-of-mass frame of the fluid, which also coincides with the centre-of-mass frame of the fluid plus electric field (the electric field, being invariant under Lorentz boosts, does not pick out any particular radial frame).

The proper time  $\tau$  in the fluid frame evolves as

$$-\frac{d\tau}{dX} = \frac{r}{\xi^1}, \quad (20.155)$$

which follows from equation (20.134) and the fact that  $\partial_0\tau = 1$ . The circumferential radius  $r$  evolves along the path of the fluid as

$$-\frac{d\ln r}{dX} = \frac{\beta_0}{\xi^1}. \quad (20.156)$$

Although it is straightforward to write down the equations governing how the tetrad frame moves through the conformal coordinates  $\eta$  and  $x$ , there is not much to be gained from this because the conformal coordinates have no fundamental physical significance.

Next, the defining equations (20.129) 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  $\xi^m$

$$-\frac{d\xi^0}{dX} = \beta_1 - rh_0, \quad (20.157a)$$

$$-\frac{d\xi^1}{dX} = -\beta_0 + rh_1, \quad (20.157b)$$

in which, in the formula for  $h_0$ , equation (20.132) has been used to convert the conformal radial derivative  $\partial_1$  to the conformal time derivative  $\partial_0$ , and thence to  $-d/dX$  by equation (20.134).

Next, the Einstein equation (20.33c) (with coordinates relabelled per (??) in the centre-of-mass frame (20.40)) yield evolution equations for the time and radial components of the vierbein coefficients  $\beta_m$

$$-\frac{d\beta_0}{dX} = -\frac{\beta_1}{\xi^0} rh_1 - \frac{r^2}{2\xi^0} G^{01}, \quad (20.158a)$$

$$-\frac{d\beta_1}{dX} = \frac{\beta_0}{\xi^1} rh_0 + \frac{r^2}{2\xi^1} G^{01}, \quad (20.158b)$$

where again, in the formula for  $\beta_0$ , equation (20.132) has been used to convert the conformal radial derivative  $\partial_1$  to the conformal time derivative  $\partial_0$ . The 4 evolution equations (20.157) and (20.158) for  $\xi^m$  and  $\beta_m$  are not independent: they are related by  $\xi^m\beta_m = v$ , a constant, equation (20.92). 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  $\beta_1$ , 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  $\Delta$ , equations (20.91) and (20.93), in a fashion that ensures numerical stability:

$$\xi^0 - \xi^1 = \frac{2v - (\xi^0 + \xi^1)(\beta_0 + \beta_1)}{\beta_0 - \beta_1}, \quad (20.159a)$$

$$\frac{2M}{r} = 1 + (\beta_0 + \beta_1)(\beta_0 - \beta_1), \quad (20.159b)$$

$$\Delta = (\xi^0 + \xi^1)(\xi^0 - \xi^1) . \quad (20.159c)$$

The evolution equations (20.157) and (20.158) involve  $h_0$  and  $h_1$ . The integrals of motion considered in §20.17.16 yield explicit expressions for  $h_0$  and  $h_1$  not involving any derivatives. For the Hubble parameter  $h_1$ , equation (20.142a) gives

$$rh_1 = -\frac{\xi^0}{\xi^1} rh_0 + \frac{1}{V} \sum_{\text{species } a} \xi_a^0 \xi_a^1 (1 + w_a) z_a , \quad (20.160)$$

where  $\varepsilon$  is the dimensionless enthalpy

$$\varepsilon_a \equiv 4\pi r^2 (1 + w_a) \rho_a . \quad (20.161)$$

For the proper acceleration  $h_0$ , a somewhat lengthy calculation starting from the integral of motion (20.144), and simplifying using the integral of motion (20.145) for  $Q$ , the expression (20.160) for  $h_1$ , Maxwell's equation (20.62b) [with the relabelling (??)], and the conductivity (20.154) in Ohm's law (20.122), gives

$$rh_0 = \frac{\xi^1 \{ 8\pi w_\perp (\beta_1 \xi^1 - w \beta_0 \xi^0) r^2 \rho + [v + (1+w) 4\pi r \sigma \xi^0] Q^2 / r^2 - w (4\pi \xi^0 \varepsilon)^2 / V \}}{4\pi \varepsilon [(\xi^1)^2 - w(\xi^0)^2]} . \quad (20.162)$$

Two more equations complete the suite. The first, which represents energy conservation for the fluid, can be written as an equation governing the entropy  $S$  of the fluid

$$-\frac{d \ln S}{dX} = \frac{\sigma Q^2}{r(1+w)\rho\xi^1} , \quad (20.163)$$

in which the  $S$  is (up to an arbitrary constant) the entropy of a comoving volume element  $V \propto r^3 \xi^1$  of the fluid

$$S \equiv r^3 \xi^1 \rho^{1/(1+w)} . \quad (20.164)$$

That equation (20.163) really is an entropy equation can be confirmed by rewriting it as

$$\frac{1}{V} \left( \frac{d\rho V}{d\tau} + p \frac{dV}{d\tau} \right) = jE = \frac{\sigma Q^2}{r^4} , \quad (20.165)$$

in which  $jE$  is recognized as the Ohmic dissipation, the rate per unit volume at which the volume element  $V$  is being heated.

The final equation represents electromagnetic energy conservation, equation (20.65a), which can be written

$$-\frac{d \ln Q}{dX} = -\frac{4\pi r \sigma}{\xi^1} . \quad (20.166)$$

The (heat) energy going into the fluid is balanced by the (free) energy coming out of the electromagnetic field.

# 21

---

## The interiors of spherical black holes

THIS CHAPTER NEEDS UPDATING.

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).

### 21.1 The mechanism of mass inflation

#### 21.1.1 Reissner-Nordström phase

Figure 21.1 illustrates how the two Einstein equations (20.59) produce the three phases of mass inflation inside a charged spherical black hole.

During the initial phase, illustrated in the top panel of Figure 21.1, the spacetime geometry is well-approximated by the vacuum, Reissner-Nordström geometry. During this phase the radial energy flux  $f$  is effectively zero, so  $\beta_1$  remains constant, according to equation (20.59b). The change in the radial velocity  $\beta_0$ , equation (20.59a), depends on the competition between the Newtonian gravitational force  $-M/r^2$ , which is always attractive (tending to make the radial velocity  $\beta_0$  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  $\beta_0$  slows down. This regime, where the (negative) radial velocity  $\beta_0$  is slowing down (becoming less negative), while  $\beta_1$  remains constant, is illustrated in the top panel of Figure 21.1.

If the initial Reissner-Nordström phase were to continue, then the radial 4-gradient  $\beta_m$  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  $\beta_m$  points in one direction for ingoing frames, and in the opposite direction for outgoing frames. If  $\beta_m$  becomes lightlike, then ingoing and outgoing

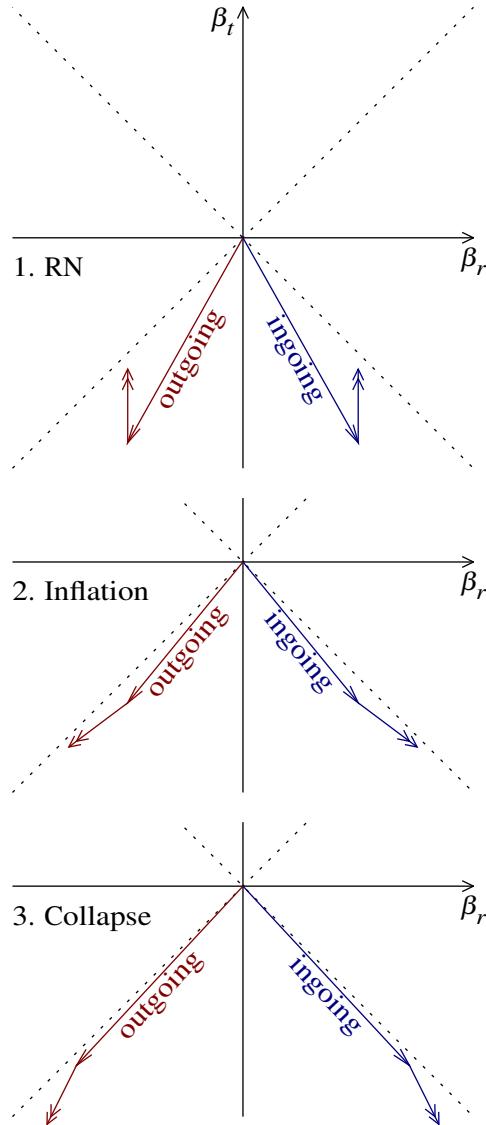


Figure 21.1 Spacetime diagrams of the tetrad-frame 4-vector  $\beta_m$ , equation (20.9), illustrating qualitatively the three successive phases of mass inflation: 1. (top) the Reissner-Nordström phase, where inflation ignites; 2. (middle) the inflationary phase itself; and 3. (bottom) the collapse phase, where inflation comes to an end. In each diagram, the arrowed lines labelled ingoing and outgoing 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 (20.59). Inside the horizon of a black hole, all locally inertial frames necessarily fall inward, so the radial velocity  $\beta_0 \equiv \partial_t r$  is always negative. A locally inertial frame is ingoing or outgoing depending on whether the proper radial gradient  $\beta_1 \equiv \partial_{rr} r$  measured in that frame is positive or negative.

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 ingoing or outgoing but not both, then  $\beta_m$  could indeed become lightlike. But if both ingoing and outgoing matter are present, even in the tiniest amount, then it is physically impossible for the ingoing and outgoing frames to stream through each other at the speed of light.

If both ingoing and outgoing 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 (20.59). This is how mass inflation is ignited.

### 21.1.2 Inflationary phase

The infalling matter now enters the second, mass inflationary phase, illustrated in the middle panel of Figure 21.1.

During this phase, the gravitational force on the right hand side of the Einstein equation (20.59a) is dominated by the pressure  $p$  produced by the counter-streaming ingoing and outgoing 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  $\beta_0$  (makes it more negative). At the same time, the radial gradient  $\beta_1$  is being driven by the energy flux  $f$ , equation (20.59b). 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  $\rho$  moving at 4-velocity  $u^m \equiv \{u^t, u^r, 0, 0\}$ , the streaming energy flux would be  $f \sim \rho u^t u^r$ , while the streaming pressure would be  $p \sim \rho (u^r)^2$ . Thus their ratio  $f/p \sim u^t/u^r$  is slightly greater than one. It follows that, as illustrated in the middle panel of Figure 21.1, the change in  $\beta_1$  slightly exceeds the change in  $\beta_0$ , which drives the 4-vector  $\beta_m$ , 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 ingoing and outgoing streams increase the gravitational force. But, as illustrated in the middle panel of Figure 21.1, the gravitational force acts in opposite directions for ingoing and outgoing 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.

### 21.1.3 Collapse phase

It might seem that inflation is locked into an exponential growth from which there is no exit. But the Einstein equations (20.59) have one more trick up their sleeve.

For the counter-streaming velocity to continue to increase requires that the change in  $\beta_1$  from equation (20.59b) continues to exceed the change in  $\beta_0$  from equation (20.59a). 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  $\beta_0$ , equation (20.59a). As will be seen in the examples of the next two sections, §§?? and ??, at least in the case of pressureless streams the mass term exponentiates slightly faster than the pressure term. 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  $\beta_m$ , 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.

## 21.2 The far future?

The Penrose diagram of a Reissner-Nordström 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 previous sections, 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.

Ingoing and outgoing 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. If the streaming 4-velocity between ingoing and outgoing streams is  $u_{ba}^m$ , equation (??), then the proper time  $d\tau_b$  that elapses on stream  $b$  observed by stream  $a$  equals the blueshift factor  $u_{ba}^t$  times the proper time  $d\tau_a$  experienced by stream  $a$ ,

$$d\tau_b = u_{ba}^t d\tau_a . \quad (21.1)$$

### 21.2.1 Inflationary phase

A physically relevant timescale for the observing stream  $a$  is how long it takes for the blueshift to increase by one  $e$ -fold, which is  $d\tau_a/d\ln u_{ba}^t$ . During the inflationary phase, stream  $a$  sees the amount of time elapsed

on stream  $b$  through one  $e$ -fold of blueshift to be

$$u_{ba}^t \frac{d\tau_a}{d \ln u_{ba}^t} = \frac{2C_b r_-}{\lambda} . \quad (21.2)$$

The right hand side of equation (21.2) is derived from  $u_{ba}^t \approx 2|u_a u_b|$ ,  $|dr_a/d\tau_a| = |\beta_{a,t}| \approx \beta u_a$ ,  $r_a \approx r_-$ , and the approximations (??) and (??) valid during the inflationary phase. The constants  $C_b$  and  $\lambda$  on the right hand side of equation (21.2) are typically of order unity, while  $r_-$  is the radius of the inner horizon where mass inflation takes place. Thus the right hand side of equation (21.2) is of the order of one black hole crossing time. In other words, stream  $a$  sees approximately one black hole crossing time elapse on stream  $b$  for each  $e$ -fold of blueshift.

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 21.10. 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 ingoing or outgoing streams is tiny compared to the other, as considered in §§??.

From a stream's own point of view, the entire inflationary episode goes by in a flash. At the onset of inflation, where  $\beta$  goes through its minimum, at  $\mu_a u_a^2 = \mu_b u_b^2 = \lambda$  according to equation (??), the blueshift  $u_{ba}^t$  is already large

$$u_{ba}^t = \frac{2\lambda}{\sqrt{\mu_a \mu_b}} , \quad (21.3)$$

thanks to the small accretion rates  $\mu_a$  and  $\mu_b$ . During the first  $e$ -fold of blueshift, each stream experiences a proper time of order  $\sqrt{\mu_a \mu_b}$  times the black hole crossing time, which is tiny. Subsequent  $e$ -folds of blueshift race by in proportionately shorter proper times.

### 21.2.2 Collapse phase

The time to reach the collapse phase is another matter. According to the estimates in §§?? and ??, reaching the collapse phase takes of order  $\sim 1/\mu_a$   $e$ -folds of blueshift, where  $\mu_a$  is the larger of the accretion rates of the ingoing and outgoing streams, equation (??). Thus in reaching the collapse phase, each stream has seen approximately  $1/\mu_a$  black hole crossing times elapse on the other stream. But  $1/\mu_a$  black hole crossing times is just the accretion time — essentially, how long the black hole has existed. This timescale, the age of the black hole, is not infinite, but it can hardly be expected that the accretion rate of a black hole would be constant over its lifetime.

If the accretion rate were in fact constant, and if quantum gravity did not intervene and the streams remained non-interacting, then indeed streams inside the black hole would reach the collapse phase, whereupon they would plunge to a spacelike singularity at zero radius. For example, in the self-similar models illustrated in Figures ?? and 21.10, outgoing baryons hitting the central singularity see ingoing dark matter accreted a factor of two into the future (specifically, for ingoing baryons and outgoing dark matter hitting the singularity, the radius of the outer horizon when the dark matter is accreted is twice that when the baryons

were accreted; the numbers are 2.11 for  $\dot{M}_\bullet = 0.03$ , 1.97 for  $\dot{M}_\bullet = 0.01$ , and unknown for  $\dot{M}_\bullet = 0.003$  because in that case the numerics overflow before the central singularity is reached). The same conclusion applies to the ultra-hard fluid models illustrated in Figure 21.8: if the accretion rate is constant, then outgoing streams see only a factor of order unity or a few into the future before hitting the central singularity.

If on the other hand the accretion rate decreases sufficiently rapidly with time, then it is possible that an outgoing stream never reaches the collapse phase, because the number of e-folds to reach the collapse phase just keeps increasing as the accretion rate decreases. By contrast, an ingoing stream is always liable to reach the collapse phase, if quantum gravity does not intervene, because an ingoing stream sees the outgoing stream from the past, when the accretion rate was liable to have been larger.

However, as already remarked in §??, in astronomically realistic black holes, it is only for large accretion rates, such as may occur when the black hole first collapses ( $\dot{M}_\bullet \gtrsim 0.01$  for the models illustrated in Fig. 21.10), that the collapse phase has a chance of being reached before the Weyl curvature exceeds the Planck scale.

To summarize, it is only streams accreted during the first few hundred or so black hole crossing times since a black hole's formation that have a chance of hitting a central spacelike singularity. Streams accreted at later times, whether ingoing or outgoing, are likely to meet their fate in the inflationary zone at the inner horizon, where the Weyl curvature exponentiates to the Planck scale and beyond.

## 21.3 Self-similar models of the interior structure of black holes

The apparatus is now in hand actually to do some real calculations of the interior structure of black holes. All the models presented in this section are spherical and self-similar. See Hamilton & Pollack (2005, PRD 71, 084031 & 084031) and Wallace, Hamilton & Polhemus (2008, arXiv:0801.4415) for more.

### 21.3.1 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 the boundary conditions outside the horizon. The policy adopted here is to set them at a sonic point, where the infalling fluid accelerates from subsonic to supersonic. The proper 3-velocity of the fluid through the self-similar frame is  $\xi^1/\xi^0$ , equation (20.100) (the velocity  $\xi^1/\xi^0$  is positive falling inward), and the sound speed is

$$\text{sound speed} = \sqrt{\frac{p_b}{\rho_b}} = \sqrt{w_b} , \quad (21.4)$$

and sonic points occur where the velocity equals the sound speed

$$\frac{\xi^1}{\xi^0} = \pm \sqrt{w_b} \quad \text{at sonic points .} \quad (21.5)$$

The denominator of the expression (20.162) for the proper acceleration  $g$  is zero at sonic points, indicating that the acceleration will diverge unless the numerator is also zero. 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 (20.162) 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  $g$  at the sonic point is not only continuous [so the numerator of (20.162) 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.

### 21.3.2 Mass and charge of the black hole

The mass  $M_\bullet$  and charge  $Q_\bullet$  of the black hole at any instant can be defined to be those that would be measured by a distant observer if there were no mass or charge outside the sonic point

$$M_\bullet = M + \frac{Q^2}{2r} , \quad Q_\bullet = Q \quad \text{at the sonic point .} \quad (21.6)$$

The mass  $M_\bullet$  in equation (21.6) 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 increases linearly with time,  $M_\bullet \propto t$ , where  $t$  is the proper time at rest far from the black hole. As discussed in §??, this time  $t$  equals the proper time  $\tau_d = r\xi_d^0/v$  recorded by dark matter clocks that free-fall radially from zero velocity at infinity. Thus the mass accretion rate  $\dot{M}_\bullet$  is

$$\dot{M}_\bullet \equiv \frac{dM_\bullet}{dt} = \frac{M_\bullet}{\tau_d} = \frac{vM_\bullet}{r\xi_d^0} \quad \text{at the sonic point .} \quad (21.7)$$

If there is no mass outside the sonic point (apart from the mass-energy in the electric field), then a freely-falling dark matter particle will have

$$\beta_{d,x} = 1 \quad \text{at the sonic point ,} \quad (21.8)$$

which can be taken as the boundary condition on the dark matter at the sonic point, for either massive or massless dark matter. Equation (21.8) follows from the facts that the geodesic equations in empty space around a charged black hole (Reissner-Nordström metric) imply that  $\beta_{d,x} = \text{constant}$  for a radially free-falling particle (the same conclusion can drawn from the Einstein equation (20.33c)), and that a particle at rest at infinity satisfies  $\beta_{d,\eta} = \partial_{d,\eta}r = 0$ , and consequently  $\beta_{d,x} = 1$  from equation (20.91) with  $r \rightarrow \infty$ .

As remarked following equation (20.94), the residual gauge freedom in the global rescaling of conformal

time  $\eta$  allows the expansion velocity  $v$  to be adjusted at will. One choice suggested by equation (21.7) is to set (but one could equally well set  $v$  to the expansion velocity of the horizon,  $v = \dot{r}_+$ , for example)

$$v = \dot{M}_\bullet , \quad (21.9)$$

which is equivalent to setting

$$\xi_d^0 = \frac{M_\bullet}{r} \quad \text{at the sonic point .} \quad (21.10)$$

Equation (21.10) is not a boundary condition: it is just a choice of units of conformal time  $\eta$ . Equation (21.10) and the boundary condition (21.8) coupled with the scalar relations (20.91) and (20.92) fully determine the dark matter 4-vectors  $\beta_{d,m}$  and  $\xi_d^m$  at the sonic point.

### 21.3.3 Equation of state

The density  $\rho_b$  and temperature  $T_b$  of an ideal relativistic baryonic fluid in thermodynamic equilibrium are related by

$$\rho_b = \frac{\pi^2 g}{30} T_b^4 , \quad (21.11)$$

where

$$g = g_B + \frac{7}{8} g_F \quad (21.12)$$

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  $g/g_P = T^\epsilon$ , with  $g_P$  the effective number of relativistic species at the Planck temperature, then the relation between density  $\rho_b$  and temperature  $T_b$  is

$$\rho_b = \frac{\pi^2 g_P}{30} T_b^{(1+w)/w} , \quad (21.13)$$

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

$$w_b = 0.32 . \quad (21.14)$$

The effective number  $g_p$  is fixed by setting the number of relativistic particles species to  $g = 5.5$  at  $T = 10$  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 not unreasonable.

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

$$S_b = \frac{(\rho_b + p_b)V}{T_b} , \quad (21.15)$$

which agrees with the earlier expression (20.164), but now has the correct normalization.

### 21.3.4 Entropy creation

One fundamentally interesting question about black hole interiors is how much entropy might be created inside the horizon. Bekenstein 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  $S_{\text{BH}}$  is, in Planck units  $c = G = \hbar = 1$ ,

$$S_{\text{BH}} = \frac{A}{4} . \quad (21.16)$$

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  $T_{\text{H}}$  equal to  $1/(2\pi)$  times the surface gravity  $g_+$  at its horizon, again in Planck units,

$$T_{\text{H}} = \frac{g_+}{2\pi} . \quad (21.17)$$

The surface gravity is defined to be the proper radial acceleration measured by a person in free-fall at the horizon. For a spherical black hole, the surface gravity is  $g_+ = -D_0\beta_0 = M/r^2 + 4\pi r p$  evaluated at the horizon, equation (20.59a).

The proper velocity of the baryonic fluid through the sonic point equals  $\xi^1/\xi^0$ , equation (20.100). Thus the entropy  $S_b$  accreted through the sonic point per unit proper time of the fluid is

$$\frac{dS_b}{d\tau} = \frac{(1+w_b)\rho_b}{T_b} \frac{4\pi r^2 \xi^1}{\xi^0} . \quad (21.18)$$

The horizon radius  $r_+$ , which is at fixed conformal radius  $x$ , expands in proportion to the conformal factor,  $r_+ \propto a$ , and the conformal factor  $a$  increases as  $d\ln a/d\tau = \partial_0 \ln a = v/(r\xi^0)$ , so the Bekenstein-Hawking entropy  $S_{\text{BH}} = \pi r_+^2$  increases as

$$\frac{dS_{\text{BH}}}{d\tau} = \frac{2\pi r_+^2 v}{r\xi^0} . \quad (21.19)$$

Thus the entropy  $S_b$  accreted through the sonic point per unit increase of the Bekenstein-Hawking entropy  $S_{\text{BH}}$  is

$$\frac{dS_b}{dS_{\text{BH}}} = \left. \frac{(1+w_b)\rho_b 4\pi r^3 \xi^1}{2\pi r_+^2 v T_b} \right|_{r=r_s} . \quad (21.20)$$

Inside the sonic point, dissipation increases the entropy according to equation (20.163). Since the entropy can diverge at a central singularity where the density diverges, and quantum gravity presumably intervenes at some point, it makes sense to truncate the production of entropy at a “splat” point where the density  $\rho_b$  hits a prescribed splat density  $\rho_{\#}$

$$\rho_b = \rho_{\#} . \quad (21.21)$$

Integrating equation (20.163) from the sonic point to the splat point yields the ratio of the entropies at the

sonic and splat points. Multiplying the accreted entropy, equation (21.20), by this ratio yields the rate of increase of the entropy of the black hole, truncated at the splat point, per unit increase of its Bekenstein-Hawking entropy

$$\frac{dS_b}{dS_{\text{BH}}} = \frac{(1+w_b)\rho_b 4\pi r^3 \xi^1}{2\pi r_+^2 v T_b} \Big|_{\rho_b=\rho_\#} . \quad (21.22)$$

### 21.3.5 Holography

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 **covariant entropy bound** (Bousso 2002, Rev. Mod. Phys. 74, 825). 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.

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 ingoing or outgoing spherical lightsheets per unit decrease in the area  $S_{\text{cov}} \equiv \pi r^2$  of the lightsheet is

$$\left| \frac{dS_b}{dS_{\text{cov}}} \right| = \frac{dS_b}{dS_{\text{BH}}} \frac{r_+^2}{r^2} \frac{v}{\xi^1 |\beta_0 \mp \beta_1|} , \quad (21.23)$$

in which the  $\mp$  sign is  $-$  for ingoing,  $+$  for outgoing lightsheets. A sufficient condition for Bousso's covariant entropy bound to be satisfied is

$$|dS_b/dS_{\text{cov}}| \leq 1 . \quad (21.24)$$

### 21.3.6 Black hole accreting a neutral relativistic plasma

The simplest case to consider is that of a black hole accreting a neutral relativistic plasma. In the 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 21.2 shows the baryonic density  $\rho_b$  and Weyl curvature  $C$  inside the uncharged black hole. The mass and accretion rate have been taken to be

$$M_\bullet = 4 \times 10^6 M_\odot , \quad \dot{M}_\bullet = 10^{-16} , \quad (21.25)$$

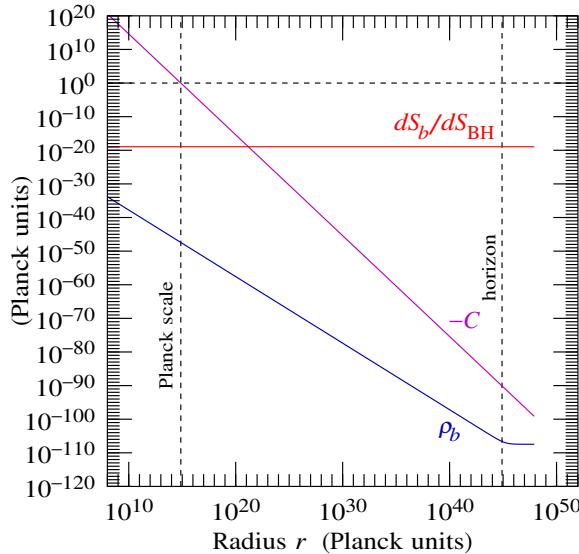


Figure 21.2 An uncharged baryonic plasma falls into an uncharged spherical black hole. The plot shows in Planck units, as a function of radius, the plasma density  $\rho_b$ , the Weyl curvature scalar  $C$  (which is negative), and the rate  $dS_b/dS_{\text{BH}}$  of increase of the plasma entropy per unit increase in the Bekenstein-Hawking entropy of the black hole. 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  $w_b = 0.32$ .

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

$$\frac{\text{Mass of MW black hole}}{\text{age of Universe}} \approx \frac{4 \times 10^6 M_\odot}{10^{10} \text{ yr}} \approx \frac{6 \times 10^{60} \text{ Planck units}}{4 \times 10^{44} \text{ Planck units}} \approx 10^{-16}. \quad (21.26)$$

Figure 21.2 shows that the baryonic plasma plunges uneventfully to a central singularity, just as in the Schwarzschild solution. The Weyl curvature scalar hits the Planck scale,  $|C| = 1$ , while the baryonic proper density  $\rho_b$  is still well below the Planck density, so this singularity is curvature-dominated.

### 21.3.7 Black hole accreting a non-conducting charged relativistic plasma

The next simplest case is that of a black hole accreting a charged but non-conducting relativistic plasma.

Figure 21.3 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 §21.3.6:  $M_\bullet = 4 \times 10^6 M_\odot$ ,  $\dot{M}_\bullet = 10^{-16}$ , 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

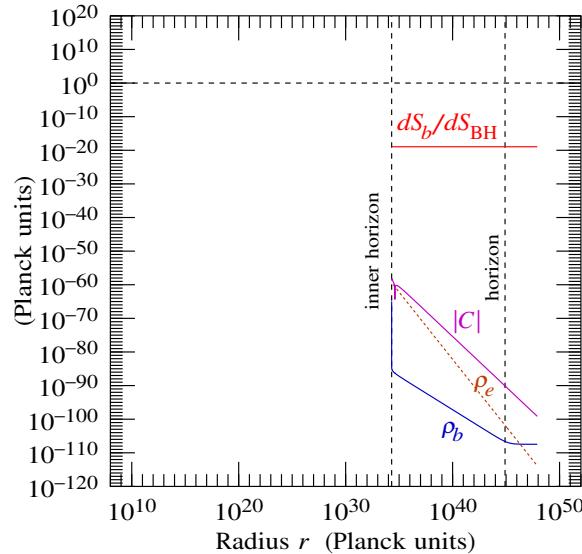


Figure 21.3 A plasma that is charged but non-conducting. The black hole has an inner horizon like the Reissner-Nordström geometry. The self-similar solution terminates at an irregular sonic point just beneath the inner horizon. The mass is  $M_\bullet = 4 \times 10^6 M_\odot$ , accretion rate  $\dot{M}_\bullet = 10^{-16}$ , equation of state  $w_b = 0.32$ , and black hole charge-to-mass  $Q_\bullet/M_\bullet = 10^{-5}$ .

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 to continue does not invalidate the solution, because the failure is hidden beneath the inner horizon, and cannot be communicated to infalling matter above it.

The solution is nevertheless not realistic, because it assumes that there is no ingoing matter, such as would inevitably be produced for example by infalling neutral dark matter. Such ingoing matter would appear infinitely blueshifted to the outgoing baryons falling through the inner horizon, which would produce mass inflation, as in §21.3.10.

### 21.3.8 Black hole accreting a conducting relativistic plasma

What happens if the baryonic plasma is not only electrically charged but also electrically conducting? If the conductivity is small, then the solutions resemble the non-conducting solutions of the previous subsection, §21.3.7. 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 §21.3.6.

Figure 21.4 shows a case in which the conductivity has been tuned to equal, within numerical accuracy, the critical conductivity  $\kappa_b = 0.35$  above which the plasma collapses to a central singularity. The parameters

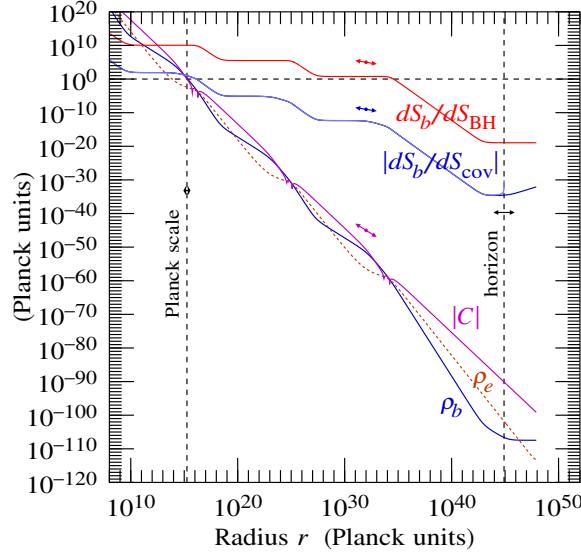


Figure 21.4 Here the baryonic plasma is charged, and electrically conducting. The conductivity is at (within numerical accuracy) the threshold above which the plasma plunges to a central singularity. The mass is  $M_\bullet = 4 \times 10^6 M_\odot$ , the accretion rate  $\dot{M}_\bullet = 10^{-16}$ , the equation of state  $w_b = 0.32$ , the charge-to-mass  $Q_\bullet/M_\bullet = 10^{-5}$ , and the conductivity parameter  $\kappa_b = 0.35$ . Arrows show how quantities vary a factor of 10 into the past and future.

are otherwise the same as in previous subsections: 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 solution at the critical conductivity exhibits the periodic self-similar behaviour first discovered in numerical simulations by Choptuik (1993, PRL 70, 9), 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 21.4 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 therefore 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. My instinct is that angular momentum transport is probably not strong enough, but I do not know this for sure. If angular momentum transport is not strong enough, then mass inflation will take place.

Figure 21.4 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, as a function of the hypothetical splat point above which entropy production is truncated. The rate is almost independent of the black hole mass  $M_\bullet$  at fixed splat density  $\rho_\#$ , so it is legitimate to interpret  $dS_b/dS_{BH}$  as the cumulative entropy created inside the black hole relative to the Bekenstein-Hawking entropy. Truncated at the Planck scale,  $|C| = 1$ , the entropy relative to Bekenstein-Hawking is  $dS_b/dS_{BH} \approx 10^{10}$ .

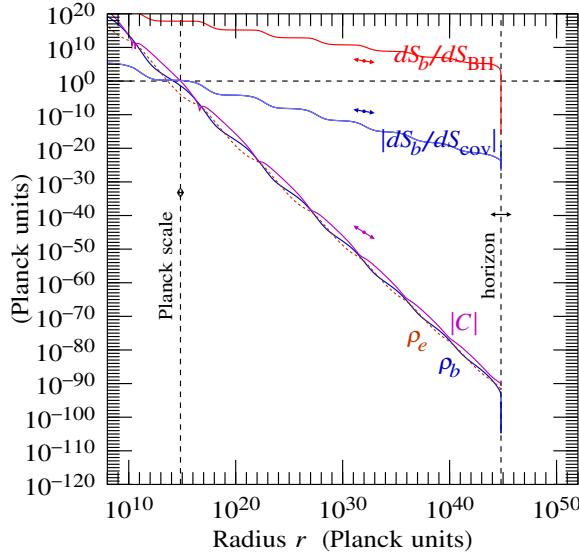


Figure 21.5 This black hole creates a lot of entropy by having a large charge-to-mass  $Q_\bullet/M_\bullet = 0.8$  and a low accretion rate  $\dot{M}_\bullet = 10^{-28}$ . The conductivity parameter  $\kappa_b = 0.35$  is again at the threshold above which the plasma plunges to a central singularity. The equation of state is  $w_b = 0.32$ .

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 21.5 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 leads to a burst of entropy production just beneath the horizon.

If the entropy created inside a black hole exceeds the Bekenstein-Hawking entropy, and the black hole later evaporates radiating only the Bekenstein-Hawking entropy, then entropy is destroyed, violating the second law of thermodynamics.

This startling conclusion is premised on the assumption that entropy created inside a black hole accumulates additively, which in turn derives from the assumption that the Hilbert space of states is multiplicative over spacelike-separated regions. This assumption, called **locality**, derives from the fundamental proposition of quantum field theory in flat space that field operators at spacelike-separated points commute. This reasoning is essentially the same as originally led Hawking (1976) to conclude that black holes must destroy information.

The same ideas that motivate holography also rescue the second law. If the future lightcones of spacelike-separated points do not intersect, then the points are permanently out of communication, and can behave like alternate quantum realities, like Schrödinger's dead-and-alive quantum cat. Just as it is not legitimate to add the entropies of the dead cat and the live cat, so also it is not legitimate to add the entropies

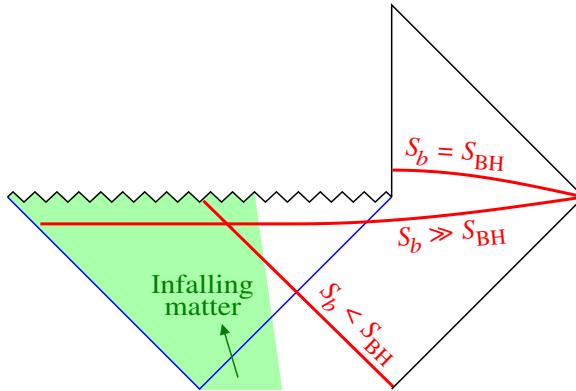


Figure 21.6 Partial Penrose diagram of the black hole. The entropy passing through the spacelike slice before the black hole evaporates exceeds that passing through the spacelike slice after the black hole evaporates, apparently violating the second law of thermodynamics. However, the entropy passing through any null slice respects the second law.

of regions inside a black hole whose future lightcones do not intersect. The states of such separated regions, instead of being distinct, are quantum entangled with each other.

Figures 21.4 and 21.5 show that the rate  $|dS_b/dS_{\text{cov}}|$  at which entropy passes through ingoing or outgoing 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. No observer actually witnesses a violation of the second law.

### 21.3.9 Black hole accreting a charged massless scalar field

The charged, non-conducting plasma considered in §21.3.7 fell through an (outgoing) inner horizon without undergoing mass inflation. This can be attributed to the fact that relativistic counter-streaming could not occur: there was only a single (outgoing) fluid, and the speed of sound in the fluid was less than the speed of light.

In reality, unless dissipation destroys the inner horizon as in §21.3.8, then relativistic counter-streaming between ingoing and outgoing fluids will undoubtedly take place, through gravitational waves if nothing else.

One way to allow relativistic counter-streaming is to let the speed of sound be the speed of light. This is true in a massless scalar (= spin-0) field  $\phi$ , which has an equation of state  $w_\phi = 1$ . Figure 21.7 shows a black hole that accretes a charged, non-conducting fluid with this equation of state. The parameters are otherwise the same as in previous subsections: 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  $\rho_\phi$  and the Weyl scalar  $C$  rapidly exponentiate up to the Planck scale and beyond.

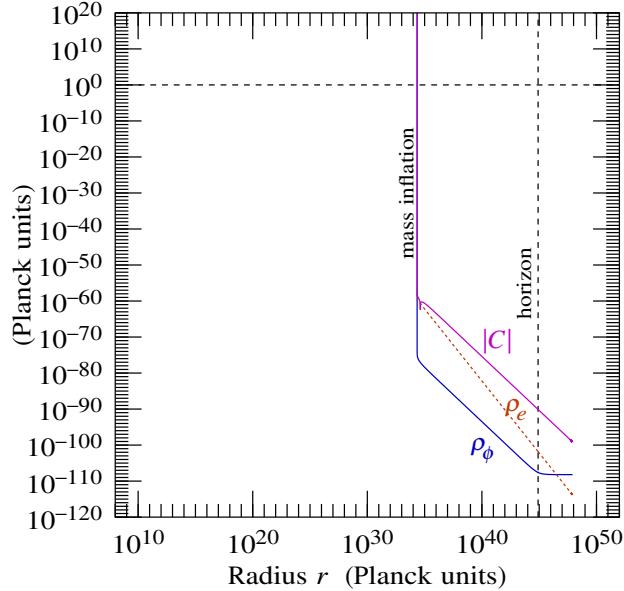


Figure 21.7 Instead of a relativistic plasma, this shows a charged scalar field  $\phi$  whose equation of state  $w_\phi = 1$  means that the speed of sound equals the speed of light. The scalar field therefore supports relativistic counter-streaming, as a result of which mass inflation occurs just 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.

One of the remarkable features of the mass inflation instability is that the smaller the accretion rate, the more violent the instability. Figure 21.8 shows mass inflation in a black hole of charge-to-mass  $Q_\bullet/M_\bullet = 0.8$  accreting a massless scalar field at rates  $\dot{M}_\bullet = 0.01, 0.003$ , and  $0.001$ . The charge-to-mass has been chosen to be largish so that the inner horizon is not too far below the outer horizon, and the accretion rates have been chosen to be large because otherwise the inflationary growth rate is too rapid to be discerned easily on the graph. The density  $\rho_\phi$  and Weyl scalar  $C$  exponentiate along with, and in proportion to, the interior mass  $M$ , which increases as the radius  $r$  decreases as

$$M \propto \exp(-\ln r/\dot{M}_\bullet) . \quad (21.27)$$

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, the closer the approach, and consequently the shorter the length scale of inflation.

### 21.3.10 Black hole accreting charged baryons and dark matter

No scalar field (massless or otherwise) has yet been observed in nature, although it is supposed that the Higgs field is a scalar field, and it is likely that cosmological inflation was driven by a scalar field. Another

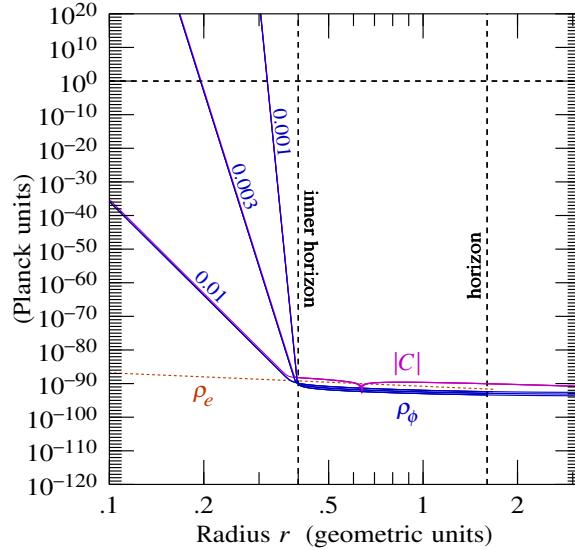


Figure 21.8 The density  $\rho_\phi$  and Weyl curvature scalar  $|C|$  inside a black hole accreting a massless scalar field. The graph shows three cases, with mass accretion rates  $\dot{M}_\bullet = 0.01, 0.003$ , and  $0.001$ . The graph illustrates that the smaller the accretion rate, the faster the density and curvature inflate. Mass inflation destroys the inner horizon: the dashed vertical line labelled “inner horizon” shows the position that the inner horizon would have if mass inflation did not occur. The black hole mass is  $M_\bullet = 4 \times 10^6 M_\odot$ , the charge-to-mass is  $Q_\bullet/M_\bullet = 0.8$ , and the conductivity is zero.

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 baryons. The charged baryons, being repelled by the electric charge of the black hole, become outgoing, while the neutral dark matter remains ingoing.

Figure 21.9 shows that relativistic counter-streaming between the baryons and the dark matter causes the centre-of-mass density  $\rho$  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 previous subsections:  $M_\bullet = 4 \times 10^6 M_\odot$ ,  $\dot{M}_\bullet = 10^{-16}$ ,  $w_b = 0.32$ ,  $Q_\bullet/M_\bullet = 10^{-5}$ , and zero conductivity. Almost all the centre-of-mass energy  $\rho$  is in the counter-streaming energy between the outgoing baryonic and ingoing dark matter. The individual densities  $\rho_b$  of baryons and  $\rho_d$  of dark matter (and  $\rho_e$  of electromagnetic energy) increase only modestly.

As in the case of the massless scalar field considered in the previous subsection, §21.3.9, the smaller the accretion rate, the shorter the length scale of inflation. Not only that, but the smaller one of the ingoing or outgoing streams is relative to the other, the shorter the length scale of inflation. Figure 21.10 shows a black hole with three different ratios of the dark-matter-to-baryon density ratio at the sonic point,  $\rho_d/\rho_b = 0.3$ ,  $0.1$ , and  $0.03$ , all with the same total accretion rate  $\dot{M}_\bullet = 10^{-2}$ . The smaller the dark matter stream, the

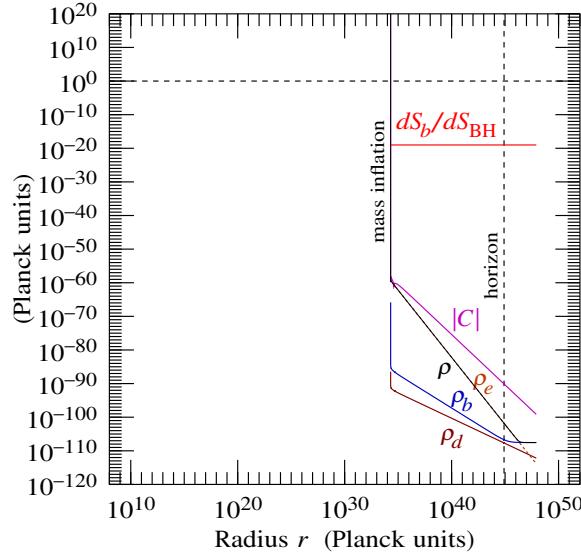


Figure 21.9 Back to the relativistic charged baryonic plasma, but now in addition the black hole accretes neutral pressureless uncharged dark matter, which streams freely through the baryonic plasma. The relativistic counter-streaming produces mass inflation just above the erstwhile inner horizon. The mass is  $M_\bullet = 4 \times 10^6 M_\odot$ , the accretion rate  $\dot{M}_\bullet = 10^{-16}$ , the 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  $\rho_d/\rho_b = 0.1$ .

faster is inflation. The accretion rate  $\dot{M}_\bullet$  and the dark-matter-to-baryon ratio  $\rho_d/\rho_b$  have been chosen to be relatively large so that the inflationary growth rate is discernible easily on the graph.

Figure 21.10 shows that, as in Figure 21.9, almost all the centre-of-mass energy is in the streaming energy between the baryons and the dark matter. For one case,  $\rho_d/\rho_b = 0.3$ , Figure 21.10 shows the individual densities  $\rho_b$  of baryons,  $\rho_d$  of dark matter, and  $\rho_e$  of electromagnetic energy, all of which remain tiny compared to the streaming energy.

### 21.3.11 The black hole particle accelerator

The previous subsection, §21.3.10, 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.

Mass inflation is an exponential instability. The nature of the black hole particle accelerator is that an individual particle spends approximately an equal interval of proper time being accelerated through each decade of collision energy.

Each baryon in the black hole collider sees a flux  $n_d u^r$  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  $u^r$  is the radial component of the proper 4-velocity, the  $\gamma v$ , of the dark matter through the baryons. The  $\gamma$  factor

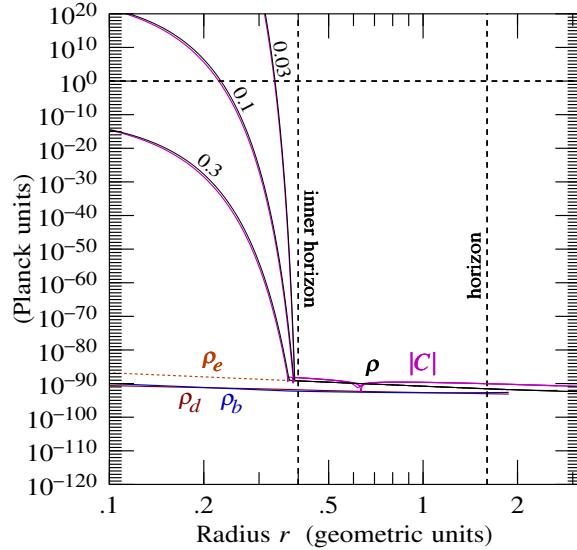


Figure 21.10 The centre-of-mass density  $\rho$  and Weyl curvature  $|C|$  inside a black hole accreting baryons and dark matter at rate  $\dot{M}_\bullet = 0.01$ . The graph shows three cases, with dark-matter-to-baryon ratio at the sonic point of  $\rho_d/\rho_b = 0.3, 0.1$ , and  $0.03$ . The smaller the ratio of dark matter to baryons, the faster the centre-of-mass density  $\rho$  and curvature  $C$  inflate. For the largest ratio,  $\rho_d/\rho_b = 0.3$  (to avoid confusion, only this case is plotted), the graph also shows the individual proper densities  $\rho_b$  of baryons,  $\rho_d$  of dark matter, and  $\rho_e$  of electromagnetic energy. During mass inflation, almost all the centre-of-mass energy  $\rho$  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.

in  $u^r$  is the relativistic beaming factor: all frequencies, including the collision frequency, are speeded up by the relativistic beaming factor  $\gamma$ . As the baryons accelerate through the collider, they spend a proper time interval  $d\tau/d\ln u^r$  in each  $e$ -fold of Lorentz factor  $u^r$ . The number of collisions per baryon per  $e$ -fold of  $u^r$  is the dark matter flux  $(\rho_d/m_d)u^r$ , multiplied by the time  $d\tau/d\ln u^r$ , multiplied by the collision cross-section  $\sigma$ . 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

$$\frac{\text{number of collisions}}{\text{e-fold of } u^r} = \frac{M_\bullet}{m_b m_d} \frac{\rho_d}{\sigma} u^r \frac{d\tau}{d\ln u^r}. \quad (21.28)$$

Figure 21.11 shows, for several different accretion rates  $\dot{M}_\bullet$ , the collision rate  $M_\bullet \rho_d u^r d\tau/d\ln u^r$  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 (21.28) 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  $\sigma$  is expressed in canonical

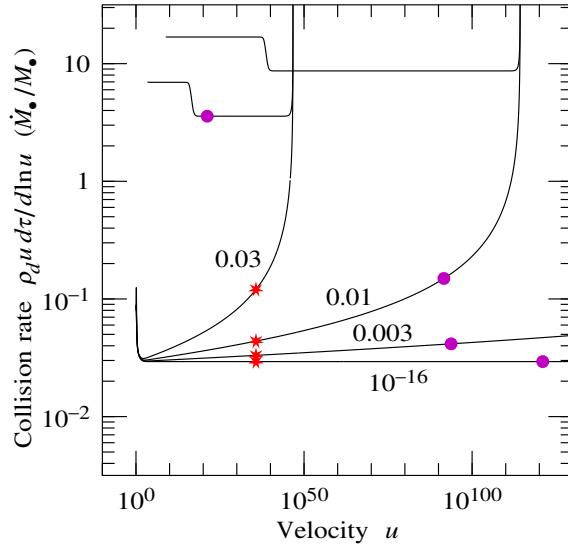


Figure 21.11 Collision rate of the black hole particle accelerator per  $e$ -fold of velocity  $u$  (meaning  $\gamma v$ ), expressed in units of the inverse black hole accretion time  $\dot{M}_\bullet/M_\bullet$ . The models illustrated are the same as those in Figure 21.10. The curves are labelled with their mass accretion rates:  $\dot{M}_\bullet = 0.03, 0.01, 0.003$ , and  $10^{-16}$ . 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.

accelerator units of femtobarns ( $1 \text{ fb} = 10^{-43} \text{ m}^2$ ) then the number of collisions (21.28) is

$$\frac{\text{number of collisions}}{e\text{-fold of } u^r} = 10^{45} \left( \frac{\sigma}{1 \text{ fb}} \right) \left( \frac{300 \text{ GeV}^2}{m_b m_d} \right) \left( \frac{\dot{M}_\bullet}{10^{-16}} \right) \left( \frac{M_\bullet \rho_d u^r d\tau/d\ln u^r}{0.03 \dot{M}_\bullet} \right). \quad (21.29)$$

Particle accelerators measure their cumulative luminosities in inverse femtobarns. Equation (21.29) shows that the black hole accelerator delivers about  $10^{45}$  femtobarns, and it does so in each  $e$ -fold of collision energy up to the Planck energy and beyond.

## 21.4 Instability at 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, is 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 otherwise, no.

# 22

---

## Ideal rotating black holes

Among the remarkable mathematical properties of the Kerr-Newman line-element is the fact that, as first shown by Carter (1968b), 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  $\Lambda$ -Kerr-Newman black holes, which are ideal charged rotating black holes in a background with a cosmological constant  $\Lambda$ .

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,  $\phi$  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  $\rho$  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  $\phi$ , the metric coefficients  $g_{tt}$ ,  $g_{t\phi}$ , and  $g_{\phi\phi}$  all have a gauge-invariant significance. The determinant of the  $2 \times 2$  submatrix of  $t$ - $\phi$  coefficients is

$$g_{tt} g_{\phi\phi} - g_{t\phi}^2 = -\frac{\rho^4}{(1 - \omega_x \omega_y)^2} \Delta_x \Delta_y. \quad (22.4)$$

The quantity  $\Delta_x$  is the horizon function. Horizons occur where the horizon function vanishes  $\Delta_x$  vanishes. The quantity  $\Delta_y$  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  $\Delta_x$  has opposite sign, trajectories cannot pass through  $\Delta_y = 0$  into a region where  $\Delta_y$  has opposite sign. Without loss of generality, the polar function  $\Delta_y$  can be taken to be positive, since the line-element (22.1) with both  $\Delta_x$  and  $\Delta_y$  flipped in sign describes the same geometry with flipped signature.

### 22.1.2 $\Lambda$ -Kerr-Newman

As shown in §22.6, the  $\Lambda$ -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 non-singular at the north and south poles,  $\theta = 0$  and  $\pi$ . For  $\Lambda$ -Kerr-Newman, the radial and angular parts  $\rho_x$  and  $\rho_y$  of the separable conformal factor are

$$\rho_x \equiv r = a \cot(ax), \quad \rho_y \equiv a \cos \theta = -ay, \quad (22.5)$$

where  $r$  is the ellipsoidal radial coordinate and  $\theta$  the polar angle, as conventionally defined, and  $a$  is the spin parameter of the black hole. Why use coordinates  $x$  and  $y$  in place of  $r$  and  $\theta$ ? 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

$$\frac{\partial}{\partial x} = -R^2 \frac{\partial}{\partial r}, \quad R \equiv \sqrt{r^2 + a^2}. \quad (22.6)$$

For  $\Lambda$ -Kerr-Newman, the coefficients  $\omega_x$  and  $\omega_y$  in the line-element (22.1) are

$$\omega_x = \frac{a}{R^2}, \quad \omega_y = a \sin^2 \theta, \quad (22.7)$$

and the horizon and polar functions  $\Delta_x$  and  $\Delta_y$  are (the horizon function  $\Delta_x$  here is related to the earlier horizon function  $\Delta$ , equation (9.3), by  $\Delta_x = R^{-2}\Delta$ )

$$\Delta_x = \frac{1}{R^2} \left( 1 - \frac{2M_\bullet r}{R^2} + \frac{Q_\bullet^2 + Q_\bullet^2}{R^2} - \frac{\Lambda r^2}{3} \right) , \quad (22.8a)$$

$$\Delta_y = \sin^2\theta \left( 1 + \frac{\Lambda a^2 \cos^2\theta}{3} \right) , \quad (22.8b)$$

where  $M_\bullet$  is the black hole's mass,  $Q_\bullet$  and  $Q_\bullet$  are its electric and magnetic charge, and  $\Lambda$  is the cosmological constant. By themselves, Maxwell's equations preclude magnetic charge, in which case  $Q_\bullet = 0$ . 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  $A_k$  is

$$A_k = \frac{1}{\rho} \left\{ -\frac{Q_\bullet r}{R^2 \sqrt{\Delta_x}}, 0, 0, -\frac{Q_\bullet \cos\theta}{\sqrt{\Delta_y}} \right\} , \quad (22.9)$$

and the radial electric and magnetic fields  $E$  and  $B$  are given by

$$E + IB \equiv F_{10} + IF_{23} = \frac{Q_\bullet + I Q_\bullet}{(\rho_x - I \rho_y)^2} , \quad (22.10)$$

where  $I$  is the pseudoscalar of the spacetime algebra, satisfying  $I^2 = -1$ . The Weyl tensor (??) has only a spin-0 component, and is

$$C = -\frac{1}{(\rho_x - I \rho_y)^3} \left( M_\bullet + I N_\bullet - \frac{Q_\bullet^2 + Q_\bullet^2}{\rho_x + I \rho_y} \right) . \quad (22.11)$$

## 22.2 Horizons

Horizons occur where the horizon function  $\Delta_x$  vanishes,

$$\Delta_x = 0 . \quad (22.12)$$

For Kerr-Newman with vanishing cosmological constant, there are outer and inner horizons  $r_\pm$  at

$$r_\pm = M_\bullet \pm \sqrt{M_\bullet^2 - a^2 - Q_\bullet^2 - Q_\bullet^2} . \quad (22.13)$$

If there is a non-zero cosmological constant  $\Lambda$ , then the horizon condition (22.12) 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.

### 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),

$$\eta^{mn} \left( e_m^{\mu} \frac{\partial S}{\partial x^{\mu}} - qA_m \right) \left( e_n^{\nu} \frac{\partial S}{\partial x^{\nu}} - qA_n \right) = -m^2. \quad (22.14)$$

The Hamilton-Jacobi equation (22.14) is a partial differential equation in the particle action  $S$ , equation (4.36). Let  $\hat{e}_m^{\mu}$  and  $\hat{A}_m$  denote the vierbein coefficients and tetrad-frame electromagnetic potential with an overall conformal factor  $\rho$  factored out:

$$\hat{e}_m^{\mu} \equiv \rho e_m^{\mu}, \quad \hat{A}_m \equiv \rho A_m. \quad (22.15)$$

With respect to the scaled vierbein  $\hat{e}_m^{\mu}$  and electromagnetic potential  $\hat{A}_m$ , the Hamilton-Jacobi equation (22.14) can be rewritten

$$\eta^{mn} \left( \hat{e}_m^{\mu} \frac{\partial S}{\partial x^{\mu}} - q\hat{A}_m \right) \left( \hat{e}_n^{\nu} \frac{\partial S}{\partial x^{\nu}} - q\hat{A}_n \right) = -m^2\rho^2. \quad (22.16)$$

To separate the Hamilton-Jacobi equation (22.16), 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 (22.16) is to impose that each of the individual factors, comprising the scaled 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 non-vanishing 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  $S_{\mu}$  each depending on a single coordinate  $x^{\mu}$ ,

$$S = \sum_{\mu} S_{\mu}(x^{\mu}). \quad (22.17)$$

Canonical momenta are equal to derivatives of the action, equation (4.105), so the condition (22.17) imposes that each canonical momentum  $\pi_{\mu}$  be a function only of the corresponding coordinate  $x^{\mu}$ ,

$$\pi_{\mu} = \frac{\partial S_{\mu}}{\partial x^{\mu}} = \text{function of } x^{\mu}. \quad (22.18)$$

A special case of the condition (22.18) occurs when a canonical momentum  $\pi_{\mu}$  is a constant, which occurs when the metric is independent of the coordinate  $x^{\mu}$ , 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  $\phi$ . The corresponding canonical momenta  $\pi_t$  and  $\pi_\phi$  are constants of motion, defining respectively the constant energy  $E$  and the azimuthal angular momentum  $L$  of the trajectory,

$$\pi_t = \frac{\partial S}{\partial t} = -E , \quad \pi_\phi = \frac{\partial S}{\partial \phi} = L . \quad (22.19)$$

If the two remaining coordinates are denoted  $x$  and  $y$ , then the particle action  $S$ , equation (22.17), is the sum

$$S = -Et + L\phi + S_x(x) + S_y(y) , \quad (22.20)$$

where  $S_x(x)$  and  $S_y(y)$  are respectively functions only of  $x$  and  $y$ .

Given that  $\pi_t$  and  $\pi_\phi$  are constants, while  $\pi_x$  and  $\pi_y$  are respectively functions of  $x$  and  $y$ , the left hand side of the Hamilton-Jacobi equation (22.16) separates as a sum of terms each of which depends only on  $x$  or only on  $y$  provided that

$$\text{for each } m, \quad \left\{ \begin{array}{l} \text{either } \hat{e}_m{}^\mu \text{ for all } \mu, \text{ and } \hat{A}_m, \text{ are functions of } x \text{ only, and } \hat{e}_m{}^y = 0 , \\ \text{or } \hat{e}_m{}^\mu \text{ for all } \mu, \text{ and } \hat{A}_m, \text{ are functions of } y \text{ only, and } \hat{e}_m{}^x = 0 . \end{array} \right. \quad (22.21)$$

The case that matches the Kerr and related geometries is the 2+2 choice

$$\text{the } \left\{ \begin{array}{l} \text{top} \\ \text{bottom} \end{array} \right\} \text{ condition of (22.21) holds for } \left\{ \begin{array}{l} m = 0 \text{ and } 1 \\ m = 2 \text{ and } 3 \end{array} \right\} . \quad (22.22)$$

Thus separability consistent with Kerr requires that

$$\hat{e}_0{}^y = \hat{e}_1{}^y = \hat{e}_2{}^x = \hat{e}_3{}^x = 0 . \quad (22.23)$$

Given the separability conditions (22.23), the 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  $\gamma_0$  and  $\gamma_1$ , and a (commuting) spatial rotation by angle  $\tan^{-1}(\hat{e}_3{}^y/\hat{e}_2{}^y)$  between tetrad axes  $\gamma_2$  and  $\gamma_3$ . Thus without loss of generality,

$$\hat{e}_0{}^x = \hat{e}_3{}^y = 0 . \quad (22.24)$$

The gauge conditions (22.24) having been effected, the 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

$$dt = dt' + \frac{\hat{e}_1{}^t}{\hat{e}_1{}^x} dx + \frac{\hat{e}_2{}^t}{\hat{e}_2{}^y} dy , \quad d\phi = d\phi' + \frac{\hat{e}_1{}^\phi}{\hat{e}_1{}^x} dx + \frac{\hat{e}_2{}^\phi}{\hat{e}_2{}^y} dy . \quad (22.25)$$

Equations (22.25) are integrable because  $\hat{e}_1{}^\mu$  and  $\hat{e}_2{}^\mu$  are respectively functions of  $x$  and  $y$  only. The transformations (22.25) of  $t$  and  $\phi$  are admissible because they preserve the Killing vectors  $\partial/\partial t$  and  $\partial/\partial\phi$ ,

$$\left. \frac{\partial}{\partial t} \right|_{x,y,\phi} = \left. \frac{\partial}{\partial t'} \right|_{x,y,\phi} , \quad \left. \frac{\partial}{\partial \phi} \right|_{t,x,y} = \left. \frac{\partial}{\partial \phi'} \right|_{t,x,y} . \quad (22.26)$$

Thus without loss of generality

$$\hat{e}_1^{\textcolor{brown}{t}} = \hat{e}_2^{\textcolor{brown}{t}} = \hat{e}_1^{\phi} = \hat{e}_2^{\phi} = 0 . \quad (22.27)$$

Finally, coordinate transformations of the  $x$  and  $y$  coordinates

$$x \rightarrow x' , \quad y \rightarrow y' , \quad (22.28)$$

can be chosen such that  $\hat{e}_1^{\textcolor{brown}{x}}$  is any function of  $x$ , and  $\hat{e}_2^{\textcolor{brown}{y}}$  is any function of  $y$ . A choice that proves advantageous in separating the Einstein and Maxwell equations is

$$\hat{e}_1^{\textcolor{brown}{x}} \hat{e}_0^{\textcolor{brown}{t}} = \hat{e}_2^{\textcolor{brown}{y}} \hat{e}_3^{\phi} = \pm 1 . \quad (22.29)$$

The separability conditions (22.21) with the 2+2 choice (22.22), which imply conditions (22.23), coupled with the gauge conditions (22.24), (22.27), and (22.29), bring the vierbein  $e_m^{\mu}$  to the form

$$e_m^{\mu} = \frac{1}{\rho} \begin{pmatrix} \frac{1}{\sqrt{\Delta_x}} & 0 & 0 & \frac{\omega_x}{\sqrt{\Delta_x}} \\ 0 & -\sqrt{\Delta_x} & 0 & 0 \\ 0 & 0 & \sqrt{\Delta_y} & 0 \\ \frac{\omega_y}{\sqrt{\Delta_y}} & 0 & 0 & \frac{1}{\sqrt{\Delta_y}} \end{pmatrix} , \quad (22.30)$$

where  $\omega_x$  and  $\Delta_x$  are some functions of  $x$ , and  $\omega_y$  and  $\Delta_y$  are some functions of  $y$ . The minus sign in  $e_1^{\textcolor{brown}{x}}$  is chosen so that, for  $\Lambda$ -Kerr-Newman, the radial tetrad basis vector  $\gamma_x$  points outward, the direction of increasing radius  $r$  but decreasing  $x$ . The corresponding inverse vierbein  $e^m_{\mu}$  is

$$e^m_{\mu} = \rho \begin{pmatrix} \frac{\sqrt{\Delta_x}}{1 - \omega_x \omega_y} & 0 & 0 & -\frac{\omega_y \sqrt{\Delta_x}}{1 - \omega_x \omega_y} \\ 0 & -\frac{1}{\sqrt{\Delta_x}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{\Delta_y}} & 0 \\ -\frac{\omega_x \sqrt{\Delta_y}}{1 - \omega_x \omega_y} & 0 & 0 & \frac{\sqrt{\Delta_y}}{1 - \omega_x \omega_y} \end{pmatrix} , \quad (22.31)$$

which implies the line-element (22.1).

The above form (22.30) of the vierbein was derived from the condition that the left hand side of the Hamilton-Jacobi equation (22.16) 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 non-zero, then the squared conformal factor  $\rho^2$  on the right hand side of the Hamilton-Jacobi equation (22.16) 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 (22.22) 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 (22.22)), 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

$$8\pi T_{mn}^e = \frac{Q_\bullet^2 + Q_\bullet^2}{\rho^4} \text{diag}(1, -1, 1, 1) . \quad (22.32)$$

The energy-momentum of a cosmological constant  $\Lambda$  is

$$8\pi T_{mn}^\Lambda = -\Lambda \eta_{mn} . \quad (22.33)$$

### 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:

$$G_{01} = G_{02} = G_{13} = G_{23} = 0 . \quad (22.34)$$

Of the remaining 6 Einstein components  $G_{mn}$ , the following 4 have zero electrovac source:

$$\rho^2 G_{12} = -2\sqrt{\Delta_x \Delta_y} \left[ \rho \frac{\partial^2(1/\rho)}{\partial x \partial y} - \frac{3}{4(1-\omega_x \omega_y)^2} \frac{d\omega_x}{dx} \frac{d\omega_y}{dy} \right], \quad (22.35a)$$

$$\rho^2 G_{03} = \frac{\sqrt{\Delta_x \Delta_y}}{2\rho^2} \left[ \frac{\partial}{\partial x} \left( \frac{\rho^2}{1-\omega_x \omega_y} \frac{d\omega_x}{dx} \right) - \frac{\partial}{\partial y} \left( \frac{\rho^2}{1-\omega_x \omega_y} \frac{d\omega_y}{dy} \right) \right], \quad (22.35b)$$

$$\rho^2 (G_{00} + G_{11}) = \frac{2\Delta_x}{1-\omega_x \omega_y} \left[ \rho \frac{\partial}{\partial x} \left( (1-\omega_x \omega_y) \frac{\partial(1/\rho)}{\partial x} \right) + \frac{1}{4(1-\omega_x \omega_y)} \left( \frac{d\omega_y}{dy} \right)^2 \right], \quad (22.35c)$$

$$\rho^2 (G_{22} - G_{33}) = \frac{2\Delta_y}{1-\omega_x \omega_y} \left[ \rho \frac{\partial}{\partial y} \left( (1-\omega_x \omega_y) \frac{\partial(1/\rho)}{\partial y} \right) + \frac{1}{4(1-\omega_x \omega_y)} \left( \frac{d\omega_x}{dx} \right)^2 \right]. \quad (22.35d)$$

If the conformal factor  $\rho^2$  is supposed to separate as a sum of radial and angular parts, equation (22.2), then the homogeneous version of equation (22.35a) reduces to

$$\frac{d(\rho_x^2)}{d\omega_x} \frac{d(\rho_y^2)}{d\omega_y} - \frac{(\rho_x^2 + \rho_y^2)^2}{(1-\omega_x \omega_y)^2} = 0. \quad (22.36)$$

Series expansion of (22.36) leads to the result that

$$\boxed{\rho^2 \equiv \rho_x^2 + \rho_y^2 = \frac{1-\omega_x \omega_y}{(f_0 + f_1 \omega_x)(f_1 + f_0 \omega_y)}}, \quad (22.37a)$$

$$\boxed{\rho_x = \sqrt{\frac{g_0 - g_1 \omega_x}{(f_0 g_1 + f_1 g_0)(f_0 + f_1 \omega_x)}}, \quad \rho_y = \sqrt{\frac{g_1 - g_0 \omega_y}{(f_0 g_1 + f_1 g_0)(f_1 + f_0 \omega_y)}}}, \quad (22.37b)$$

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 either  $\rho_x$  or  $\rho_y$ : 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  $\rho_x$  and  $\rho_y$ , and the relative sizes of  $g_0$  and  $g_1$  can be changed by adjusting an arbitrary constant in the split between  $\rho_x^2$  and  $\rho_y^2$ . Given the expression (22.37) for the conformal factor  $\rho$ , the Einstein component  $G_{03}$ , equation (22.35b), reduces to

$$\rho^2 G_{03} = \frac{\sqrt{\Delta_x \Delta_y}}{2(1-\omega_x \omega_y)} \left[ \frac{d\omega_x}{dx} \frac{d}{dx} \ln \left( \frac{d\omega_x/dx}{f_0 + f_1 \omega_x} \right) - \frac{d\omega_y}{dy} \frac{d}{dy} \ln \left( \frac{d\omega_y/dy}{f_1 + f_0 \omega_y} \right) \right]. \quad (22.38)$$

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

$$\frac{d\omega_x}{dx} = 2\sqrt{(f_0 + f_1 \omega_x) \left[ g_0 + \frac{1}{f_0} (f_1 g_0 + f_2) \omega_x \right]}, \quad (22.39a)$$

$$\frac{d\omega_y}{dy} = 2\sqrt{(f_1 + f_0 \omega_y) \left[ g_1 + \frac{1}{f_1} (f_0 g_1 + f_2) \omega_y \right]}, \quad (22.39b)$$

for some constants  $g_0$  and  $g_1$ , which can be taken without loss of generality to equal those in the conformal factor (22.37). With the conformal factor  $\rho$  given by equation (22.37) and  $d\omega_x/dx$  and  $d\omega_y/dy$  given by equations (22.39), the Einstein components  $G_{00} + G_{11}$  and  $G_{22} - G_{33}$  reduce to

$$\rho^2 (G_{00} + G_{11}) = 2\Delta_x \frac{(f_2 + f_0 g_1 + f_1 g_0)(f_0 + f_1 \omega_x)^2}{f_0 f_1 (1 - \omega_x \omega_y)^2}, \quad (22.40a)$$

$$\rho^2 (G_{22} - G_{33}) = 2\Delta_y \frac{(f_2 + f_0 g_1 + f_1 g_0)(f_1 + f_0 \omega_y)^2}{f_0 f_1 (1 - \omega_x \omega_y)^2}. \quad (22.40b)$$

These vanish provided that the constant  $f_2$  satisfies

$$f_2 = -(f_0 g_1 + f_1 g_0). \quad (22.41)$$

Inserting this value into equations (22.39) implies

$$\boxed{\frac{d\omega_x}{dx} = 2\sqrt{(f_0 + f_1 \omega_x)(g_0 - g_1 \omega_x)}}, \quad (22.42a)$$

$$\boxed{\frac{d\omega_y}{dy} = 2\sqrt{(f_1 + f_0 \omega_y)(g_1 - g_0 \omega_y)}}. \quad (22.42b)$$

The sign of the square root for  $d\omega_x/dx$  is the same as that for  $\rho_x$ , while the sign of the square root for  $d\omega_y/dy$  is the same as that for  $\rho_y$ .

#### 22.4.3 Separation of the remaining 2 Einstein equations

Define  $Y_x$  and  $Y_y$  by

$$Y_x \equiv \frac{d\Delta_x}{dx} - \Delta_x \frac{d}{dx} \ln \left[ (f_0 + f_1 \omega_x) \frac{d\omega_x}{dx} \right], \quad (22.43a)$$

$$Y_y \equiv \frac{d\Delta_y}{dy} - \Delta_y \frac{d}{dy} \ln \left[ (f_1 + f_0 \omega_y) \frac{d\omega_y}{dy} \right]. \quad (22.43b)$$

In terms of  $Y_x$  and  $Y_y$ , the Einstein components  $G_{00} - G_{11}$  and  $G_{22} + G_{33}$  are

$$\rho^2 (G_{00} - G_{11}) \quad (22.44a)$$

$$= \frac{1}{1 - \omega_x \omega_y} \left( Y_x \frac{d \ln \omega_x}{dx} - Y_y \frac{d \ln \omega_y}{dy} \right) + Y_x \frac{d}{dx} \ln \left( \frac{f_0 + f_1 \omega_x}{\omega_x} \right) - \frac{\partial Y_y}{\partial y} + Y_y \frac{d}{dy} \ln \left[ \frac{\omega_y (f_1 + f_0 \omega_y)}{d\omega_y/dy} \right],$$

$$\rho^2 (G_{22} + G_{33}) \quad (22.44b)$$

$$= \frac{1}{1 - \omega_x \omega_y} \left( Y_x \frac{d \ln \omega_x}{dx} - Y_y \frac{d \ln \omega_y}{dy} \right) - Y_y \frac{d}{dy} \ln \left( \frac{f_1 + f_0 \omega_y}{\omega_y} \right) + \frac{\partial Y_x}{\partial x} - Y_x \frac{d}{dx} \ln \left[ \frac{\omega_x (f_0 + f_1 \omega_x)}{d\omega_x/dx} \right].$$

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

$$\frac{1}{(1 - \omega_x \omega_y)} \left( \frac{f_0 h_0 + h_2 \omega_x + f_1 h_1 \omega_x^2}{\omega_x} - \frac{f_1 h_1 + h_2 \omega_y + f_0 h_0 \omega_y^2}{\omega_y} \right) - \frac{f_0 h_0 + h_3 \omega_x}{\omega_x} + \frac{f_1 h_1 + h_3 \omega_y}{\omega_y} = 0 , \quad (22.45)$$

for some constants  $h_0$ ,  $h_1$ ,  $h_2$ , and  $h_3$ . Separating each of equations (22.44) according to the pattern of equation (22.45) leads to the homogeneous solutions

$$Y_x = \frac{(f_0 + f_1 \omega_x)(h_0 + h_1 \omega_x)}{d\omega_x/dx} , \quad Y_y = \frac{(f_1 + f_0 \omega_y)(h_1 + h_0 \omega_y)}{d\omega_y/dy} . \quad (22.46)$$

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 (22.44) and separating that difference in the pattern of equation (22.45). The solution is a sum of a homogeneous solution (22.46) and a particular solution

$$Y_x = \frac{2(Q_\bullet^2 + Q_\bullet^2)(f_0 + f_1 \omega_x)^2}{d\omega_x/dx} , \quad Y_y = 0 . \quad (22.47)$$

Inserting equations (22.47) into the Einstein expressions (22.44) yields Einstein components that have precisely the form (22.32) 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 (22.44) in the pattern of equation (22.45). A particular solution is

$$Y_x = \frac{2\Lambda}{f_1^2 d\omega_x/dx} , \quad Y_\phi = \frac{2\Lambda \omega_y^2}{f_1^2 d\omega_x/dx} . \quad (22.48)$$

Inserting equations (22.48) into the Einstein expressions (22.44) yields Einstein components that have precisely the form of a cosmological constant,  $G_{mn} = -\Lambda \eta_{mn}$ .

Solving equations (22.43a) and (22.43b) with  $Y_x$  and  $Y_y$  given by a sum of the homogeneous, electromagnetic, and vacuum contributions, equations (22.46), (22.47), and (22.48), yields the general electrovac solution for the horizon and polar functions  $\Delta_x$  and  $\Delta_y$ ,

$$\begin{aligned} \Delta_x &= (f_0 + f_1 \omega_x) \left[ (k_0 + k_1 \omega_x) - \frac{2M_\bullet \sqrt{(f_0 + f_1 \omega_x)(g_0 - g_1 \omega_x)}}{(f_0 g_1 + f_1 g_0)^{3/2}} + \frac{(Q_\bullet^2 + Q_\bullet^2)(f_0 + f_1 \omega_x)}{f_0 g_1 + f_1 g_0} \right] \\ &\quad - \frac{\Lambda(g_0 - g_1 \omega_x)}{3f_1(f_0 g_1 + f_1 g_0)^2} , \end{aligned} \quad (22.49a)$$

$$\Delta_y = (f_1 + f_0 \omega_y) \left[ (k_1 + k_0 \omega_y) - \frac{2N_\bullet \sqrt{(f_1 + f_0 \omega_y)(g_1 - g_0 \omega_y)}}{(f_0 g_1 + f_1 g_0)^{3/2}} \right] + \frac{\Lambda \omega_y(g_1 - g_0 \omega_y)}{3f_1(f_0 g_1 + f_1 g_0)^2} , \quad (22.49b)$$

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 (22.49a) for  $\Delta_x$  is the black hole's mass. The constant  $N_\bullet$  in the expression (22.49b) for  $\Delta_y$  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  $A_k$  in terms of a scaled electromagnetic potential  $\mathcal{A}_k$ , equation (23.2). Separability of the Hamilton-Jacobi equations requires, equations (22.21) and (22.22), that

$$\begin{aligned} \mathcal{A}_t, \quad \mathcal{A}_x &\text{ are functions of } x \text{ only ,} \\ \mathcal{A}_y, \quad \mathcal{A}_\phi &\text{ are functions of } y \text{ only .} \end{aligned} \quad (22.50)$$

For the line-element (22.1), and with the conditions (22.50), the non-vanishing components of the tetrad-frame electromagnetic field  $F_{mn}$  are the radial electric  $E$  and magnetic  $B$  fields

$$E \equiv F_{10} = -\frac{1}{\rho^2} \left( \frac{d\mathcal{A}_t}{dx} + \frac{\omega_y \mathcal{A}_t - \mathcal{A}_\phi}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right) , \quad (22.51a)$$

$$B \equiv F_{23} = \frac{1}{\rho^2} \left( \frac{d\mathcal{A}_\phi}{dy} + \frac{\omega_x \mathcal{A}_\phi - \mathcal{A}_t}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right) . \quad (22.51b)$$

The remaining components of the electromagnetic field vanish identically,

$$F_{02} = F_{03} = F_{12} = F_{13} = 0 . \quad (22.52)$$

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,

$$\mathcal{A}_x = \mathcal{A}_y = 0 . \quad (22.53)$$

The electromagnetic field given by equations (22.51) and (22.52) satisfies automatically the source-free Maxwell's equations. The sourced Maxwell's equations are

$$D^m F_{mn} = 4\pi j_n . \quad (22.54)$$

Stationary solutions require vanishing current,  $j_n = 0$ . Two of the sourced Maxwell equations vanish identically, and the corresponding currents vanish automatically:

$$j_1 = j_2 = 0 . \quad (22.55)$$

Given the expressions (22.42) for  $d\omega_x/dx$  and  $d\omega_y/dy$ , the remaining two sourced Maxwell's equations can be written

$$-\frac{\sqrt{\Delta_x}}{\rho^3} \left[ \frac{\partial Z_t}{\partial x} + Z_t \frac{\partial}{\partial x} \ln \left( \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right) - Z_\phi \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right] = 4\pi j_0 , \quad (22.56a)$$

$$-\frac{\sqrt{\Delta_y}}{\rho^3} \left[ \frac{\partial Z_\phi}{\partial y} + Z_\phi \frac{\partial}{\partial y} \ln \left( \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_y}{dy} \right) - Z_t \frac{1}{1 - \omega_x \omega_y} \frac{d\omega_x}{dx} \right] = 4\pi j_3 , \quad (22.56b)$$

where  $Z_t$  and  $Z_\phi$  are defined to be

$$Z_t \equiv \frac{d\omega_x}{dx} \frac{\partial}{\partial x} \left( \frac{\mathcal{A}_t}{d\omega_x/dx} \right) , \quad (22.57a)$$

$$Z_\phi \equiv \frac{d\omega_y}{dy} \frac{\partial}{\partial y} \left( \frac{\mathcal{A}_\phi}{d\omega_y/dy} \right) . \quad (22.57b)$$

The homogeneous solutions of equations (22.56) are

$$Z_t = Z_\phi = 0 . \quad (22.58)$$

Homogeneous solution of equations (22.57) yields

$$\frac{\mathcal{A}_t}{d\omega_x/dx} \equiv -\frac{Q_\bullet}{2(f_0g_1 + f_1g_0)} , \quad (22.59a)$$

$$\frac{\mathcal{A}_\phi}{d\omega_y/dy} \equiv -\frac{Q_\bullet}{2(f_0g_1 + f_1g_0)} , \quad (22.59b)$$

where  $Q_\bullet$  and  $Q_\bullet$  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 (22.59) for  $\mathcal{A}_k$  into the expressions (22.51) yields the electric and magnetic fields (22.10).

### 22.5.2 Separation of Maxwell's equations

The form (22.56) of the Maxwell equations for  $j_0$  and  $j_3$  assumed that  $d\omega_x/dx$  and  $d\omega_y/dy$ , satisfy the equations (22.42) obtained by separating Einstein's equations. However, the Maxwell equations can also be separated directly, and the conditions (22.64) that result are consistent with the Einstein conditions (22.42).

If one provisionally supposes that  $\mathcal{A}_\phi = 0$ , then the Maxwell equation for the angular current  $j_3$  is

$$\begin{aligned} & \frac{\mathcal{A}_t \sqrt{\Delta_y}}{\rho^3(1 - \omega_x \omega_y)^2} \left\{ \frac{d\omega_x}{dx} \left[ (1 - \omega_x \omega_y) \frac{d \ln(\mathcal{A}_t / \omega_x)}{dx} + \frac{d \ln \omega_x}{dx} \right] + \frac{d\omega_y}{dy} \left[ (1 - \omega_x \omega_y) \frac{d}{dy} \ln \left( \frac{d \ln \omega_y}{dy} \right) + \frac{d \ln \omega_y}{dy} \right] \right\} \\ &= 4\pi j_3 . \end{aligned} \quad (22.60)$$

Conversely, if one provisionally supposes that  $\mathcal{A}_t = 0$ , then the Maxwell equation for the radial current  $j_0$  is

$$\begin{aligned} & \frac{\mathcal{A}_\phi \sqrt{\Delta_x}}{\rho^3(1 - \omega_x \omega_y)^2} \left\{ \frac{d\omega_y}{dy} \left[ (1 - \omega_x \omega_y) \frac{d \ln(\mathcal{A}_\phi / \omega_y)}{dy} + \frac{d \ln \omega_y}{dy} \right] + \frac{d\omega_x}{dx} \left[ (1 - \omega_x \omega_y) \frac{d}{dx} \ln \left( \frac{d \ln \omega_x}{dx} \right) + \frac{d \ln \omega_x}{dx} \right] \right\} \\ &= 4\pi j_0 . \end{aligned} \quad (22.61)$$

The homogeneous solutions of equations (22.60) and (22.61) prove to be the homogeneous solutions of the full equations without any restriction on  $\mathcal{A}_t$  or  $\mathcal{A}_\phi$ . Equation (22.60) separates with 4 separation constants  $q_i$  as

$$(1 - \omega_x \omega_y) \left( \frac{-q_0 + q_3 \omega_x}{\omega_x} \right) + \frac{q_0 - 2q_1 \omega_x - q_2 \omega_x^2}{\omega_x} + (1 - \omega_x \omega_y) \left( \frac{-q_2 - q_3 \omega_y}{\omega_y} \right) + \frac{q_2 + 2q_1 \omega_y - q_0 \omega_y^2}{\omega_y} = 0 , \quad (22.62)$$

and equation (22.60) separates in a similar fashion, the vanishing of the second term inside braces in either of equations (22.60) or (22.61) requiring that

$$q_3 = q_1 . \quad (22.63)$$

The separated solutions for  $\omega_x$  and  $\omega_y$  are

$$\frac{d\omega_x}{dx} = 2\sqrt{q_0 - 2q_1\omega_x - q_2\omega_x^2} , \quad (22.64a)$$

$$\frac{d\omega_y}{dy} = 2\sqrt{q_2 + 2q_1\omega_y - q_0\omega_y^2} . \quad (22.64b)$$

These are consistent with the separated solution (22.42) found for Einstein's equations provided that

$$q_0 = f_0 g_0 , \quad 2q_1 = f_0 g_1 - f_1 g_0 , \quad q_2 = f_1 g_1 . \quad (22.65)$$

## 22.6 $\Lambda$ -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 (22.49a) for the radial horizon function  $\Delta_x$  shows that the effect of mass and charge becomes negligible where  $f_0 + f_1\omega_x \rightarrow 0$ . Expression (22.37) for the separable conformal factor  $\rho$  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  $\omega_x$  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  $\omega_x$  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 (22.75), it is shown that the situation where  $f_0$  is non-zero 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  $g_{\phi\phi}$  vanish at the poles of the rotation axis,  $\theta = 0$  and  $\pi$ , to avoid singular behaviour at the poles. The three metric coefficients  $g_{tt}$ ,  $g_{t\phi}$ ,

and  $g_{\phi\phi}$  are gauge-invariant in view of the invariant character of  $t$  and  $\phi$ , and they are

$$g_{tt} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (-\Delta_x + \omega_x^2 \Delta_y) , \quad (22.66a)$$

$$g_{t\phi} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (\omega_y \Delta_x - \omega_x \Delta_y) , \quad (22.66b)$$

$$g_{\phi\phi} = \frac{\rho^2}{(1 - \omega_x \omega_y)^2} (-\omega_y^2 \Delta_x + \Delta_y) . \quad (22.66c)$$

The vanishing of  $g_{t\phi}$  and  $g_{\phi\phi}$  at the poles requires that both  $\omega_y$  and  $\Delta_y$  must vanish at the poles.

Connection with familiar polar coordinates  $\{r, \theta, \phi\}$  may be established by requiring that the metric coefficients (22.66a) go over to their asymptotic expressions in the absence of a cosmological constant or NUT parameter,

$$g_{tt} \rightarrow -1 , \quad g_{t\phi} \rightarrow 0 , \quad g_{\phi\phi} \rightarrow r^2 \sin^2 \theta \quad \text{as } r \rightarrow \infty . \quad (22.67)$$

The expressions (22.49) for the horizon functions  $\Delta_x$  and  $\Delta_y$  then imply that, in the absence of a cosmological constant or NUT parameter,

$$\Delta_y = \frac{\omega_y}{a} = \sin^2 \theta , \quad \Delta_x \rightarrow \frac{\omega_x}{a} \rightarrow \frac{1}{r^2} \quad \text{as } r \rightarrow \infty , \quad (22.68)$$

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  $\rho$  fixes  $f_1 = a^{-1/2}$ .

Integrating the relation (22.42b) between  $\omega_y$  and  $y$  with  $f_0 = 0$  establishes that  $\omega_y$  is quadratic in  $y$ . Requiring that the polar part of the metric  $g_{yy} dy^2 \equiv \rho^2 dy^2 / \Delta_y$  be non-singular 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

$$y = -\cos \theta , \quad (22.69)$$

where the sign has been chosen so that  $y$  increases as  $\theta$  increases. Expression (22.69) can be imposed also in the presence of a cosmological constant and a NUT parameter. For  $\Lambda$ -Kerr-Newman with no NUT parameter,

$$\omega_y = a \sin^2 \theta . \quad (22.70)$$

Equation (22.70) is not true if the NUT parameter is non-vanishing, a case deferred to §22.7. The expressions (22.69) and (22.70) are consistent with the relation (22.42b) between them provided that  $g_1 = a g_0$  and  $g_0 = a/f_1$ . The complete set of constants in equations (22.49) is

$$f_0 = 0 , \quad f_1 = a^{-1/2} , \quad g_0 = a^{3/2} , \quad g_1 = a^{5/2} , \quad k_0 = a^{-1/2} , \quad k_1 = a^{-3/2} Q_\bullet^2 . \quad (22.71)$$

The non-zero value for  $k_1$  for non-zero magnetic charge  $Q_\bullet$  ensures non-singular behaviour at the poles.

The radial variable  $x$  analogous to the angular variable  $y$  of equation (22.69) comes from solving equation (22.42a),  $d\omega_x/dx = 2\sqrt{a\omega_x(1-a\omega_x)}$ , which gives

$$x = \frac{1}{a} \sin^{-1} \sqrt{a\omega_x} . \quad (22.72)$$

A pair of radial and angular variables that emerged naturally from the analysis, besides  $x$  and  $y$ , are  $\rho_x$  and  $\rho_y$  defined by equation (22.37b). In terms of  $\omega_x$ , 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  $\rho_x$ , which is consistent with the asymptotic behaviour  $\rho \rightarrow r$  as  $r \rightarrow \infty$ , in which case

$$\omega_x = \frac{a}{R^2} , \quad R = \sqrt{r^2 + a^2} . \quad (22.73)$$

The radial and angular variables  $\rho_x$  and  $\rho_y$  are

$$\rho_x = r , \quad \rho_y = a \cos \theta . \quad (22.74)$$

This completes the derivation of the  $\Lambda$ -Kerr-Newman solutions.

### 22.6.1 There are no solutions that rotate at infinity

The  $\Lambda$ -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 non-zero? If  $f_0$  is non-zero, then the tetrad rotates through the coordinates with some constant finite angular velocity  $\omega_\infty$  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  $\phi$  so that it corotates at infinity. The line-element can then be brought to standard  $\Lambda$ -Kerr-Newman form with  $f_0 = 0$  by a coordinate transformation of the angular coordinate  $y$ . Specifically, the coordinate transformations

$$dy' = (1 - \omega_\infty \omega_y) dy , \quad \phi' = \phi + \omega_\infty t , \quad (22.75)$$

bring the line-element with  $\omega_\infty \neq 0$  to the standard  $\Lambda$ -Kerr-Newman form with  $\omega_\infty = 0$ ,

$$ds^2 = \rho^2 \left[ -\frac{\Delta_x}{(1 - \omega'_x \omega'_y)^2} (dt - \omega'_y d\phi)^2 + \frac{dx^2}{\Delta_x} + \frac{dy'^2}{\Delta'_y} + \frac{\Delta'_y}{(1 - \omega'_x \omega'_y)^2} (d\phi' - \omega'_x dt)^2 \right] , \quad (22.76)$$

with primed quantities

$$\omega'_x \equiv \omega_x - \omega_\infty , \quad \omega'_y \equiv \frac{\omega_y}{1 - \omega_\infty \omega_y} , \quad \Delta'_y \equiv \frac{\Delta_y}{(1 - \omega_\infty \omega_y)^2} . \quad (22.77)$$

Thus, at least among separable electrovac solutions, there are no solutions that physically rotate at infinity.

## 22.7 Taub-NUT geometry

Spacetimes with a finite NUT parameter  $N_\bullet$  (Taub, 1951; Newman, Tamburino, and Unti, 1963; Stephani et al., 2003; Kagramanova et al., 2010) have closed timelike curves circulating around one or both polar axes, so they are fun, but not physically realistic.

### 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 (22.71) generalize to

$$\begin{aligned} f_0 &= 0 , \quad f_1 = a^{-1/2} , \quad g_0 = a^{3/2} , \quad g_1 = a^{1/2} b^2 , \\ k_0 &= a^{-1/2} [1 - \frac{1}{3}\Lambda(a^2 - 2c_\bullet N_\bullet + N_\bullet^2)] , \quad k_1 = a^{-3/2} [\mathcal{Q}_\bullet^2 - 2N_\bullet(N_\bullet + ac_\bullet + \frac{2}{3}a^2\Lambda N_\bullet(c_\bullet^2 - 1))] , \end{aligned} \quad (22.78)$$

where

$$b \equiv \sqrt{a^2 + 2ac_\bullet N_\bullet + N_\bullet^2} . \quad (22.79)$$

Besides the NUT parameter  $N_\bullet$ , there is an additional constant of integration  $c_\bullet$ .

The resulting Taub-NUT line-element takes the separable form (22.1), in which the radial and angular parts  $\rho_x$  and  $\rho_y$  of the conformal factor are

$$\rho_x \equiv r = b \cot(bx) , \quad \rho_y \equiv N_\bullet + a \cos \theta = N_\bullet - ay , \quad (22.80)$$

the coefficients  $\omega_x$  and  $\omega_y$  are

$$\omega_x = \frac{a}{R^2} , \quad R \equiv \sqrt{r^2 + b^2} , \quad \omega_y = a \sin^2 \theta + 2N_\bullet(c_\bullet - \cos \theta) , \quad (22.81)$$

and the horizon and polar functions  $\Delta_x$  and  $\Delta_y$  are

$$\Delta_x = -\frac{1}{3}\Lambda + \frac{1}{R^2} \left(1 + \frac{4}{3}a\Lambda c_\bullet N_\bullet\right) + \frac{1}{R^4} \left[-2M_\bullet r + Q_\bullet^2 + \mathcal{Q}_\bullet^2 - 2N_\bullet(N_\bullet + ac_\bullet + \frac{2}{3}a^2\Lambda N_\bullet(c_\bullet^2 - 1))\right] , \quad (22.82a)$$

$$\Delta_y = \sin^2 \theta \left[1 - \frac{1}{3}a^2\Lambda \sin^2 \theta + \frac{4}{3}\Lambda N_\bullet(N_\bullet + a \cos \theta)\right] . \quad (22.82b)$$

Poles occur where  $\Delta_y = 0$ , that is, at  $\theta = 0$  or  $\pi$ . A sisytube encircles any pole where  $\omega_y$  fails to vanish. If the NUT parameter  $N_\bullet$  is non-zero, then generically sisytubes encircle both poles, but for the special cases  $c_\bullet = \pm 1$ , a sisytube encircles only one of the two poles. Horizons occur where  $\Delta_x = 0$ . For vanishing  $\Lambda$ , there are outer and inner horizons at

$$r_\pm = M_\bullet \pm \sqrt{M_\bullet^2 + N_\bullet^2 - Q_\bullet^2 - \mathcal{Q}_\bullet^2 - a^2} . \quad (22.83)$$

The horizon and polar functions  $\Delta_x$  and  $\Delta_y$  given by equations (22.82) do not agree with the earlier expressions (22.8) for non-zero  $\Lambda$  and vanishing  $N_\bullet$ , but this is not a misprint. The difference arises from

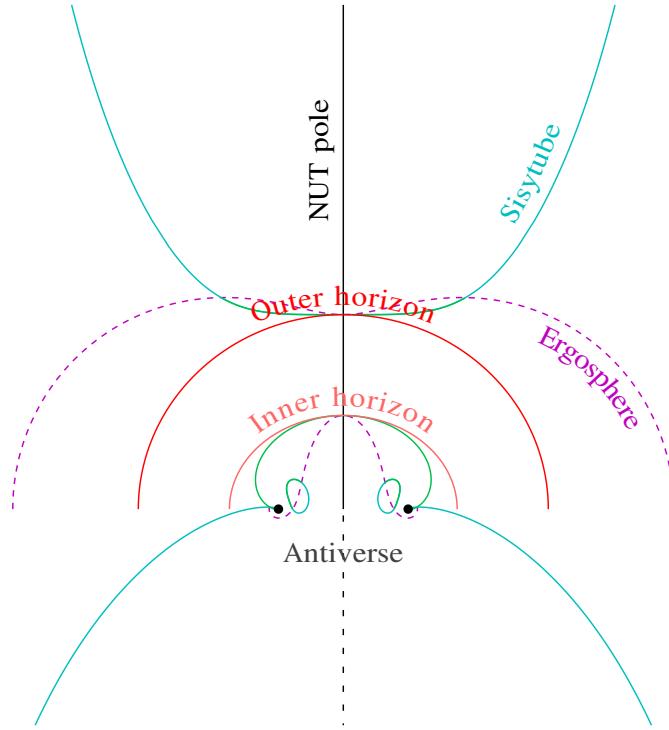


Figure 22.1 Geometry of a Kerr-NUT black hole, with NUT parameters  $N_\bullet = 0.75M$  and  $c_\bullet = -1$ , and spin parameter  $a = 1.2M$ . The outer sisyphus tube encircles the northern polar axis outside the outer ergosphere, while the inner sisyphus tube encircles the extension of the northern axis into the Antiverse inside the inner ergosphere. The choice  $c_\bullet = -1$  means that there is no sisyphus tube 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  $g_{\phi\phi} = 0$ . Sisyphus tubes, the regions containing closed timelike curves, are where  $g_{tt} \geq 0$  and  $g_{\phi\phi} \geq 0$ .

an arbitrariness in the choice of homogeneous solution when a cosmological constant  $\Lambda$  is present, equations (22.49). The tetrad-frame electromagnetic potential  $A_k$  is

$$A_k = \frac{1}{\rho} \left\{ -\frac{Q_\bullet r}{R^2 \sqrt{\Delta_x}}, 0, 0, -\frac{Q_\bullet (a \cos \theta + N_\bullet)}{a \sqrt{\Delta_y}} \right\}. \quad (22.84)$$

### 22.7.2 Closed timelike curves in Taub-NUT

Sisyphus tubes, containing closed timelike curves, occur in regions where  $g_{tt} \geq 0$  and  $g_{\phi\phi} \geq 0$ , Exercise (23.3). A sisyphus tube encircles any pole where  $\omega_y$  fails to vanish.

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 sisyphus tube encircles the north pole but not the south pole.

**Exercise 22.3 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

$$dl^2 = \rho^2 \left[ \frac{dy^2}{\Delta_y} + \frac{(\Delta_y - \omega_y^2 \Delta_x) d\phi^2}{(1 - \omega_x \omega_y)^2} \right]. \quad (22.85)$$

The angular line-element is diagonal, with proper distances in the two orthogonal  $y$  and  $\phi$  directions

$$\frac{\rho dy}{\sqrt{\Delta_y}}, \quad \frac{\rho \sqrt{\Delta_y - \omega_y^2 \Delta_x} d\phi}{1 - \omega_x \omega_y}. \quad (22.86)$$

The area of the angular  $y$ - $\phi$  surface at fixed radius  $x$  and time  $t$  is obtained by integrating the product of the proper distances over the surface,

$$A = \int \frac{\rho^2 \sqrt{(1 - \omega_y^2 \Delta_x / \Delta_y)}}{(1 - \omega_x \omega_y)} dy d\phi. \quad (22.87)$$

Horizons occur where the horizon function vanishes,  $\Delta_x = 0$ , in which case the area simplifies to

$$\begin{aligned} A &= \int \frac{\rho^2}{(1 - \omega_x \omega_y)} dy d\phi \\ &= 2\pi \int \frac{1}{(f_0 + f_1 \omega_x)(f_1 + f_0 \omega_y)} dy \\ &= \frac{2\pi}{f_0 + f_1 \omega_x} \int \frac{d\omega_y}{2\sqrt{(f_1 + f_0 \omega_y)^3 (g_1 - g_0 \omega_y)}} \\ &= \frac{2\pi}{(f_0 g_1 + f_1 g_0)(f_0 + f_1 \omega_x)} \left[ \sqrt{\frac{g_1 - g_0 \omega_y}{f_1 + f_0 \omega_y}} \right]. \end{aligned} \quad (22.88)$$

The second line of equations invokes equation (22.37a), while the third line uses equation (22.42b) to transform the integral over  $y$  to an integral over  $\omega_y$ . The constants are given by equation (22.71) for  $\Lambda$ -Kerr-Newman, and equations (22.78) for Taub-NUT. The integration over  $y$  is from  $-1$  to  $1$ , south to north pole. For  $\Lambda$ -Kerr-Newman,  $\omega_y = 0$  at both poles, but for Taub-NUT,  $\omega_y = 2N_\bullet(c_\bullet \pm 1)$  at the poles, equation (22.81). In either case, for both  $\Lambda$ -Kerr-Newman and Taub-NUT, the area of the horizon is

$$A = 4\pi R^2. \quad (22.89)$$


---

## 23

---

### Trajectories in ideal rotating black holes

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  $\Lambda$ -Kerr-Newmann geometry can be brought to a simple form (23.7) by writing the covariant tetrad-frame momentum  $p_k$  of a particle in terms of a set of Hamilton-Jacobi parameters  $P_k$ ,

$$p_k \equiv \frac{1}{\rho} \left\{ \frac{P_t}{\sqrt{\Delta_x}}, \frac{P_x}{\sqrt{\Delta_x}}, \frac{P_y}{\sqrt{\Delta_y}}, \frac{P_\phi}{\sqrt{\Delta_y}} \right\}, \quad (23.1)$$

and the covariant tetrad-frame electromagnetic potential  $A_k$  in terms of a set of Hamilton-Jacobi potentials  $\mathcal{A}_k$ ,

$$A_k \equiv \frac{1}{\rho} \left\{ \frac{\mathcal{A}_t}{\sqrt{\Delta_x}}, \frac{\mathcal{A}_x}{\sqrt{\Delta_x}}, \frac{\mathcal{A}_y}{\sqrt{\Delta_y}}, \frac{\mathcal{A}_\phi}{\sqrt{\Delta_y}} \right\}. \quad (23.2)$$

The contravariant coordinate momenta  $dx^\kappa/d\lambda = e_k{}^\kappa p^k$  are related to the Hamilton-Jacobi parameters  $P_k$  by

$$\frac{dx^\kappa}{d\lambda} = \frac{1}{\rho^2} \left\{ -\frac{P_t}{\Delta_x} + \frac{\omega_y P_\phi}{\Delta_y}, -P_x, P_y, -\frac{\omega_x P_t}{\Delta_x} + \frac{P_\phi}{\Delta_y} \right\}. \quad (23.3)$$

The tetrad-frame momenta  $p_k$  are related to the generalized momenta  $\pi_\kappa$  by  $p_k = e_k^\kappa \pi_\kappa - qA_k$ , which implies that the Hamilton-Jacobi parameters  $P_k$  are related to the canonical momenta  $\pi_\kappa$  by

$$P_t \equiv \pi_{\textcolor{brown}{t}} + \pi_\phi \omega_x - q\mathcal{A}_t , \quad (23.4\text{a})$$

$$P_x \equiv -\Delta_x \pi_{\textcolor{brown}{x}} - q\mathcal{A}_x , \quad (23.4\text{b})$$

$$P_y \equiv \Delta_y \pi_{\textcolor{brown}{y}} - q\mathcal{A}_y , \quad (23.4\text{c})$$

$$P_\phi \equiv \pi_\phi + \pi_t \omega_y - q\mathcal{A}_\phi . \quad (23.4\text{d})$$

Time translation symmetry and axisymmetry imply that  $\pi_{\textcolor{brown}{t}}$  and  $\pi_\phi$  are constants of motion, equation (22.19),

$$\pi_{\textcolor{brown}{t}} = -E , \quad \pi_\phi = L . \quad (23.5)$$

The separability conditions derived in §22.3 imply that

$$\begin{aligned} P_t , \quad P_x &\quad \text{are functions of } x \text{ only ,} \\ P_y , \quad P_\phi &\quad \text{are functions of } y \text{ only .} \end{aligned} \quad (23.6)$$

In terms of the Hamilton-Jacobi parameters  $P_k$ , the Hamilton-Jacobi equation (22.16) is

$$\frac{-P_t^2 + P_x^2}{\Delta_x} + \frac{P_y^2 + P_\phi^2}{\Delta_y} = -m^2 \rho^2 . \quad (23.7)$$

Separability for massive particles,  $m \neq 0$ , requires that the conformal factor  $\rho$  separate as equation (22.2). The Hamilton-Jacobi equation (23.7) then separates as

$$-\left(\frac{-P_t^2 + P_x^2}{\Delta_x} + m^2 \rho_x^2\right) = \frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho_y^2 = \mathcal{K} , \quad (23.8)$$

with  $\mathcal{K}$  a separation constant, the **Carter constant**. The separated Hamilton-Jacobi equations (23.8) imply that

$$P_x = \pm \sqrt{P_t^2 - (\mathcal{K} + m^2 \rho_x^2) \Delta_x} , \quad (23.9\text{a})$$

$$P_y = \pm \sqrt{-P_\phi^2 + (\mathcal{K} - m^2 \rho_y^2) \Delta_y} . \quad (23.9\text{b})$$

From the expression (23.3) for the coordinate momenta  $dx^\kappa/d\lambda$ , the trajectory of a freely-falling particle follows from integrating  $dy/dx = -P_y/P_x$ , equivalent to the implicit equation

$$-\frac{dx}{P_x} = \frac{dy}{P_y} . \quad (23.10)$$

Again from expression (23.3), the time and azimuthal coordinates  $t$  and  $\phi$  along the trajectory are then obtained by quadratures,

$$dt = \frac{P_t dx}{P_x \Delta_x} + \frac{\omega_y P_\phi dy}{P_y \Delta_y} , \quad d\phi = \frac{\omega_x P_t dx}{P_x \Delta_x} + \frac{P_\phi dy}{P_y \Delta_y} . \quad (23.11)$$

Again from expression (23.3), the affine parameter  $\lambda$  along the trajectory satisfies  $d\lambda/\rho^2 = -dx/P_x = dy/P_y$ , so similarly reduces to quadratures,

$$d\lambda = -\frac{\rho_x^2 dx}{P_x} + \frac{\rho_y^2 dy}{P_y} . \quad (23.12)$$

In the limiting case of trajectories at constant latitude  $y$ , where  $dy/P_y$  is zero divided by zero, expressions for  $t$ ,  $\phi$ , and  $\lambda$  along the trajectory are obtained by replacing  $dy/P_y \rightarrow -dx/P_x$  in equations (23.11) and (23.12). Similarly for circular trajectories, where  $dx/P_x$  is zero divided by zero, expressions for  $t$ ,  $\phi$ , and  $\lambda$  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 (22.10), obtained by multiplying by the pseudoscalar  $I$ , equation (13.22), is the same expression with the electric  $Q_\bullet$  and magnetic  $\mathcal{Q}_\bullet$  charges of the black hole exchanged according to

$$Q_\bullet \rightarrow -\mathcal{Q}_\bullet , \quad \mathcal{Q}_\bullet \rightarrow Q_\bullet . \quad (23.13)$$

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

$$qQ_\bullet \rightarrow q_e Q_\bullet + q_m \mathcal{Q}_\bullet , \quad q\mathcal{Q}_\bullet \rightarrow q_e \mathcal{Q}_\bullet - q_m Q_\bullet , \quad (23.14)$$

in the expressions (23.4a) and (23.4d) for  $P_t$  and  $P_\phi$ . The particle magnetic charge  $q_m$  is set to zero hereafter; it can be reincorporated by making the transformation (23.14) of constants.

## 23.3 Killing vectors and Killing tensor

The Kerr-Newman geometry is stationary and axisymmetric. As such it has two Killing vectors  $e_t$  and  $e_\phi$ , §7.32. The symmetries imply conservation of energy  $E = -\pi_t$  and azimuthal angular momentum  $L = \pi_\phi$  of freely-falling particles, equations (22.19).

The separability of the Kerr-Newman geometry means that it also has a Killing tensor  $K^{mn}$ . The Hamilton-Jacobi equation (23.8) can be written in terms of the tetrad-frame momenta  $p_k$ , equation (23.1), as

$$K^{mn} p_m p_n = \mathcal{K} , \quad (23.15)$$

where  $K^{mn}$  is

$$K^{mn} = \text{diag}(-\rho_y^2, \rho_y^2, \rho_x^2, \rho_x^2) . \quad (23.16)$$

The Killing tensor  $K^{mn}$  satisfies Killing's equation

$$D_{(k} K_{mn)} = 0 . \quad (23.17)$$

### 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 (23.10), which with the expressions (23.9) for  $P_x$  and  $P_y$  can be written

$$\frac{d\lambda}{\rho^2} = -\frac{dx}{\sqrt{P_t^2 - (\mathcal{K} + m^2 \rho_x^2) \Delta_x}} = \frac{dy}{\sqrt{-P_\phi^2 + (\mathcal{K} - m^2 \rho_y^2) \Delta_y}} . \quad (23.18)$$

At points where the polar function vanishes,  $\Delta_y = 0$ , the Hamilton-Jacobi equation (23.8) implies that

$$P_y = P_\phi = 0 \quad \text{at } \Delta_y = 0 . \quad (23.19)$$

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  $\Delta_y$  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.

### 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 (23.7) rearranges to

$$P_t^2 - P_x^2 = \left( \frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho^2 \right) \Delta_x . \quad (23.20)$$

This shows that the Hamilton-Jacobi parameters  $P_t$  and  $P_x$  must satisfy

$$\begin{aligned} |P_t| &> |P_x| & \text{if } \Delta_x > 0 , \\ |P_t| &= |P_x| & \text{if } \Delta_x = 0 , \\ |P_t| &< |P_x| & \text{if } \Delta_x < 0 . \end{aligned} \quad (23.21)$$

The Hamilton-Jacobi parameters must be continuous, including across horizons. Thus  $P_t$  must have the same sign everywhere throughout any connected region where  $\Delta_x$  is positive, which in the Kerr-Newman geometry

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	$P_t = P_x > 0$

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  $\Delta_x$  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 (23.8) shows that the Carter constant  $\mathcal{K}$  is necessarily positive. The vanishing of the Carter constant,

$$\mathcal{K} = 0 , \quad (23.22)$$

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

$$P_y = P_\phi = 0 . \quad (23.23)$$

They further satisfy  $P_t^2 = P_x^2$ . Outgoing and ingoing geodesics are distinguished by the relative signs of  $P_t$  and  $P_x$ ,

$$\begin{aligned} P_t &= -P_x && \text{outgoing} \\ P_t &= P_x && \text{ingoing} . \end{aligned} \quad (23.24)$$

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

$$\frac{L}{E} = \omega_y . \quad (23.25)$$

The affine parameter  $\lambda$  along a principal null congruence satisfies

$$d\lambda \propto \frac{\rho^2 dx}{P_t/E} = \frac{\rho^2 dx}{1 - \omega_x \omega_y} = \frac{dr}{\sqrt{f_0 g_1 + f_1 g_0} (f_1 + f_0 \omega_y)} , \quad (23.26)$$

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

$$\lambda = \pm r . \quad (23.27)$$

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  $u^\mu$  of the tetrad frame through the coordinates is

$$u^\mu = e_0^\mu = \frac{1}{\rho \sqrt{\Delta_x}} \{1, 0, 0, \omega_x\} . \quad (23.28)$$

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

$$\mathcal{K} = \mathcal{Q} + \left. \frac{P_\phi^2}{\Delta_y} \right|_{\rho_y=0} , \quad (23.29)$$

which has the property that  $\mathcal{Q} = 0$  for orbits in the equatorial plane,  $\rho_y = 0$ . For  $\Lambda$ -Kerr-Newman, the Carter integral is

$$\mathcal{Q} = \mathcal{K} - (L - aE)^2. \quad (23.30)$$

**Exercise 23.1 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

$$\mathcal{Q} \geq 0, \quad (23.31)$$

where  $\mathcal{Q}$  is the Carter integral defined by equation (23.30).

2. Argue that for a particle to reach the singularity at  $r = 0$ , positivity of  $P_x^2$  requires that

$$Q_\bullet^2(L - aE)^2 + (Q_\bullet^2 + a^2)\mathcal{Q} \leq 0. \quad (23.32)$$

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,

$$|q| \geq |m|. \quad (23.33)$$

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,

$$|q| \geq |m| \sqrt{\frac{Q_\bullet^2 + a^2}{Q_\bullet^2}}, \quad (23.34)$$

which generalizes the Reissner-Nordström condition (23.33).

**Solution.** Equation (23.31) comes from

$$\mathcal{K} = \frac{P_y^2 + P_\phi^2}{\Delta_y} + m^2 \rho_y^2 \geq \frac{P_\phi^2}{\Delta_y}, \quad (23.35)$$

and taking the limit  $y \rightarrow 0$ . Equation (23.32) comes from

$$R^4 \{ -P_t^2 + [\mathcal{Q} + (L - aE)^2 + m^2 \rho_x^2] \Delta_x \} = -R^4 P_x^2 \leq 0 , \quad (23.36)$$

and taking the limit  $r \rightarrow 0$ .

**Exercise 23.2 When must  $t$  and  $\phi$  progress forwards on a geodesic?** Under what circumstances must the time coordinate  $t$  or azimuthal angle  $\phi$  progress forwards along a geodesic?

1. Show that, in regions where  $\Delta_x \geq 0$ ,

$$\frac{P_\phi^2}{\Delta_y} \leq \mathcal{K} \leq \frac{P_t^2}{\Delta_x} . \quad (23.37)$$

Hence show that in the Universe, Wormhole, and Antiverse regions outside the horizons, where  $P_t < 0$ ,

$$\frac{dt}{d\lambda} \geq \frac{(1 - \omega_x \omega_y)^2 \sqrt{\mathcal{K}}}{\rho^4 \sqrt{\Delta_x \Delta_y} (\omega_y \sqrt{\Delta_x} + \sqrt{\Delta_y})} g_{\phi\phi} = -\frac{\sqrt{\mathcal{K} \Delta_x \Delta_y}}{\omega_y \sqrt{\Delta_x} + \sqrt{\Delta_y}} g_{tt} , \quad (23.38)$$

and

$$\frac{d\phi}{d\lambda} \geq \frac{(1 - \omega_x \omega_y)^2 \sqrt{\mathcal{K}}}{\rho^4 \sqrt{\Delta_x \Delta_y} (\sqrt{\Delta_x} + \omega_x \sqrt{\Delta_y})} g_{tt} = -\frac{\sqrt{\mathcal{K} \Delta_x \Delta_y}}{\sqrt{\Delta_x} + \omega_x \sqrt{\Delta_y}} g_{\phi\phi} . \quad (23.39)$$

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  $\phi$  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 (23.38) and (23.39) 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  $\phi$  must progress backwards between the ergospheres.

**Exercise 23.3 Inside the sisytube.** The sisytube, §9.10, is the region where  $g_{tt} \geq 0$  and  $g_{\phi\phi} \leq 0$ . Show that, for null geodesics at a turnaround point in both radius and latitude inside the sisytube,  $dx = dy = 0$ ,

$$\frac{dt}{d\phi} = \frac{\omega_y \sqrt{\Delta_x} \pm \sqrt{\Delta_y}}{\sqrt{\Delta_x} \pm \omega_x \sqrt{\Delta_y}} = \frac{\omega_y \sqrt{\Delta_x} + \sqrt{\Delta_y}}{\sqrt{\Delta_x} + \omega_x \sqrt{\Delta_y}} \left\{ 1 \quad \text{or} \quad \frac{g_{\phi\phi}}{g_{tt}} \right\} . \quad (23.40)$$

Sketch the lightcone structure in the  $t$ - $\phi$  plane inside the sisytube. When does the time  $t$  progress forwards, backwards, or not at all? To go backwards in time, must a particle go prograde or retrograde?

**Solution.** Retrograde.

**Exercise 23.4 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

$$ds^2 = - (dt - \omega_y d\phi)^2 + dx^2 + \frac{dy^2}{\Delta_y} + \Delta_y d\phi^2 . \quad (23.41)$$

Show that the tetrad-frame energy-momentum tensor is diagonal provided that  $\omega_y$  is linear in  $y$ . Show that the energy-momentum is constant everywhere provided that  $\Delta_y$  is quadratic in  $y$ . Show that the energy-momentum takes perfect fluid form with an ultra-hard equation of state (pressure = energy density) if

$$\omega_y = \frac{2y}{y_s}, \quad \Delta_y = 2y \left(1 + \frac{y}{y_s}\right), \quad (23.42)$$

where  $y_s$  is the boundary of the sisytube. Explore Gödel's Universe.

---

## 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 23.5 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

$$aL + qQr \leq -R^2 \sqrt{\left(\frac{L^2}{\Delta_y} + m^2\rho^2\right)} \Delta_x. \quad (23.43)$$

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 §23.16.

---

## 23.9 Constant latitude trajectories in the Kerr-Newman geometry

For simplicity, the next several sections, up to and including §23.19, 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  $\theta$ , or equivalently at constant  $y \equiv -\cos \theta$ ,

$$y = \text{constant}. \quad (23.44)$$

Constant latitude orbits occur where the angular potential  $P_y^2$ , equation (23.9b), not only vanishes, but is an extremum,

$$P_y^2 = \frac{dP_y^2}{dy} = 0 , \quad (23.45)$$

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  $L$

$$E = \pm \sqrt{1 + \frac{L^2}{a^2 \sin^4 \theta}} . \quad (23.46)$$

Solutions at any polar angle  $\theta$  and any angular momentum  $L$  exist, ranging from  $E = \pm 1$  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,

$$\mathcal{Q}_\bullet = \Lambda = 0 . \quad (23.47)$$

For brevity, the black hole subscripts will be dropped from the black hole mass and electric charge  $M_\bullet$  and  $Q_\bullet$ ,

$$M_\bullet = M , \quad Q_\bullet = Q . \quad (23.48)$$

### 23.10.1 Condition for a circular orbit

An orbit can be termed circular if it is at constant radius  $r$ ,

$$r = \text{constant} . \quad (23.49)$$

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 (23.9a), vanishes. Circular orbits occur where the radial potential  $P_x^2$  not only vanishes, but is an extremum,

$$P_x^2 = \frac{dP_x^2}{dr} = 0 , \quad (23.50)$$

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

$$\frac{d^2P_x^2}{dr^2} , \quad (23.51)$$

with  $-$  for stable,  $+$  for unstable circular orbits. Marginally stable orbits occur where  $d^2P_x^2/dr^2 = 0$ .

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  $\alpha$  by

$$\alpha \equiv y_{\max}^2 = \cos^2\theta_{\min} , \quad (23.52)$$

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 (23.9b) allows the Carter constant  $\mathcal{K}$ , or equivalently the Carter integral  $\mathcal{Q}$ , equation (23.30), to be eliminated in favour of the inclination parameter  $\alpha$ , equation (23.52)

$$\mathcal{Q} = \mathcal{K} - (L - aE)^2 = \alpha \left[ a^2(m^2 - E^2) + \frac{L^2}{1 - \alpha} \right] . \quad (23.53)$$

Equation (23.53) is a quadratic equation in  $\alpha$ , so has two roots for  $\alpha$  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  $\alpha \in [0, 1]$ ),

$$\begin{aligned} \mathcal{Q} > 0 & \quad 1 \text{ root in } \alpha \in [0, 1] , \\ \mathcal{Q} < 0 & \quad 2 \text{ roots in } \alpha \in [0, 1] . \end{aligned} \quad (23.54)$$

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  $\alpha$  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 (23.53) 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.

## 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.

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 (23.9a), eliminate the Carter integral  $\mathcal{K}$  in favour of the inclination parameter  $\alpha$  using equation (23.53). Furthermore, eliminate the energy  $E \equiv -\pi_t$  in favour of  $P_t$ , equation (23.4a). 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 (23.4a) for  $P_t$  is

$$P_t = -E + \frac{aL}{R^2} + \frac{qQr}{R^2}. \quad (23.55)$$

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 (23.50), is a quadratic equation in the azimuthal angular momentum  $L \equiv \pi_\phi$ , whose solutions are

$$\frac{L}{\sqrt{1-\alpha}} = \frac{R^2}{r^2 + a^2\alpha} \left[ a\sqrt{1-\alpha} \left( -P_t + \frac{qQr}{R^2} \right) \pm \sqrt{\frac{P_t^2}{\Delta_x} - (r^2 + a^2\alpha)} \right]. \quad (23.56)$$

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 (23.56) 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 (23.50), yields a quartic equation

$$\boxed{p_0 + p_1 P + p_2 P^2 + p_3 P^3 + p_4 P^4 = 0}, \quad (23.57)$$

for the dimensionless quantity  $P$  (not to be confused with  $P_t$  or  $P_x$ ) defined by

$$P \equiv -\frac{P_t}{R^2 \Delta_x}. \quad (23.58)$$

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 (23.57) are

$$p_0 \equiv r^2(r^2 + a^2\alpha)^2 , \quad (23.59a)$$

$$p_1 \equiv -2qQr(r^2 - a^2\alpha)(r^2 + a^2\alpha) , \quad (23.59b)$$

$$p_2 \equiv -2r^2(r^2 + a^2\alpha)(r^2 - 3Mr + 2Q^2 + a^2\alpha + a^2\alpha M/r) + q^2Q^2(r^2 - a^2\alpha)^2 , \quad (23.59c)$$

$$p_3 \equiv 2qQr(r^2 - a^2\alpha)(r^2 - 3Mr + 2Q^2 + 2a^2 - a^2\alpha + a^2\alpha M/r) , \quad (23.59d)$$

$$\begin{aligned} p_4 \equiv & [r^6 - 6Mr^5 + (9M^2 + 4Q^2 + 2a^2\alpha)r^4 - 4M(3Q^2 + a^2)r^3 \\ & + (4Q^4 - 6a^2\alpha M + 4a^2Q^2 + a^4\alpha^2)r^2 + 2a^2\alpha(2Q^2 + 2a^2 - a^2\alpha)Mr + a^4\alpha^2M^2] . \end{aligned} \quad (23.59e)$$

The quartic (23.57) is the condition for an orbit at radius  $r$  to be circular. Physical solutions  $P$  must be real. Barring degenerate cases, the quartic (23.57) has either zero, two, or four real solutions at any one radius  $r$ . Numerically, it is better to solve the quartic (23.57) 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 (23.57) as a function of radius are illustrated in Figure 23.1 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 (23.57),

$$\begin{aligned} \frac{L}{\sqrt{1-\alpha}} &= \frac{1}{2a\sqrt{1-\alpha}} \left[ R^2P^{-1} - \frac{qQ(r^2 - a^2)}{r} - (R^2 - 3Mr + 2Q^2 + a^2M/r)P \right] \\ &= \pm \frac{1}{r^2 + a^2\alpha} \sqrt{l_{-1}P^{-1} + l_0 + l_1P + l_2P^2} , \end{aligned} \quad (23.60a)$$

$$E = \frac{1}{2} [P^{-1} + qQ/r + (1 - M/r)P] , \quad (23.60b)$$

$$\frac{d^2P_x^2}{dr^2} = \frac{2}{(r^2 + a^2\alpha)^2} (q_{-1}P^{-1} + q_0 + q_1P + q_2P^2) , \quad (23.60c)$$

where the coefficients  $l_i$  and  $q_i$  are

$$l_{-1} \equiv qQrR^2(r^2 + a^2\alpha) , \quad (23.61a)$$

$$l_0 \equiv -R^2(r^2 + a^2\alpha)(2Mr - Q^2) - q^2Q^2(r^4 - a^4\alpha) , \quad (23.61b)$$

$$\begin{aligned} l_1 \equiv & -qQr [2r^4 - 5Mr^3 + 3(Q^2 + a^2)r^2 - a^2(1+\alpha)Mr + a^2(Q^2 + \alpha Q^2 + a^2 - a^2\alpha) \\ & + 3a^4\alpha M/r - a^4\alpha(Q^2 + a^2)/r^2] , \end{aligned} \quad (23.61c)$$

$$l_2 \equiv [3Mr^3 - 2Q^2r^2 + a^2(1+\alpha)Mr - a^2(1+\alpha)Q^2 - a^4\alpha M/r] R^4 \Delta_x , \quad (23.61d)$$

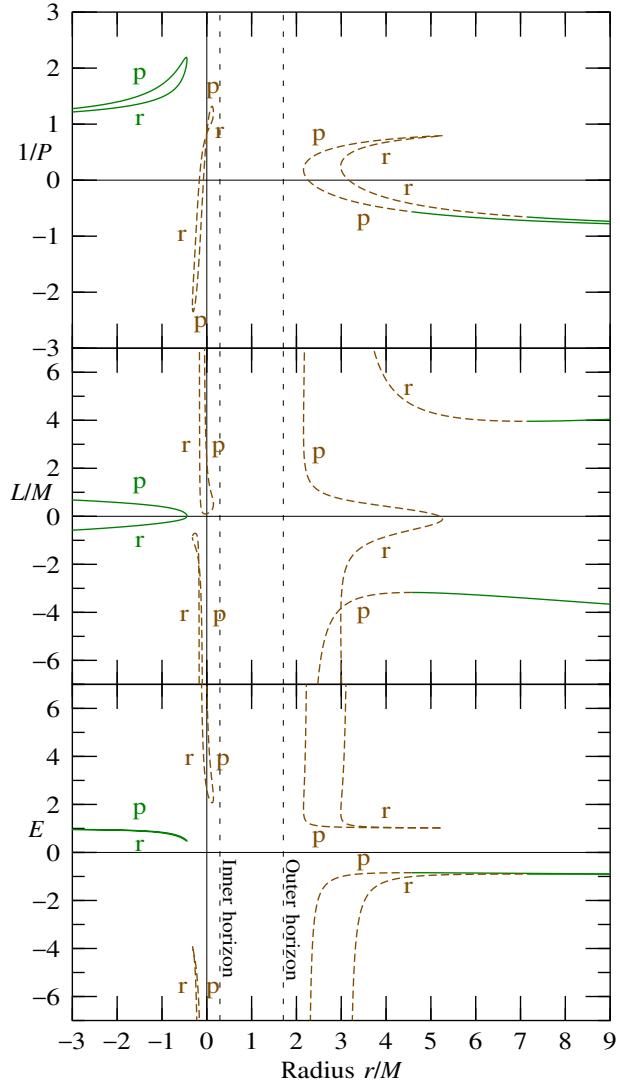


Figure 23.1 Values of  $1/P$ , equation (23.58), angular momentum  $L$ , and energy  $E$ , for circular orbits at radius  $r$  of a charged particle about a Kerr-Newman black hole. The parameters are illustrative: the black hole has spin parameter  $a/M = 0.5$  and charge  $Q/M = 0.5$ , and the particle 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 (23.57); 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 (23.66); 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  $qQ < mM$ .

and

$$q_{-1} \equiv 2qQr(r^2 - a^2\alpha)(r^2 + a^2\alpha) , \quad (23.62a)$$

$$q_0 \equiv -4(r^2 + a^2\alpha)(Mr^3 - Q^2r^2 - a^2\alpha Mr) - q^2Q^2(r^2 - a^2\alpha)^2 , \quad (23.62b)$$

$$q_1 \equiv -qQr [r^4 - 4Mr^3 + 3(Q^2 + a^2 - 2a^2\alpha)r^2 + 12a^2\alpha Mr - a^2\alpha(6Q^2 + 6a^2 - a^2\alpha) - a^4\alpha^2(Q^2 + a^2)/r^2] , \quad (23.62c)$$

$$q_2 \equiv (3Mr^3 - 4Q^2r^2 - 6a^2\alpha Mr - a^4\alpha^2 M/r) R^4 \Delta_x . \quad (23.62d)$$

Equations (23.60) determine the values of  $L$ ,  $E$ , and  $d^2P_x^2/dr^2$  uniquely for any given root  $P$  of the quartic (23.57). The expression on the second line of equations (23.60a) for  $L$  is equivalent to the expression on the first line, the sign of the second expression being chosen to agree with the first expression. The second expression has the virtue of remaining well-behaved in the limit  $a \rightarrow 0$  or  $1 - \alpha \rightarrow 0$ . The two expressions (23.60a) for  $L$  are moreover equivalent to the expression (23.56) with one of the two choices of sign in the latter.

For non-zero  $a$ , the reality of a solution  $P$  of the quartic (23.57) is a necessary and sufficient condition for a corresponding circular orbit to exist. In particular, the argument of the square root in the expression (23.60a) for  $L$  is guaranteed to be positive. For zero  $a$ , however, the quartic (23.60a), which reduces in this case to the square of a quadratic, §23.19, 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 (23.60a) 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 (23.60b) 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 must also have positive energy  $E$  in the Universe region outside the horizon. As discussed in §23.16, equation (23.84), 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 (23.57), the derivative  $dP_y^2/d\alpha$  is

$$\frac{dP_y^2}{d\alpha} = \frac{1}{r^2 + a^2\alpha} (k_{-1}P^{-1} + k_0 + k_1P + k_2P^2) , \quad (23.63)$$

where the coefficients  $k_i$  are

$$k_{-1} \equiv -qQr(r^2 + a^2\alpha) , \quad (23.64a)$$

$$k_0 \equiv (r^2 + a^2\alpha)(2Mr - Q^2) + q^2Q^2(r^2 - a^2\alpha) , \quad (23.64b)$$

$$k_1 \equiv \frac{qQ}{r} [r^2(2r^2 + 3a^2 - 5Mr + 3Q^2) - a^2\alpha(2r^2 + a^2 - 3Mr + Q^2)] , \quad (23.64c)$$

$$k_2 \equiv -(3Mr - 2Q^2 - a^2\alpha M/r) R^4 \Delta_x , \quad (23.64d)$$

### 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,

$$P \leftrightarrow -P, \quad Q \leftrightarrow -Q, \quad L \leftrightarrow -L, \quad E \leftrightarrow -E. \quad (23.65)$$

The radial flip exchanges Universes and Antiverses,

$$r \leftrightarrow -r, \quad M \leftrightarrow -M, \quad Q \leftrightarrow -Q. \quad (23.66)$$

### 23.11.2 Prograde and retrograde orbits

At zero spin,  $a = 0$ , the quartic (23.57) reduces to the square of a quadratic (this is the Reissner-Nordström case considered in §23.19). 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 (23.65); 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 23.1, 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 (23.57) for a circular orbit vanish, and the quartic reduces to a quadratic in  $P^2$ . Solving the quadratic yields two possible solutions

$$1/P^2 = \frac{F_\pm}{r^2 + a^2\alpha}, \quad (23.67)$$

where  $F_\pm$  are

$$F_\pm \equiv r^2 - 3Mr + 2Q^2 + a^2\alpha(1 + M/r) \pm 2a\sqrt{(1-\alpha)(Mr - Q^2 - a^2\alpha M/r)}. \quad (23.68)$$

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 non-zero spin  $a \neq 0$ , the necessary

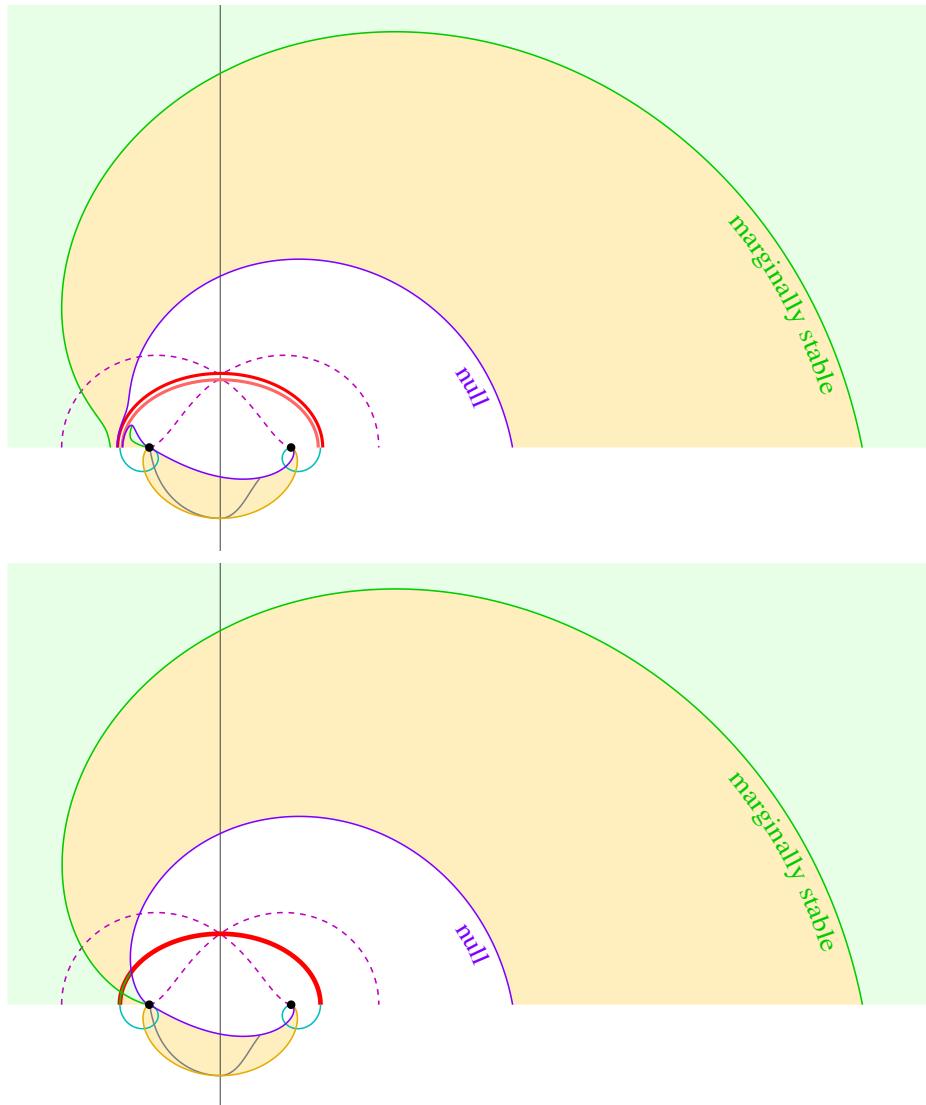


Figure 23.2 Location of stable (shaded green) and unstable (shaded amber) circular orbits in a Kerr black hole with spin (top) slightly sub-extremal ( $a = 0.99M$ ), and (bottom) extremal ( $a = M$ ). The plotted latitude of each circular orbit is its inclination, the maximum latitude reached by the orbit. Null (violet), marginally stable (green), and constant-latitude (grey; inside the Antiverse, at  $r < 0$ ) circular orbits are marked. 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), sisyntubes (cyan), and singularities (black) are shown as in Figures 9.1 and 9.3.

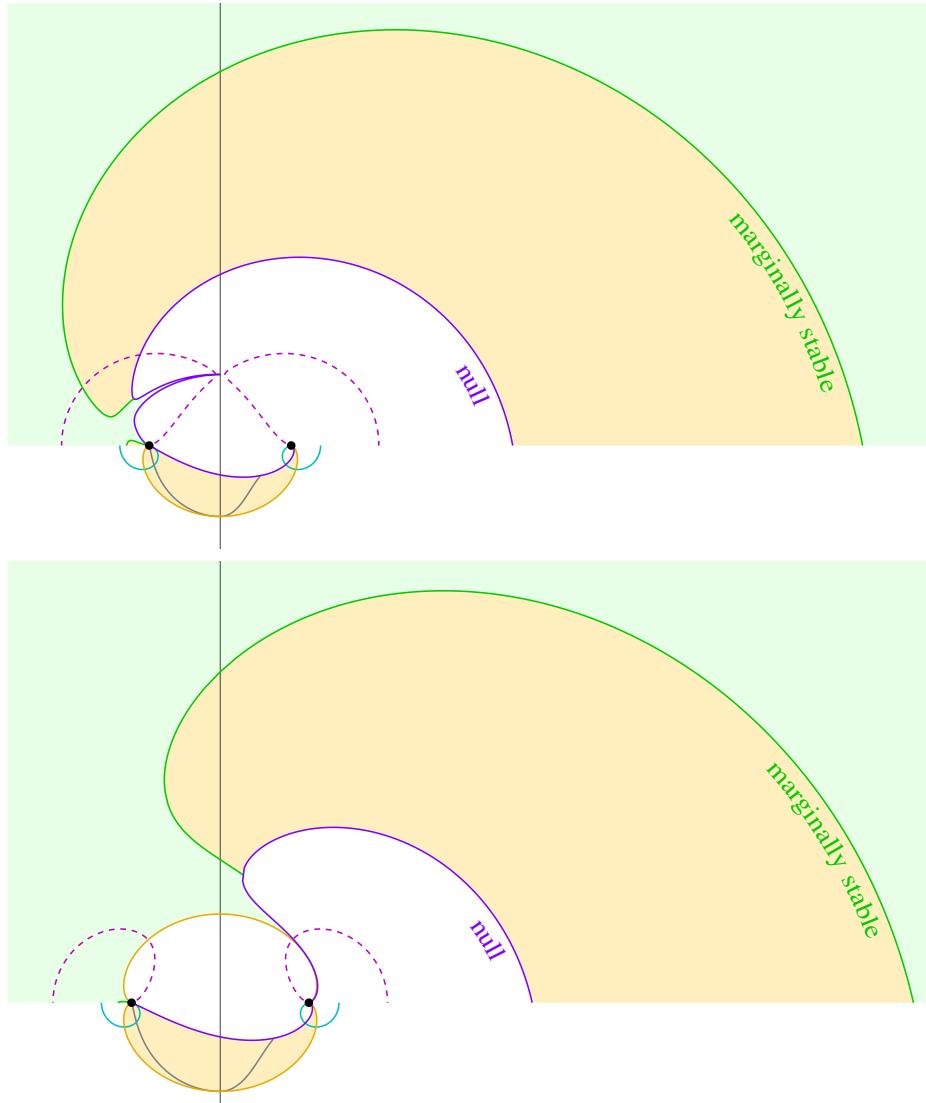


Figure 23.3 As Figure 23.2, but for a Kerr black hole with spin (top) slightly super-extremal ( $a = 1.001M$ ), and (bottom) super-extremal ( $a = 1.25M$ ). Ergospheres, sisyntubes, and singularities are shown as in Figure 9.4.

and sufficient condition for the existence of a circular orbit is that  $P$  be real, which requires that  $F_{\pm}$  be real and positive,

$$Mr - Q^2 - a^2 \alpha M/r \geq 0 \quad \text{and} \quad F_{\pm} \geq 0 . \quad (23.69)$$

The conditions (23.69) remain necessary and sufficient in the limit  $a = 0$  of zero spin (where  $P$  is real even without the first of the two conditions (23.69)). For zero electric charge, the expressions (23.60) for the angular momentum  $L$ , energy  $E$ , and stability  $d^2P_x^2/dr^2$  of a circular orbit, and the expression (23.63) for the angular derivative  $dP_y^2/d\alpha$  of the angular potential, simplify to

$$\frac{L}{\sqrt{1-\alpha}} = \frac{1}{2a\sqrt{1-\alpha}} [R^2 P^{-1} - (R^2 - 3Mr + 2Q^2 + a^2 M/r)P] \\ = \pm \frac{1}{r^2 + a^2\alpha} \sqrt{l_0 + l_2 P^2}, \quad (23.70a)$$

$$E = \frac{1}{2} [P^{-1} + (1 - M/r)P], \quad (23.70b)$$

$$\frac{d^2P_x^2}{dr^2} = \frac{2}{(r^2 + a^2\alpha)^2} (q_0 + q_2 P^2), \quad (23.70c)$$

$$\frac{dP_y^2}{d\alpha} = \frac{1}{r^2 + a^2\alpha} (k_0 + k_2 P^2). \quad (23.70d)$$

The coefficients  $l_i$  and  $q_i$  from equations (23.61) and (23.62) reduce to

$$l_0 \equiv -R^2(r^2 + a^2\alpha)(2Mr - Q^2), \quad (23.71a)$$

$$l_2 \equiv [3Mr^3 - 2Q^2r^2 + a^2(1+\alpha)Mr - a^2(1+\alpha)Q^2 - a^4\alpha M/r] R^4 \Delta_x, \quad (23.71b)$$

and

$$q_0 \equiv -4(r^2 + a^2\alpha)(Mr^3 - Q^2r^2 - a^2\alpha Mr), \quad (23.72a)$$

$$q_2 \equiv (3Mr^3 - 4Q^2r^2 - 6a^2\alpha Mr - a^4\alpha^2 M/r) R^4 \Delta_x. \quad (23.72b)$$

while the coefficients  $k_i$  from equations (23.64) reduce to

$$k_0 \equiv (r^2 + a^2\alpha)(2Mr - Q^2), \quad (23.73a)$$

$$k_2 \equiv -(3Mr - 2Q^2 - a^2\alpha M/r) R^4 \Delta_x. \quad (23.73b)$$

Figures 23.2 and 23.3 illustrate the location of stable and unstable circular orbits in the Kerr geometry ( $Q = 0$ ) with sub-extremal and extremal spins (Figure 23.2), and super-extremal spin (Figure 23.3). 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 (23.69). The boundaries corresponding to the two conditions (23.69) are marked respectively by solid amber and violet lines in Figures 23.2 and 23.3. As discussed further in §23.13, the boundary of the second of the two conditions (23.69),  $F_{\pm} = 0$ , 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 (23.54). 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 are marked by solid violet lines in Figures 23.2 and 23.3. 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 (23.57) 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 (23.59), factors as

$$\frac{p_4}{p_0} = \frac{F_+ F_-}{(r^2 + a^2\alpha)^2}, \quad (23.74)$$

where  $F_{\pm}$  are defined by equation (23.68). A null circular orbit thus occurs at a radius  $r$  such that

$$F_+ = 0 \quad \text{or} \quad F_- = 0, \quad (23.75)$$

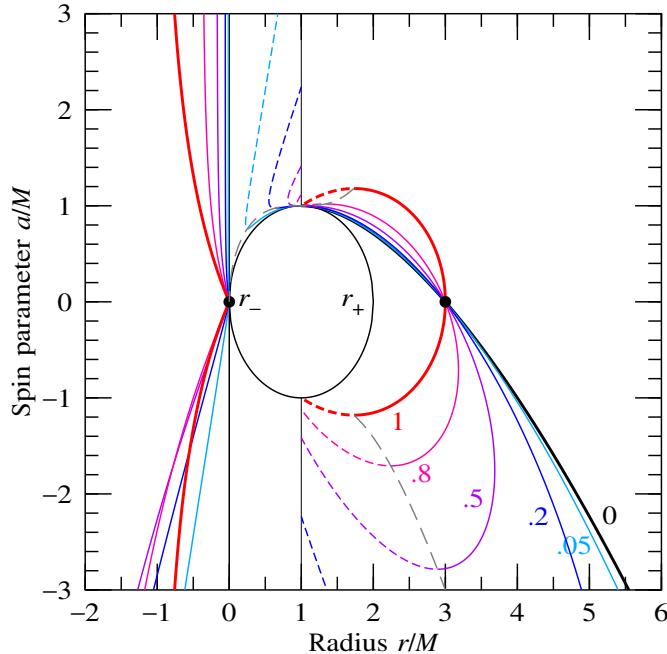


Figure 23.4 Radii of null circular orbits (generalization of the photon sphere) for a Kerr black hole with various spin parameters  $a$ , including super-extremal spin parameters,  $|a/M| > 1$ . Positive and negative  $a/M$  signify prograde ( $F_+ = 0$ ) and retrograde ( $F_- = 0$ ) orbits respectively. Lines are labelled with values of the inclination parameter  $\alpha$ , 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.

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 azimuthal angular momentum  $J \equiv L/E$  per unit energy on the null circular orbit is, from equations (23.60a) and (23.60b) in the limit  $P_t \rightarrow \pm\infty$ ,

$$\frac{J}{\sqrt{1-\alpha}} \equiv \frac{L/E}{\sqrt{1-\alpha}} = \pm \frac{2\sqrt{l_2}}{(r^2 + a^2\alpha)^2(1 - M/r)} . \quad (23.76)$$

This too is independent of the particle charge  $q$ .

Figure 23.4 illustrates the radii of null circular orbits for a Kerr (uncharged) black hole, for various spin and inclination parameters  $a$  and  $\alpha$ , 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  $r = 3M$ , the so-called **photon sphere**. 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. At any inclination there are always 2 null circular orbits at negative radius, one prograde and one retrograde (in Figure 23.4, 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 generally 2 null circular orbits at positive radius, one prograde and one retrograde. If the black hole is sufficiently near extremal, then there are a further 2 null circular orbits at positive radius. If the black hole is sub-extremal ( $|a/M| < 1$ ), then the additional 2 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 orbits exist at large inclinations,  $\alpha > -3 + 2\sqrt{3} \approx 0.464$ ; one orbit is prograde, the other retrograde.

## 23.14 Marginally stable circular orbits

Figure 23.5 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  $\alpha$ . Marginally stable circular orbits are marked by solid green lines in Figures 23.2 and 23.3.

## 23.15 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 23.2 and 23.3. None of these orbits lies inside the retrograde sisytube, so all of them progress forwards, not backwards, in Boyer-Lindquist time  $t$ .

As Figures 23.2 and 23.3 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

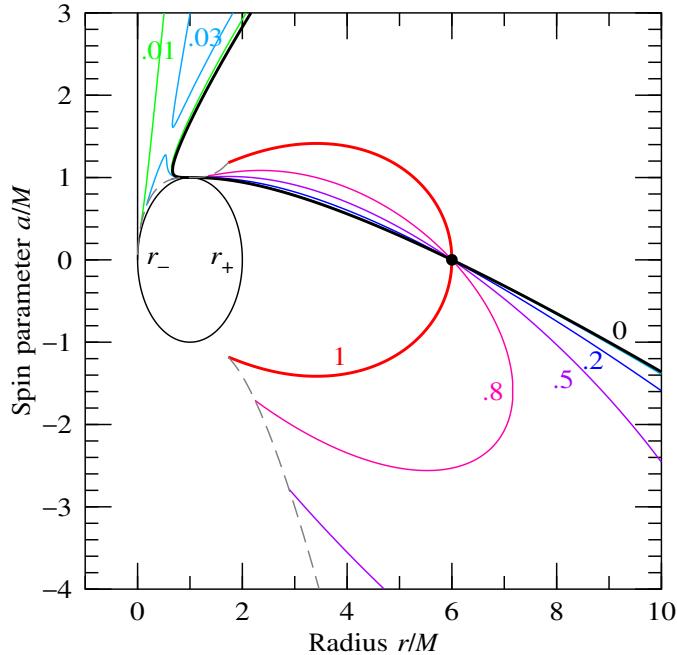


Figure 23.5 Radii of marginally stable circular orbits for a Kerr black hole with various spin parameters  $a$ , including super-extremal spin parameters,  $|a/M| > 1$ . As in Figure 23.4, positive and negative  $a/M$  signify prograde and retrograde orbits respectively. Lines are labelled with values of the inclination parameter  $\alpha$ , varying from equatorial orbits ( $\alpha = 0$ ) to polar orbits ( $\alpha = 1$ ). Long dashed lines, which are the same as in Figure 23.4, mark where marginally stable orbits become null and terminate, examples of which are illustrated in Figures 23.2 and 23.3. 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.

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  $\phi$  has incremented by  $-2\pi$ , constitute Closed Timelike Curves.

### 23.16 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  $F_{\pm}$  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

$$\tilde{P} \equiv -\frac{P_t}{M\sqrt{\Delta_x}} . \quad (23.77)$$

For circular orbits on the horizon of an extremal black hole, where  $r = M$  and  $M^2 = Q^2 + a^2$ , the quartic condition (23.57) reduces to a quadratic

$$\tilde{p}_2 + \tilde{p}_3 \tilde{P} + \tilde{p}_4 \tilde{P}^2 = 0 , \quad (23.78)$$

where the coefficients  $\tilde{p}_i$  are

$$\tilde{p}_2 \equiv 4a^2(1-\alpha)(M^2 + a^2\alpha) + (qQ/M)^2(M^2 - a^2\alpha)^2 , \quad (23.79a)$$

$$\tilde{p}_3 \equiv -2(qQ/M)(M^2 - a^2\alpha)(M^2 + a^2\alpha) , \quad (23.79b)$$

$$\tilde{p}_4 \equiv (M^2 + a^2)^2 - a^2(1-\alpha)(6M^2 + a^2 + a^2\alpha) . \quad (23.79c)$$

The azimuthal angular momentum  $L$ , energy  $E$ , and stability  $d^2P_x^2/dr^2$  of circular orbits on the horizon are

$$\frac{L}{\sqrt{1-\alpha}} = \frac{1}{2a} \left[ (a^2 - M^2)(qQ/M) + (M^2 + a^2)\tilde{P} \right] = \pm \frac{1}{M^2 + a^2\alpha} \sqrt{\tilde{l}_0 + \tilde{l}_1 \tilde{P} + \tilde{l}_2 \tilde{P}^2} , \quad (23.80a)$$

$$E = \frac{1}{2} \left( \tilde{P} + qQ/M \right) , \quad (23.80b)$$

$$\frac{d^2P_x^2}{dr^2} = 0 , \quad (23.80c)$$

where the coefficients  $\tilde{l}_i$  are

$$\tilde{l}_0 \equiv -(M^2 + a^2)^2(M^2 + a^2\alpha) - q^2Q^2(M^4 - a^4\alpha) , \quad (23.81a)$$

$$\tilde{l}_1 \equiv qQM(M^2 + a^2)(M^2 + a^2\alpha) , \quad (23.81b)$$

$$\tilde{l}_2 \equiv M^2(M^2 + a^2)^2 . \quad (23.81c)$$

Circular orbits on the horizon are always marginally stable, equation (23.80c). 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  $\alpha$ . For neutral particles,  $qQ = 0$ , the coefficient  $\tilde{p}_3$  vanishes, and  $\tilde{p}_2$  is positive, so the quadratic (23.78) has a real root  $\tilde{P}$  only as long as  $\tilde{p}_4$  is negative. This imposes the condition that

$$\alpha \leq \frac{M(4a^2 - M^2)}{a^2(3M + 2\sqrt{2M^2 + a^2})} . \quad (23.82)$$

For a Kerr (uncharged) black hole, where  $a = M$ , the inclination must be less than

$$\alpha \leq -3 + 2\sqrt{3} = 0.464 , \quad (23.83)$$

as illustrated in the bottom panel of Figure 23.2.

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,

$$E = \frac{qQ}{3M} + \sqrt{\frac{1}{3} + \left(\frac{qQ}{3M}\right)^2}. \quad (23.84)$$

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  $-\infty$ ,

$$E \rightarrow 0 \quad \text{as } qQ \rightarrow -\infty. \quad (23.85)$$

The smallest possible energy for a circular orbit for a neutral particle,  $q = 0$ , is

$$E = \frac{1}{\sqrt{3}}. \quad (23.86)$$

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 23.6 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 is that, as discussed §23.16, an extremal black hole may have circular orbits at its horizon; but these orbits have  $P_t = 0$ , and are not null.

### 23.17 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.

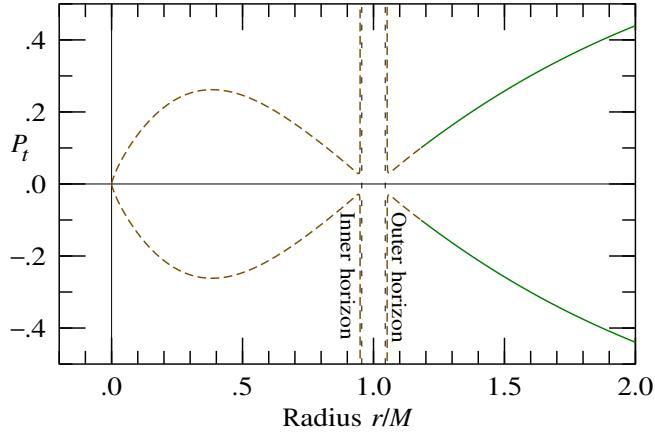


Figure 23.6 Values of the Hamilton-Jacobi parameter  $P_t$  for circular orbits at radius  $r$  in the equatorial plane of a near-extremal Kerr black hole, with black hole spin parameter  $a = 0.999M$ . The diagram illustrates that as the orbital radius  $r$  approaches the horizon,  $P_t$  first approaches zero, but then increases sharply to infinity, corresponding to null circular orbits. In the case of an exactly extremal black hole,  $P_t$  goes 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.

For circular orbits in the equatorial plane,  $\alpha = 0$ , of an uncharged black hole,  $Q = 0$ , the solution (23.67) for  $P$  simplifies to

$$1/P^2 = \frac{F_{\pm}}{r^2} \quad (23.87)$$

where  $F_{\pm}$ , equation (23.68), reduce to

$$F_{\pm} \equiv r^2 - 3Mr \pm 2a\sqrt{Mr}, \quad (23.88)$$

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  $r \rightarrow 4M$  (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 23.6.

### 23.17.1 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 (23.70c) to the case of equatorial

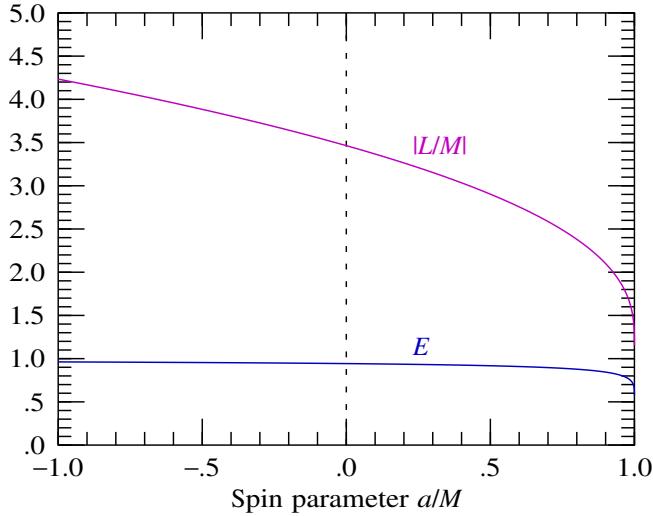


Figure 23.7 Energy  $E$  and azimuthal angular momentum  $|L/M|$  of circular orbits on the ISCO of a Kerr black hole as a function of the spin parameter  $a/M$ . The angular momentum  $L$  is positive for  $a > 0$  (prograde), negative for  $a < 0$  (retrograde).

orbits,  $\alpha = 0$ , and zero black hole charge,  $Q = 0$ , yields the condition of marginal stability

$$r^2 - 6Mr - 3a^2 \pm 8a\sqrt{Mr} = 0. \quad (23.89)$$

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 the ISCO are

$$E = \sqrt{1 - \frac{2M}{3r}} \quad (23.90a)$$

$$L = \pm \frac{2M}{3\sqrt{3}} \sqrt{\frac{12r}{M} - 7 + 4\sqrt{\frac{3r}{M} - 2}}, \quad (23.90b)$$

which are illustrated in Figure 23.7.

### 23.18 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

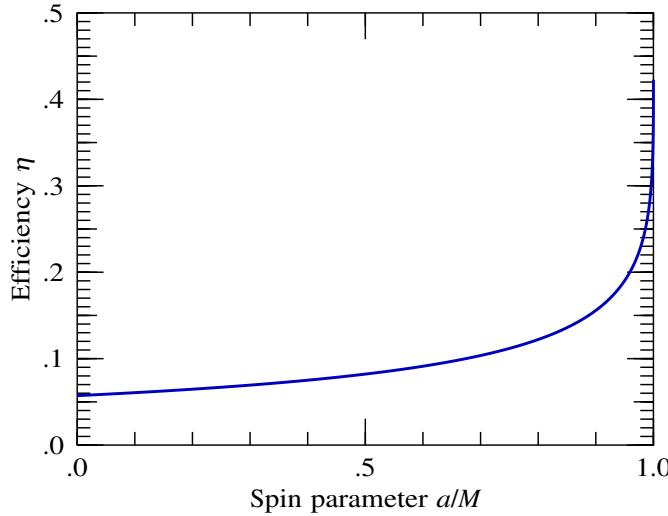


Figure 23.8 Efficiency of accretion on to a Kerr black hole, equation (23.91). The efficiency varies from  $\eta = 0.06$  at  $a = 0$  to  $\eta = 0.42$  at  $a = M$ .

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 (23.90a), a particle must lose fractional energy

$$\eta \equiv \frac{E_\infty - E_{\text{ISCO}}}{E_\infty} = 1 - \sqrt{1 - \frac{2M}{3r}} . \quad (23.91)$$

In the simple thin-disk model, particles in the disk lose energy by emitting radiation, which astronomers can detect. The fractional energy  $\eta$  represents the efficiency with which rest mass energy is converted to radiation. The efficiency  $\eta$ , illustrated in Figure 23.8, 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.

### 23.18.1 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 23.7, 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 23.7 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 (23.3) 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  $\tau$  experienced by Icarus is

$$\frac{dt}{d\tau} = \frac{1}{\rho^2} \left( -\frac{P_t}{\Delta_x} + \frac{\omega_y P_\phi}{\Delta_y} \right) = \frac{1}{r^2 + a^2 \cos^2\theta} [R^2 P + a(L - aE \sin^2\theta)] . \quad (23.92)$$

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 (23.92) dominates. For large  $P$ , equation (23.70b) shows that

$$P \approx \frac{2E}{1 - M/r} . \quad (23.93)$$

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  $E = 1$ ,

$$P \approx \frac{2}{1 - M/r} . \quad (23.94)$$

Equation (23.94) 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 (23.92) simplifies to

$$\frac{dt}{d\tau} = \frac{r \pm a\sqrt{M/r}}{\sqrt{F_{\pm}}} . \quad (23.95)$$

For  $E = 1$ , equation (23.95) becomes

$$\frac{dt}{d\tau} = 1 + \frac{2}{\sqrt{1-a/M}(1+\sqrt{1-a/M})} \approx \frac{2}{\sqrt{1-a/M}} . \quad (23.96)$$

At the Thorne limit  $a = 0.998M$ , the time dilation factor is

$$\frac{dt}{d\tau} = 44 . \quad (23.97)$$

**Exercise 23.8 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 (23.92) on the prograde equatorial ISCO is, to lowest order in  $1 - a/M$ ,

$$\frac{dt}{d\tau} \approx \frac{2^{4/3}}{\sqrt{3}(1-a/M)^{1/3}} . \quad (23.98)$$

To achieve the required time dilation factor requires, to lowest order,

$$1 - a/M \approx \frac{16}{3\sqrt{3}(dt/d\tau)^3} , \quad (23.99)$$

which for  $dt/d\tau \approx 60,000$  is

$$1 - a/M \approx 10^{-14} , \quad (23.100)$$

or  $a \approx 0.99999999999999M$ . This is much closer to extremality than the Thorne limit. At the Thorne limit  $a = 0.998M$ , the time dilation factor is

$$\frac{dt}{d\tau} = 11 . \quad (23.101)$$

### 23.19 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 non-rotating black hole,  $a = 0$ . For a non-rotating 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  $L_{\text{tot}}$ . For non-equatorial orbits, the relation between azimuthal and total angular momentum is  $L = \pm\sqrt{1-\alpha}L_{\text{tot}}$ .

For a non-rotating black hole,  $a = 0$ , the quartic condition (23.57) for a circular orbit of a particle of rest mass  $m = 1$  and electric charge  $q$  reduces to the square of a quadratic,

$$r^2 - qQrP - (r^2 - 3Mr + 2Q^2)P^2 = 0 . \quad (23.102)$$

Solving the quadratic (23.102) yields two solutions

$$1/P = \frac{qQ}{2r} \pm \sqrt{1 - \frac{3M}{r} + \frac{2Q^2}{r^2} + \frac{q^2Q^2}{4r^2}} . \quad (23.103)$$

The sign of  $P$ , equation (23.58), 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 (23.103)  $P$  of the quadratic,

$$L = \sqrt{P^2 R^4 \Delta_x - r^2} , \quad (23.104a)$$

$$E = \frac{PR^4 \Delta_x}{r^2} + \frac{qQ}{r} , \quad (23.104b)$$

$$\frac{d^2P_x^2}{dr^2} = 2(r^2 - 6Mr + 5Q^2 + q^2Q^2) - 2\left(1 - \frac{6M}{r} + \frac{6Q^2}{r^2}\right)P^2R^2\Delta_x . \quad (23.104c)$$

For massless particles, circular orbits occur where the solution (23.103) for  $1/P$  vanishes, which occurs when

$$r^2 - 3Mr + 2Q^2 = 0 , \quad (23.105)$$

independent of the charge  $q$  of the particle. The condition (23.105) is consistent with the Kerr-Newman condition for a null circular orbit, the vanishing of  $F_\pm$  given by equation (23.68). However, for Kerr-Newman, the argument of the square root on the right hand side of equation (23.68) 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 (23.105) 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

$$r = \frac{3M}{2} \left(1 + \sqrt{1 - \frac{8Q^2}{9}}\right) \quad \text{for } 0 \leq Q^2 \leq M^2 . \quad (23.106)$$

In (unphysical) cases  $Q^2 < 0$  or  $M^2 < Q^2 \leq (9/8)M^2$ , both solutions of equation (23.105) are valid.

## 23.20 Hypersurface-orthogonal congruences

The Hamilton-Jacobi separated solution makes it possible to construct congruences (§18.1) of timelike or null geodesics in the  $\Lambda$ -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 §§18.6 and 18.7.

It should be remarked from the outset that the principal null congruences of the  $\Lambda$ -Kerr-Newman geometry are *not* hypersurface-orthogonal, except in the special case of spherical symmetry.

### 23.20.1 Hypersurface-orthogonality condition

As discussed in §18.6, a timelike hypersurface-orthogonal congruence is constructed by picking an arbitrary spacelike 3-dimensional hypersurface 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 §18.7, 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  $\Lambda$ -Kerr-Newman geometry is the separated expression (22.20) for the particle action  $S$ , with generalized momenta  $\pi_x$  and  $\pi_y$  coming from equations (23.4b) and (23.4c),

$$S = \int \left( -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy \right) . \quad (23.107)$$

The Hamilton-Jacobi parameters  $P_x$  and  $P_y$ , equations (23.9), depend on the particle mass  $m$  and on the constants of motion  $C_\alpha \equiv \{E, L, \mathcal{K}\}$ . The mass  $m$ , which can be scaled to a constant without loss of generality, may be either positive or zero. If the mass is zero, then the energy  $E$  can be scaled to a constant without loss of generality. If the constants of motion  $C_\alpha$  are constant everywhere, then the integrand on the right hand side of equation (23.107) is manifestly integrable, being a sum of 4 terms each depending on only one of each of the 4 coordinates  $t, \phi, x, y$ . However, equation (23.107) remains valid for a general congruence of geodesics, the integral being understood as being taken along the geodesics. The constants of motion  $C_\alpha$  are by definition constant along each geodesic, but may vary (smoothly) from one geodesic to another. Since the integral (23.107) is along geodesics, and the constants of motion are constant along geodesics, the action integrates to the separated expression

$$S = -E(t - t_i) + L(\phi - \phi_i) - \int_{x_i} \frac{P_x}{\Delta_x} dx + \int_{y_i} \frac{P_y}{\Delta_y} dy , \quad (23.108)$$

in which the constants of motion  $C_\alpha$  are held constant in the integrals over  $x$  and  $y$  even when those constants vary across geodesics. The quantities  $x_i^\mu \equiv \{t_i, x_i, y_i, \phi_i\}$  are constant along geodesics, but can be chosen arbitrarily across geodesics. For example,  $x_i^\mu$  could be chosen to be constants (perhaps zero), or they could be chosen to be the values of the coordinates  $x^\mu$  on some initial 3-dimensional hypersurface, or whatever other choice may be desirable.

Derivatives of the action  $S$  with respect to the constants of motion yield comoving spatial coordinates  $X^\alpha \equiv \{X^E, X^L, X^K\}$  defined by (see §23.21 for the special case  $P_y = 0$ )

$$X^E \equiv \frac{\partial S}{\partial E} = \int \left( -dt + \frac{P_t dx}{P_x \Delta_x} + \frac{\omega_y P_\phi dy}{P_y \Delta_y} \right) = -(t - t_i) + \int_{x_i} \frac{P_t dx}{P_x \Delta_x} + \int_{y_i} \frac{\omega_y P_\phi dy}{P_y \Delta_y} , \quad (23.109a)$$

$$X^L \equiv \frac{\partial S}{\partial L} = \int \left( d\phi - \frac{\omega_x P_t dx}{P_x \Delta_x} - \frac{P_\phi dy}{P_y \Delta_y} \right) = \phi - \phi_i - \int_{x_i} \frac{\omega_x P_t dx}{P_x \Delta_x} - \int_{y_i} \frac{P_\phi dy}{P_y \Delta_y} , \quad (23.109b)$$

$$X^K \equiv \frac{\partial S}{\partial K} = \int \left( \frac{dx}{2P_x} + \frac{dy}{2P_y} \right) = \int_{x_i} \frac{dx}{2P_x} + \int_{y_i} \frac{dy}{2P_y} . \quad (23.109c)$$

As in the action (23.107), the integrals in the definitions (23.109) are to be understood as being taken along geodesics. And as in the integrated action (23.108), because the constants of motion are constant along geodesics, the coordinates  $X^\alpha$  integrate to the separated expressions on the rightmost sides of equations (23.109) with the constants of motion held constant even when those constants vary across geodesics. As is evident from equations (23.10) and (23.11), the comoving coordinates  $X^\alpha$  are constant along geodesics,

$$dX^\alpha = 0 . \quad (23.110)$$

By adjusting the quantities  $x_i^\mu$ , which are constant along geodesics but which can vary arbitrarily across geodesics, values of the comoving coordinates may be chosen arbitrarily on an initial 3-dimensional hypersurface.

The total derivative of the action (23.107) is

$$dS = -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy + X^\alpha dC_\alpha , \quad (23.111)$$

in which the last term  $X^\alpha dC_\alpha$  takes into account the possible variation of constants of motion across geodesics. Timelike geodesics are orthogonal to hypersurfaces of constant action if

$$p_\mu = \frac{\partial S}{\partial x^\mu} . \quad (23.112)$$

Equation (23.112) is equivalent to the condition that the total derivative (23.111) of the action is

$$dS = -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy . \quad (23.113)$$

Comparing equations (23.111) and (23.113) shows that a timelike congruence is hypersurface-orthogonal if and only if

$$X^\alpha dC_\alpha = 0 . \quad (23.114)$$

The condition (23.113), or equivalently (23.114), can continue to be imposed in the massless limit, where the congruence becomes null. However, for a null congruence the condition (23.113) need not be equivalent to the condition (23.112). As discussed in §18.7, in the massless limit the momentum is not only orthogonal but also tangent to the limiting null hypersurface, and equation (23.112) need be imposed only over each 3-dimensional null hypersurface, not over the entire 4-dimensional spacetime.

For massive particles,  $m \neq 0$ , the action  $S$  can be taken to be zero over an initial 3-dimensional hypersurface, which amounts to choosing  $x_i^\mu$  in equation (23.108) to be the values of the coordinates on the initial hypersurface. The comoving coordinates defined by equations (23.109) then vanish identically,  $X^\alpha = 0$ , and the hypersurface-orthogonal condition (23.114) is satisfied, regardless of whether the constants of motion  $C_\alpha$  vary across geodesics.

For massless particles,  $m = 0$ , the energy  $E$  in the integrated action (23.108) can be scaled to  $\pm 1$  without loss of generality, with  $+1$  in the Universe, and  $-1$  in the Parallel Universe. The hypersurface-orthogonal condition (23.114) can then be satisfied by arranging that  $X^L = X^K = 0$ , which can be accomplished by choosing  $x_i$ ,  $y_i$ , and  $\phi_i$  to be the values of the coordinates on the initial 3-dimensional hypersurface. The action  $S_i$  on the initial 3-dimensional hypersurface is then

$$S_i = -E(t - t_i) . \quad (23.115)$$

The freedom to adjust the parameter  $t_i$  arbitrarily allows the initial action to be chosen at will over the initial hypersurface.

Although for a null congruence the initial 3-dimensional hypersurface and the initial action on it can be chosen arbitrarily, time translation symmetry picks out preferred choices. Time translation symmetry is respected if the initial 3-dimensional hypersurface on which the action is defined is a tensor product of the time axis and any 2-dimensional spatial surface, and the action is taken to be constant on the 2-dimensional surfaces at constant time  $t$ . Setting  $t_i = 0$  yields the action  $S_i$  on the initial 3-dimensional hypersurface as

$$S_i = -Et . \quad (23.116)$$

### 23.20.2 Hypersurface-orthogonal timelike outgoing and ingoing congruences

Given the symmetries of the  $\Lambda$ -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 (23.122) that describes hypersurface-orthogonal outgoing and ingoing congruences. If the action (23.108) 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

$$T \equiv E(t - t_i) - L(\phi - \phi_i) , \quad (23.117a)$$

$$Z \equiv - \int_{x_i} \frac{P_x}{\Delta_x} dx + \int_{y_i} \frac{P_y}{\Delta_y} dy , \quad (23.117b)$$

which are constructed so that the actions  $S_\pm$  for the outgoing and ingoing congruences are

$$S_\pm = -T \pm Z . \quad (23.118)$$

The quantities  $x_i^\mu$  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 (23.109c) 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  $x_i^\mu$  are the same for both outgoing and ingoing actions, as required by the definitions (23.117) 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 (23.114) 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 (23.113), and the total derivatives of  $T$  and  $Z$  are

$$dT \equiv E dt - L d\phi , \quad (23.119a)$$

$$dZ \equiv -\frac{P_x}{\Delta_x} dx + \frac{P_y}{\Delta_y} dy . \quad (23.119b)$$

The other two coordinates in the line-element of the hypersurface-orthogonal congruences can be taken to be  $\phi$  and either  $X^{\mathcal{K}}$  if  $\mathcal{K}$  is constant, or  $\mathcal{K}$  if  $\mathcal{K}$  varies. Specifically,

$$\frac{dx}{2P_x} + \frac{dy}{2P_y} = dX^{\mathcal{K}} - \frac{\partial X^{\mathcal{K}}}{\partial \mathcal{K}} d\mathcal{K} , \quad (23.120)$$

where

$$-\frac{\partial X^{\mathcal{K}}}{\partial \mathcal{K}} = \int_{x_i} \frac{\Delta_x dx}{4P_x^3} + \int_{y_i} \frac{\Delta_y dy}{4P_y^3} . \quad (23.121)$$

The right hand side of equation (23.120) 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  $T, Z, \phi$ , and either  $X^{\mathcal{K}}$  or  $\mathcal{K}$ , is

$$ds^2 = \frac{\rho^2}{P_x^2 \Delta_y + P_y^2 \Delta_x} \left\{ \Delta_x \Delta_y (-C^2 dT^2 + dZ^2) + 4P_x^2 P_y^2 \begin{cases} dX^{\mathcal{K}} & \text{if } \mathcal{K} \text{ constant} \\ -\frac{\partial X^{\mathcal{K}}}{\partial \mathcal{K}} d\mathcal{K} & \text{if } \mathcal{K} \text{ varies} \end{cases} \right\}^2 + \frac{C^2}{E^2(1 - \omega_x \omega_y)^2} [(P_t^2 \Delta_y - P_\phi^2 \Delta_x) d\phi + (\omega_x P_t \Delta_y - P_\phi \Delta_x) dT]^2 \} , \quad (23.122)$$

where the coefficient  $C$  is

$$C \equiv \left( \frac{P_x^2 \Delta_y + P_y^2 \Delta_x}{P_t^2 \Delta_y - P_\phi^2 \Delta_x} \right)^{1/2} = \left( 1 - \frac{m^2 \rho^2 \Delta_x \Delta_y}{P_t^2 \Delta_y - P_\phi^2 \Delta_x} \right)^{1/2} = \left( 1 + \frac{m^2 \rho^2 \Delta_x \Delta_y}{P_x^2 \Delta_y + P_y^2 \Delta_x} \right)^{-1/2} , \quad (23.123)$$

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 (23.122) 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 (23.122) all vanish along both outgoing and ingoing geodesics. Thus the only part of the line-element (23.122) that varies along geodesics is the part proportional to  $-C^2dT^2+dZ^2$ . The proper times  $\tau$  along the timelike geodesics of the outgoing and ingoing congruences satisfy  $m\tau = T \mp Z$ .

### 23.20.3 Double-null hypersurface-orthogonal congruences

The line-element (23.122) for hypersurface-orthogonal congruences remains well-defined in the limit of zero particle mass,  $m = 0$ . In the massless limit, the coefficient  $C$ , equation (23.123), is unity,

$$C = 1 . \quad (23.124)$$

Moreover, for massless particles the energy  $E$  can be scaled to  $\pm 1$  without loss of generality,

$$|E| = 1 . \quad (23.125)$$

Define outgoing (+) and ingoing (-) null coordinates  $V_{\pm}$  by

$$V_{\pm} \equiv T \pm Z = -S_{\mp} , \quad (23.126)$$

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 23.9.

For massless particles, the line-element (23.122) in terms of the coordinates  $V_+$ ,  $V_-$ ,  $\phi$ , and either  $X^{\mathcal{K}}$  or  $\mathcal{K}$ , takes the double-null form

$$ds^2 = \frac{\rho^2}{P_x^2 \Delta_y + P_y^2 \Delta_x} \left\{ -\Delta_x \Delta_y dV_- dV_+ + 4P_x^2 P_y^2 \begin{cases} dX^{\mathcal{K}} & \text{if } \mathcal{K} \text{ constant} \\ -\frac{\partial X^{\mathcal{K}}}{\partial \mathcal{K}} d\mathcal{K} & \text{if } \mathcal{K} \text{ varies} \end{cases} \right\}^2 + \frac{1}{(1 - \omega_x \omega_y)^2} \left[ (P_t^2 \Delta_y - P_\phi^2 \Delta_x) d\phi + \frac{1}{2} (P_t \Delta_y - P_\phi \Delta_x) (dV_+ + dV_-) \right]^2 . \quad (23.127)$$

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 (23.127) all vanish along both outgoing and ingoing null geodesics. The affine parameter  $\lambda_{\pm}$  along outgoing (+) or ingoing (-) geodesics satisfies

$$d\lambda_{\pm} = \frac{\rho^2 \Delta_x \Delta_y}{2(P_x^2 \Delta_y + P_y^2 \Delta_x)} dV_{\pm} . \quad (23.128)$$

Taking the massless limit of the line-element (23.122) does not preserve the condition (23.112) that the momenta along geodesics are orthogonal to hypersurfaces of constant action throughout the 4-dimensional

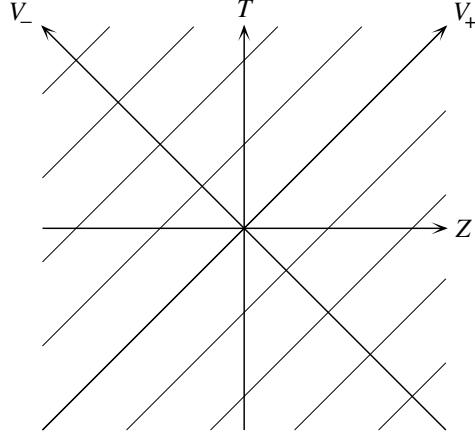


Figure 23.9 Null coordinates  $V_{\pm}$ , equation (23.126), on a spacetime diagram in  $T$  and  $Z$ . The diagonal grid of outgoing null lines, which increase in the outgoing null  $V_+$  direction and are lines of constant ingoing null coordinate  $V_-$ , are lines of constant phase for an outgoing null wave in the geometric optics (high frequency) limit.

spacetime. Rather, the massless limit of the condition (23.112) 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 §18.7.

### 23.21 The principal null and Doran congruences

The geodesics in the principal null congruence, and in the Doran congruence, follow lines of constant latitude  $y$ , and accordingly the Hamilton-Jacobi parameter  $P_y$  vanishes identically,

$$P_y = 0 . \quad (23.129)$$

The integrated action (23.107) then reduces to

$$S = -E(t - t_i) + L(\phi - \phi_i) - \int_{x_i} \frac{P_x}{\Delta_x} dx . \quad (23.130)$$

The total derivative (23.111) of the action reduces to

$$dS = -E dt + L d\phi - \frac{P_x}{\Delta_x} dx + X^{\alpha} dC_{\alpha} , \quad (23.131)$$

in which the coordinates  $X^{\alpha}$  are given by equations (23.109) with the integrals over  $dy$  omitted. As can be checked from equation (23.3), for  $P_y = 0$  the coordinates  $X^{\alpha}$  are generically no longer comoving coordinates, meaning that they no longer satisfy  $dX^{\alpha} = 0$  along geodesics. For  $P_y = 0$ , the coordinates  $X^E$  and  $X^L$  are comoving only in the special case that  $P_{\phi} = 0$ , while  $X^K$  is never a comoving coordinate. Thus for  $P_y = 0$

the hypersurface-orthogonality condition (23.114) can be fulfilled only if  $\mathcal{K}$  is constant and either  $P_\phi = 0$ , or  $E$  and  $L$  are both constant.

For  $P_y$  and  $P_\phi$  both to vanish identically,  $m$  and  $\mathcal{K}$  must both be zero. These conditions are precisely those that define the principal null congruences. For massless particles the energy  $E$  can be scaled to  $\pm 1$  without loss of generality. The condition  $P_\phi = 0$  requires that  $L = E\omega_y$ , so  $L$  cannot be constant. The hypersurface-orthogonality condition (23.114) then holds provided that the comoving coordinate  $X^L$  is arranged to vanish everywhere. 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 principle 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 23.9 explores the vorticity  $\varpi$  and other components of the extrinsic curvature along the principal null congruences of the  $\Lambda$ -Kerr-Newman geometry.

The other way to satisfy the hypersurface-orthogonality condition (23.114) with  $P_y = 0$  is for  $E$ ,  $L$ , and  $\mathcal{K}$  all to be constant. The condition for  $P_y$  to vanish with all constants of motion constant across geodesics is achieved in the Kerr-Newman geometry (with zero cosmological constant) by and only by the conditions

$$|E| = m , \quad L = 0 , \quad \mathcal{K} = a^2m^2 , \quad (23.132)$$

which are precisely those of the Doran timelike congruence. For the Doran congruence, the time and spatial coordinates  $T$  and  $Z$  defined by equation (23.117) are

$$T \equiv mt , \quad (23.133a)$$

$$Z \equiv - \int_{x_1} \frac{P_x}{\Delta_x} dx . \quad (23.133b)$$

The outgoing and ingoing actions  $S_\pm$  are

$$S_\pm = -T \pm Z = -m \left( t \pm \int \frac{\beta dr}{1 - \beta^2} \right) , \quad (23.134)$$

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 (23.122) of the Doran congruences in hypersurface-orthogonal form is

$$ds^2 = \rho^2 \left[ \frac{\Delta_x}{m^2 \beta^2} (-C^2 dT^2 + dZ^2) + \frac{dy^2}{\Delta_y} + \frac{\Delta_y - \omega_y^2 \Delta_x}{(1 - \omega_x \omega_y)^2} \left( d\phi - \frac{\omega_x \Delta_y - \omega_y \Delta_x}{\Delta_y - \omega_y^2 \Delta_x} \frac{dT}{m} \right)^2 \right] , \quad (23.135)$$

where the coefficient  $C$  is

$$C \equiv \left( \frac{\beta^2 \Delta_y}{\Delta_y - \omega_y^2 \Delta_x} \right)^{1/2} = \left( 1 - \frac{\rho^2 \Delta_x \Delta_y}{\Delta_y - \omega_y^2 \Delta_x} \right)^{1/2} = \left( 1 + \frac{\rho^2 \Delta_x}{\beta^2} \right)^{-1/2} . \quad (23.136)$$

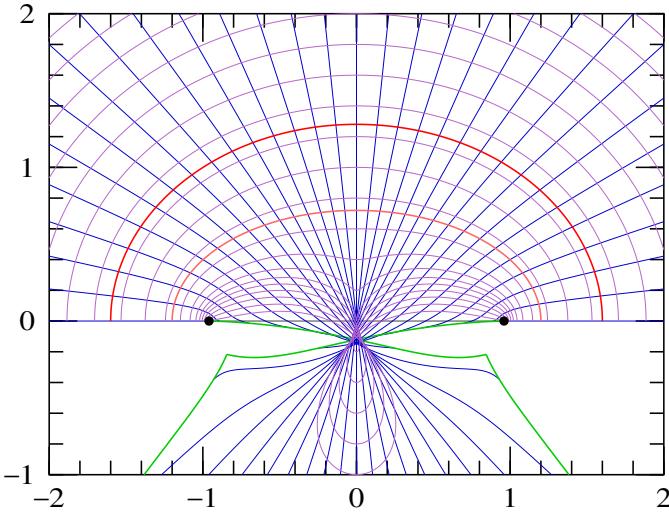


Figure 23.10 Null geodesics (blue lines) and surfaces of constant outgoing and ingoing action (purple lines) in the Pretorius and Israel (1998) double-null hypersurface-orthogonal congruence, for a Kerr black hole with  $a = 0.96M$ . The coordinates are Boyer-Lindquist, and the units are geometric ( $c = G = M = 1$ ). 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. 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. Thick (reddish) lines mark the outer and inner horizons, and filled circles mark the ring singularity.

## 23.22 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 §23.21, there is no double-null hypersurface-orthogonal congruence with  $P_y = 0$ . As discussed in §23.20.2, for  $P_y \neq 0$ , the hypersurface-orthogonality condition (23.114) 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  $\pm 1$ , with  $E = +1$  in the Universe part of the geometry.

In the  $\Lambda$ -Kerr-Newman geometry, motion in latitude extends to the south and north poles only if

$$L = 0 . \quad (23.137)$$

In addition, in order to avoid the geodesics turning around and therefore crossing, the Carter constant must satisfy

$$\mathcal{K} \geq \frac{\omega_y^2}{\Delta_y} \quad (23.138)$$

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 (23.138). 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 (23.138). This is achieved by choosing  $\mathcal{K}$  such that  $P_y$  vanishes at infinite radius, which imposes

$$\mathcal{K} = \left. \frac{\omega_y^2}{\Delta_y} \right|_{\infty} . \quad (23.139)$$

For Kerr-Newman without a cosmological constant, this is

$$\mathcal{K} = a^2 \sin^2 \theta_{\infty} , \quad (23.140)$$

where  $\theta_{\infty}$  is the polar angle of the geodesic at infinite radius. Null geodesics with  $L = 0$  and  $\mathcal{K}$  given by equation (23.139) vary in latitude from a minimum latitude that exhausts the condition (23.138) at infinite radius. The Pretorius-Israel double null line-element is equation (23.127) with  $L = 0$  and non-constant  $\mathcal{K}$  given by equation (23.139). For  $E = 1$  and  $L = 0$ , the time and azimuth Hamilton-Jacobi parameters are  $P_t = -1$  and  $P_{\phi} = -\omega_y$ .

Figure 23.10 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 (23.117). The initial 3-dimensional hypersurface from which the null hypersurfaces project is a sphere of constant radius at infinity for the ingoing congruence, and a sphere of constant radius at the outer horizon for the outgoing congruence. As discussed at the end of §23.20.1, the parameters  $x_i$ ,  $y_i$ , and  $\phi_i$  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  $S_i = -t$ , equation (23.116). The hypersurfaces of constant action, or phase, shown in Figure 23.10 are lines of constant  $Z$  at fixed  $t$ .

### 23.22.1 Application of the null singularity theorem to a $\Lambda$ -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  $\Lambda$ -Kerr-Newman geometry.

The surfaces of constant action in Figure 23.10 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  $\vartheta$  along both outgoing and ingoing congruences. If the dog-leg proposition (§18.9.1) held, then the future would terminate at the caustic surface of first crossing marked in Figure 23.10 by the upper thick (green) line inside the Antiverse ( $r < 0$ ). However, as discussed in Concept question 18.2, 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  $\Lambda$ -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.

**Exercise 23.9 Expansion, vorticity, and shear along the principal null congruences of the  $\Lambda$ -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  $\gamma_v$  and  $\gamma_u$  of the corresponding Newman-Penrose tetrad, equations (36.1).

1. Show that the expansion, vorticity, and shear along the outgoing (upper sign) and ingoing (lower sign) principal null congruences are

$$\vartheta + i\varpi = s \frac{R^2 \sqrt{|\Delta_x|} (\pm \rho_x + i\rho_y)}{\sqrt{2} \rho^3}, \quad \sigma = 0, \quad (23.141)$$

where the overall sign  $s$  is  $+$ , except that  $s$  is  $-$  along the outgoing congruence inside the horizon ( $\Delta_x < 0$ ).

2. Define

$$\lambda_{\pm} \equiv (\rho_y \pm i\rho_x) \sqrt{\Delta_y} \frac{\partial \ln \rho^2 / \partial y}{d\omega_y / dy}, \quad \nu \equiv \ln \left( \frac{\rho \sqrt{\Delta_x}}{1 - \omega_x \omega_y} \right), \quad \mu \equiv \tan^{-1} \left( \frac{\rho_y}{\rho_x} \right). \quad (23.142)$$

Show that the following Lorentz transformation of the tetrad

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda^2 & 1 & \lambda_- & \lambda_+ \\ \lambda_+ & 0 & 1 & 0 \\ \lambda_- & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-\nu} & 0 & 0 & 0 \\ 0 & e^{\nu} & 0 & 0 \\ 0 & 0 & e^{-i\mu} & 0 \\ 0 & 0 & 0 & e^{i\mu} \end{pmatrix} \begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \quad (23.143)$$

brings the tetrad to a form parallel-transported along the outgoing principal null direction  $\gamma_v$  with tetrad-frame connections

$$\Gamma_{kmv} = 0 \quad \text{for all } km , \quad (23.144)$$

and similarly that the Lorentz transformation

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \lambda^2 & -\lambda_+ & -\lambda_- \\ 0 & 1 & 0 & 0 \\ 0 & -\lambda_- & 1 & 0 \\ 0 & -\lambda_+ & 0 & 1 \end{pmatrix} \begin{pmatrix} e^\nu & 0 & 0 & 0 \\ 0 & e^{-\nu} & 0 & 0 \\ 0 & 0 & e^{i\mu} & 0 \\ 0 & 0 & 0 & e^{-i\mu} \end{pmatrix} \begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} \quad (23.145)$$

brings the tetrad to a form parallel-transported along the ingoing principal null direction  $\gamma_u$  with tetrad-frame connections

$$\Gamma_{kmu} = 0 \quad \text{for all } km . \quad (23.146)$$

The rightmost of the two Lorentz transformations in equations (23.143) and (23.145) 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  $\gamma_v$  (respectively ingoing  $\gamma_u$ ) axis unchanged, transforming the remaining axes. The Lorentz-transformed frames are no longer principal null. In the outgoing (respectively ingoing) transformed frame, the non-vanishing components of the Weyl tensor are its spin 0, -1, and -2 (respectively 0, +1, and +2) components.

---

## 24

---

### 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 (1968b), 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,  $\phi$  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  $\Gamma_{azb}$  along the radial directions  $z = t$  and  $x$ , and  $\Gamma_{yaz}$  along the angular directions  $a = y$  and  $\phi$ . 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 ingoing and outgoing 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 (ingoing or outgoing) 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 ingoing and outgoing 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  $\Delta r$  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 particle accelerator

A good way to think conceptually about mass inflation is that it acts like a particle accelerator. The counter-streaming pressure accelerates ingoing and outgoing 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 particle accelerator. 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 ingoing and outgoing streams at the inner horizon?** The inflationary instability is driven by ingoing and outgoing 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 (non-baryonic cold dark matter particles). Calculate the maximum latitude to which particles falling from infinite radius can reach and be either outgoing or ingoing.

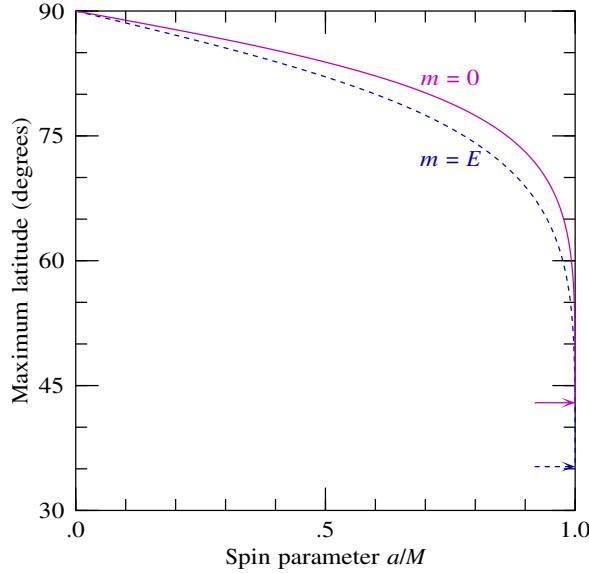


Figure 24.1 Massless (solid) or massive (dashed) collisionless particles falling from infinite radius can reach the inner horizon and be either outgoing or ingoing only up to a certain maximum latitude, shown here. The maximum accessible latitude depends on the spin of the black hole. The maximum latitude varies from  $90^\circ$  (all latitudes are accessible) for a Schwarzschild (non-spinning) black hole, to  $\arcsin \sqrt{-3 + 2\sqrt{3}} = 42.9^\circ$  ( $m = 0$ ) or  $\arcsin \sqrt{1/3} = 35.3^\circ$  ( $m = E$ ), 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  $P_t$  is positive or negative, §23.5. The division between outgoing and ingoing at the inner horizon  $r = r_-$  occurs when  $P_t$  given by equation (23.4a) is zero at the inner horizon, which for a Kerr black hole occurs when the ratio of the azimuthal angular momentum to energy of the particle is

$$\frac{L}{E} = \omega_x(r_-) . \quad (24.10)$$

Particles are outgoing at the inner horizon if their angular momentum per unit energy  $L/E$  is greater than the critical value (24.10), ingoing if less.

2. To reach a given latitude  $\theta$  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  $\infty$  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 be at the transition between outgoing and ingoing at the inner horizon at a given latitude is then

$$\min \left( \frac{P_t^2}{\Delta_x} - m^2 \rho_x^2 \right) \geq \frac{P_\phi^2}{\Delta_y} + m^2 \rho_y^2 , \quad (24.13)$$

where the minimum on the left hand side is over radius  $r$  from  $r_+$  to  $\infty$ . The condition (24.13) along with (24.10) 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 (23.19). 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 (23.4a) is a constant,  $P_t = \pi_{\text{t}} = -E$ . 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 principle 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 (23.25), 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.

---

---

## Concept Questions

1. Why do general relativistic perturbation theory using the tetrad formalism as opposed to the coordinate approach?
2. Why is the tetrad metric  $\gamma_{mn}$  assumed fixed in the presence of perturbations?
3. Are the tetrad axes  $\gamma_m$  fixed under a perturbation?
4. Is it true that the tetrad components  $\varphi_{mn}$  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  $\mathring{g}_{\mu\nu}$ , change under an infinitesimal coordinate gauge transformation?
6. How can the vierbein perturbation  $\varphi_{mn}$  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 (26.76) defines the mass  $M$  of a body as what a distant observer would measure from its gravitational potential. Similarly equation (26.84) 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  $\Psi$  and  $\Phi$  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  $\chi$  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,  $\pm 1$ , and  $\pm 2$ . The spin- $\pm 2$  components describe propagating gravitational waves, while the spin-0 and spin- $\pm 1$  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.

# 25

---

## 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.

### 25.1 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.

### 25.2 Vierbein perturbation

Let the vierbein perturbation  $\varphi_{mn}$  be defined so that the perturbed vierbein is

$$e_m^{\mu} = (\delta_m^n + \varphi_m{}^n)^0 e_n^{\mu} , \quad (25.1)$$

with corresponding inverse

$$e_m^m = (\delta_n^m - \varphi_n{}^m)^0 e_n^{\mu} . \quad (25.2)$$

Since the perturbation  $\varphi_m{}^n$  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  $\varphi_{mn}$  of the vierbein perturbation

$$\varphi_{mn} = \gamma_{nl} \phi_m{}^l . \quad (25.3)$$

In terms of the covariant perturbation  $\varphi_{mn}$ , the perturbed vierbein (25.1) is

$$e_m^{\mu} = (\gamma_{mn} + \varphi_{mn}){}^0 e^{n\mu}, \quad (25.4)$$

The perturbation  $\varphi_{mn}$  can be regarded as a tetrad tensor field defined on the unperturbed background.

### 25.3 Gauge transformations

The vierbein perturbation  $\varphi_{mn}$  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.

### 25.4 Tetrad metric assumed constant

In the tetrad formalism, tetrad axes  $\gamma_m$  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. 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

$$\boxed{\gamma_{mn} = \dot{\gamma}_{mn} = \text{constant}}. \quad (25.5)$$

For example, if the tetrad is orthonormal, then the tetrad metric is constant, the Minkowski metric  $\eta_{mn}$ . However, the tetrad could also be some other tetrad for which the tetrad metric is constant, such as a spin tetrad (§12.1.1), or a Newman-Penrose tetrad (§36.1.1).

## 25.5 Perturbed coordinate metric

The perturbed coordinate metric is

$$\begin{aligned} g_{\mu\nu} &= \gamma_{mn} e^m{}_\mu e^n{}_\nu \\ &= \gamma_{kl} (\delta_m^k - \varphi_m{}^k) \overset{0}{e}{}^m{}_\mu (\delta_n^l - \varphi_n{}^l) \overset{0}{e}{}^n{}_\nu \\ &= \overset{0}{g}_{\mu\nu} - (\varphi_{\mu\nu} + \varphi_{\nu\mu}) . \end{aligned} \quad (25.6)$$

Thus the perturbation of the coordinate metric depends only on the symmetric part of the vierbein perturbation  $\varphi_{mn}$ , not the antisymmetric part

$$\overset{1}{g}_{\mu\nu} = -(\varphi_{\mu\nu} + \varphi_{\nu\mu}) . \quad (25.7)$$

## 25.6 Tetrad gauge transformations

Under an infinitesimal tetrad (Lorentz) transformation, the covariant vierbein perturbations  $\varphi_{mn}$  transform as

$$\varphi_{mn} \rightarrow \varphi_{mn} + \epsilon_{mn} , \quad (25.8)$$

where  $\epsilon_{mn}$  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  $\varphi_{mn}$  is arbitrarily adjustable through an infinitesimal tetrad transformation, while the symmetric part  $\varphi_{mn} + \varphi_{nm}$  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 (25.7) 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  $\varphi_{mn}$ ) altogether, and work only with the tetrad gauge-invariant part (the symmetric part of  $\varphi_{mn}$ )? 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  $\varphi_{mn}$ .

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 25.1 Non-infinitesimal tetrad transformations in perturbation theory?** In perturbation theory, can tetrad gauge transformations be non-infinitesimal?

---

## 25.7 Coordinate gauge transformations

A coordinate gauge transformation is a transformation of the coordinates  $x^\mu$  by an infinitesimal shift  $\epsilon^\mu$

$$x^\mu \rightarrow x'^\mu = x^\mu + \epsilon^\mu. \quad (25.9)$$

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  $\epsilon^\mu$  is, like the vierbein perturbations  $\varphi_{mn}$ , 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  $\epsilon^\mu$  can be regarded as a vector field defined on the unperturbed background. The tetrad components  $\epsilon^m$  of the shift  $\epsilon^\mu$  are

$$\epsilon^m = \delta^m_\mu \epsilon^\mu. \quad (25.10)$$

Physically, the tetrad-frame shift  $\epsilon^m$  is the shift measured in locally inertial coordinates  $\xi^m$ ,

$$\xi^m \rightarrow \xi'^m = \xi^m + \epsilon^m. \quad (25.11)$$

### 25.7.1 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  $A_{\mu\nu\dots}^{\kappa\lambda\dots}(x)$  under a coordinate gauge transformation (25.9) is minus its Lie derivative  $\mathcal{L}_\epsilon$  with respect to the infinitesimal shift  $\epsilon$ ,

$$A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) \rightarrow A'_{\mu\nu\dots}^{\kappa\lambda\dots}(x) = A_{\mu\nu\dots}^{\kappa\lambda\dots}(x) - \mathcal{L}_\epsilon A_{\mu\nu\dots}^{\kappa\lambda\dots}. \quad (25.12)$$

The Lie derivative  $\mathcal{L}_\epsilon A_{\mu\nu\dots}^{\kappa\lambda\dots}$  is given by formula (7.129). Under a coordinate gauge transformation (25.9), 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  $A'_{\mu\nu\dots}^{\kappa\lambda\dots}(x)$  is evaluated at fixed coordinate position  $x$ , not at fixed physical position.

### 25.7.2 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 (25.9). 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 (25.9), a tetrad-frame 4-vector  $A^m$  transforms as

$$A^m(x) \rightarrow A'^m(x) = A^m(x) - \mathcal{L}_\epsilon A^m, \quad (25.13)$$

where the Lie derivative is, equation (7.116),

$$\mathcal{L}_\epsilon A^m = \epsilon^\kappa \frac{\partial A^m}{\partial x^\kappa} \quad \text{not a tetrad tensor}. \quad (25.14)$$

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  $A^{kl\dots}_{mn\dots}$  transforms under a coordinate gauge transformation (25.9) as

$$A^{kl\dots}_{mn\dots}(x) \rightarrow A'^{kl\dots}_{mn\dots}(x) = A^{kl\dots}_{mn\dots}(x) - \mathcal{L}_\epsilon A^{kl\dots}_{mn\dots}, \quad (25.15)$$

where the Lie derivative is

$$\mathcal{L}_\epsilon A^{kl\dots}_{mn\dots} = \epsilon^\pi \partial_\pi A^{kl\dots}_{mn\dots} \quad \text{not a tensor}. \quad (25.16)$$

Again, the change  $-\epsilon^\pi \partial_\pi A^{kl\dots}_{mn\dots}$  is a coordinate tensor (a coordinate scalar), but not a tetrad tensor.

### Concept question 25.2 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 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.

#### 25.7.3 Coordinate gauge transformation of the vierbein

The vierbein  $e_m^\mu$  is a coordinate vector and a tetrad vector. It transforms under a coordinate gauge transformation (25.9) as

$$e_m^\mu(x) \rightarrow e'_m{}^\mu(x) - \mathcal{L}_\epsilon e_m{}^\mu, \quad (25.17)$$

where the Lie derivative of the vierbein is, equation (7.120),

$$\begin{aligned} \mathcal{L}_\epsilon e_m{}^\mu &= -e_m{}^\kappa \frac{\partial \epsilon^\mu}{\partial x^\kappa} + \epsilon^\kappa \frac{\partial e_m{}^\mu}{\partial x^\kappa} \\ &= -\partial_m(e^n{}^\mu \epsilon_n) + \epsilon^k \partial_k e_m{}^\mu \\ &= -e^{n\mu} (\partial_m \epsilon_n - \epsilon^k d_{nkm} + \epsilon^k d_{nmk}) \\ &= -e^{n\mu} [\partial_m \epsilon_n + \epsilon^k (\mathring{\Gamma}_{nkm} - \mathring{\Gamma}_{nmk})]. \end{aligned} \quad (25.18)$$

On the third line the vierbein derivatives have been replaced by  $d_{nkm}$  defined by equation (11.32), while on the fourth line  $\mathring{\Gamma}_{nkm}$  is the torsion-free tetrad-frame connection, defined in terms of the vierbein derivatives by equation (11.54).

### 25.7.4 Coordinate transformation of the vierbein perturbation

According to equation (25.1), the perturbation  $\dot{e}_m^{\mu}$  of the vierbein may be expressed in terms of a covariant vierbein perturbation field  $\varphi_{mn}$ ,

$$\dot{e}_m^{\mu} = \varphi_{mn} \overset{\text{def}}{=} e_m^{\mu}. \quad (25.19)$$

The perturbation induced by a coordinate gauge transformation (25.9) equals the Lie derivative given by equation (25.18),  $\dot{e}_m^{\mu} = \mathcal{L}_{\epsilon} e_m^{\mu}$ . Consequently the vierbein perturbation  $\varphi_{mn}$  transforms under a coordinate gauge transformation (25.9) as

$$\boxed{\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \partial_m \epsilon_n + (\overset{\circ}{\Gamma}_{nkm} - \overset{\circ}{\Gamma}_{nmk}) \epsilon^k}, \quad (25.20)$$

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 25.3 Variation of unperturbed quantities under coordinate gauge transformations?** How does an unperturbed quantity, such as the unperturbed coordinate metric  $\overset{\circ}{g}_{\mu\nu}$ , vary under an infinitesimal coordinate gauge transformation? **Answer.** It doesn't. The variation is considered to be part of the perturbation.

## 25.8 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.

### 25.8.1 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

$$\boxed{\mathbf{w} = \underset{\text{scalar}}{\nabla w_{\parallel}} + \underset{\text{vector}}{\mathbf{w}_{\perp}}}. \quad (25.21)$$

In this context, the term **vector** signifies a 3-vector  $\mathbf{w}_\perp$  that is transverse, that is to say, it has vanishing divergence,

$$\boxed{\nabla \cdot \mathbf{w}_\perp = 0} . \quad (25.22)$$

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 (25.21)

$$\nabla \cdot \mathbf{w} = \nabla^2 w_\parallel . \quad (25.23)$$

The operator  $\nabla^2$  on the right hand side of equation (25.23) is the 3D Laplacian. The solution of equation (25.23) is

$$w_\parallel(\mathbf{x}) = - \int \frac{\nabla' \cdot \mathbf{w}(\mathbf{x}')}{|\mathbf{x}' - \mathbf{x}|} \frac{d^3 x'}{4\pi} . \quad (25.24)$$

The solution (25.24) 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 (25.21) 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 (25.21) is

$$\mathbf{w} = \underset{\text{scalar}}{\mathbf{w}_\parallel} + \underset{\text{vector}}{\mathbf{w}_\perp} . \quad (25.25)$$

### 25.8.2 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  $\nabla$  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  $2\pi$  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)

$$a(\mathbf{k}) = \int a(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x , \quad a(\mathbf{x}) = \int a(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}} \frac{d^3k}{(2\pi)^3} . \quad (25.26)$$

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  $\nabla$  in real space is equivalent to multiplying by  $-ik$  in Fourier space

$$\boxed{\nabla \rightarrow -ik} . \quad (25.27)$$

Thus the decomposition (25.21) of the 3D vector  $\mathbf{w}$  translates into Fourier space as

$$\mathbf{w} = \underset{\text{scalar}}{-i\mathbf{k}} w_{\parallel} + \underset{\text{vector}}{\mathbf{w}_{\perp}} , \quad (25.28)$$

where the vector part  $\mathbf{w}_{\perp}$  satisfies

$$\mathbf{k} \cdot \mathbf{w}_{\perp} = 0 . \quad (25.29)$$

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  $\mathbf{k}$ .

### 25.8.3 Decomposition of a tensor in flat 3D space

Similarly, the 9 components of a  $3 \times 3$  spatial matrix  $h_{ab}$  can be decomposed into 3 scalars, 2 vectors, and 1 tensor:

$$h_{ab} = \underset{\text{scalar}}{\delta_{ab} \phi} + \underset{\text{scalar}}{\nabla_a \nabla_b h} + \underset{\text{scalar}}{\varepsilon_{abc} \nabla_c \tilde{h}} + \underset{\text{vector}}{\nabla_a h_b} + \underset{\text{vector}}{\nabla_b \tilde{h}_a} + \underset{\text{tensor}}{h_{ab}^T} . \quad (25.30)$$

In this context, the term **tensor** signifies a  $3 \times 3$  matrix  $h_{ab}^T$  that is traceless, symmetric, and transverse:

$$h_a^{Ta} = 0 , \quad h_{ab}^T = h_{ba}^T , \quad \nabla_a h_{ab}^T = 0 . \quad (25.31)$$

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,

$$\nabla_a h_a = \nabla_a \tilde{h}_a = 0 . \quad (25.32)$$

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  $3 \times 3$  matrix  $h_{ab}$  is

$$h_a^a = 3\phi + \nabla^2 h . \quad (25.33)$$

# 26

---

## 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

$${}^0 g_{\mu\nu} = \eta_{\mu\nu} . \quad (26.1)$$

In this chapter the tetrad  $\gamma_m$  is taken to be orthonormal, and aligned with the unperturbed coordinate axes  ${}^0 e_\mu$ , so that the unperturbed vierbein is the unit matrix

$${}^0 e_m{}^\mu = \delta_m^\mu . \quad (26.2)$$

Let overdot denote partial differentiation with respect to time  $t$ ,

$$\text{overdot} \equiv \frac{\partial}{\partial t} , \quad (26.3)$$

and let  $\nabla$  denote the spatial gradient

$$\nabla \equiv \frac{\partial}{\partial x} \equiv \nabla_a \equiv \frac{\partial}{\partial x^a} . \quad (26.4)$$

Sometimes it will also be convenient to use  $\nabla_m$  to denote the 4-dimensional spacetime derivative

$$\nabla_m \equiv \left\{ \frac{\partial}{\partial t}, \nabla \right\} . \quad (26.5)$$

### 26.1 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 (26.13), that there are 6 coordinate and tetrad gauge-invariant perturbations, comprising 2 scalars  $\Psi$  and  $\Phi$ , 1 vector  $W_a$  containing 2 degrees of freedom, and 1 tensor  $h_{ab}$  containing 2 degrees of freedom.

The vierbein perturbations  $\varphi_{mn}$  defined by equation (25.1) decompose, §25.8, into 6 scalars, 4 vectors, and 1 tensor, a total of  $6 + 4 \times 2 + 1 \times 2 = 16$  degrees of freedom,

$$\varphi_{00} = \begin{array}{c} \psi \\ \text{scalar} \end{array}, \quad (26.6a)$$

$$\varphi_{0a} = \begin{array}{c} \nabla_a w \\ \text{scalar} \end{array} + \begin{array}{c} w_a \\ \text{vector} \end{array}, \quad (26.6b)$$

$$\varphi_{a0} = \begin{array}{c} \nabla_a \tilde{w} \\ \text{scalar} \end{array} + \begin{array}{c} \tilde{w}_a \\ \text{vector} \end{array}, \quad (26.6c)$$

$$\varphi_{ab} = \begin{array}{c} \delta_{ab} \Phi \\ \text{scalar} \end{array} + \begin{array}{c} \nabla_a \nabla_b h \\ \text{scalar} \end{array} + \varepsilon_{abc} \nabla_c \tilde{h} + \begin{array}{c} \nabla_a h_b \\ \text{vector} \end{array} + \begin{array}{c} \nabla_b \tilde{h}_a \\ \text{vector} \end{array} + \begin{array}{c} h_{ab} \\ \text{tensor} \end{array}. \quad (26.6d)$$

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 (26.6) are

$$\varphi_{mn} = \begin{pmatrix} \psi & w_x & w_y & \nabla_z w \\ \tilde{w}_x & \Phi + h_{xx} & h_{xy} + \nabla_z \tilde{h} & \nabla_z \tilde{h}_x \\ \tilde{w}_y & h_{xy} - \nabla_z \tilde{h} & \Phi - h_{xx} & \nabla_z \tilde{h}_y \\ \nabla_z \tilde{w} & \nabla_z h_x & \nabla_z h_y & \Phi + \nabla_z^2 h \end{pmatrix}. \quad (26.7)$$

To identify coordinate gauge-invariant quantities, it is necessary to consider infinitesimal coordinate gauge transformations (25.9). The 4 tetrad-frame components  $\epsilon_m$  of the coordinate shift of the coordinate gauge transformation decompose into 2 scalars and 1 vector

$$\epsilon_m = \left\{ \begin{array}{c} \epsilon_0 \\ \text{scalar} \end{array}, \quad \begin{array}{c} \nabla_a \epsilon \\ \text{scalar} \end{array} + \begin{array}{c} \epsilon_a \\ \text{vector} \end{array} \right\}. \quad (26.8)$$

In the flat space background space being considered, the coordinate gauge transformation (25.20) of the vierbein perturbation simplifies to

$$\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \nabla_m \epsilon_n. \quad (26.9)$$

In terms of the scalar, vector, and tensor potentials introduced in equations (26.6), the gauge transformations (26.9) are

$$\varphi_{00} \rightarrow \begin{array}{c} \psi + \dot{\epsilon}_0 \\ \text{scalar} \end{array}, \quad (26.10a)$$

$$\varphi_{0a} \rightarrow \begin{array}{c} \nabla_a (w + \dot{\epsilon}) + (w_a + \dot{\epsilon}_a) \\ \text{scalar} \end{array} + \begin{array}{c} \text{vector} \end{array}, \quad (26.10b)$$

$$\varphi_{a0} \rightarrow \begin{array}{c} \nabla_a (\tilde{w} + \epsilon_0) + \tilde{w}_a \\ \text{scalar} \end{array} + \begin{array}{c} \text{vector} \end{array}, \quad (26.10c)$$

$$\varphi_{ab} \rightarrow \begin{array}{c} \delta_{ab} \Phi \\ \text{scalar} \end{array} + \begin{array}{c} \nabla_a \nabla_b (h + \epsilon) \\ \text{scalar} \end{array} + \varepsilon_{abc} \nabla_c \tilde{h} + \begin{array}{c} \nabla_a (h_b + \epsilon_b) \\ \text{vector} \end{array} + \begin{array}{c} \nabla_b \tilde{h}_a \\ \text{vector} \end{array} + \begin{array}{c} h_{ab} \\ \text{tensor} \end{array}. \quad (26.10d)$$

Equations (26.10a) imply that under an infinitesimal coordinate gauge transformation the potentials trans-

form as

$$\psi \rightarrow \psi + \dot{\epsilon}_0 , \quad (26.11a)$$

$$w \rightarrow w + \dot{\epsilon} , \quad w_a \rightarrow w_a + \dot{\epsilon}_a , \quad (26.11b)$$

$$\tilde{w} \rightarrow \tilde{w} + \epsilon_0 , \quad \tilde{w}_a \rightarrow \tilde{w}_a , \quad (26.11c)$$

$$\Phi \rightarrow \Phi , \quad h \rightarrow h + \epsilon , \quad \tilde{h} \rightarrow \tilde{h} , \quad h_a \rightarrow h_a + \epsilon_a , \quad \tilde{h}_a \rightarrow \tilde{h}_a , \quad h_{ab} \rightarrow h_{ab} . \quad (26.11d)$$

Eliminating the coordinate shift  $\epsilon_m$  from the transformations (26.11) yields 12 coordinate gauge-invariant combinations of the potentials

$$\begin{array}{ccccccccc} \psi - \dot{\tilde{w}} & , & w - \dot{h} & , & w_a - \dot{h}_a & , & \tilde{w}_a & , & \Phi \\ \text{scalar} & & \text{scalar} & & \text{vector} & & \text{vector} & & \text{scalar} \\ & & & & & & & & \text{scalar} \\ & & & & & & & & \tilde{h} \\ & & & & & & & & \text{vector} \\ & & & & & & & & \tilde{h}_a \\ & & & & & & & & \text{tensor} \\ & & & & & & & & h_{ab} \end{array} . \quad (26.12)$$

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, §25.6. There are 6 combinations of the coordinate gauge-invariant perturbations (26.12) 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

$$\boxed{\Psi \underset{\text{scalar}}{\equiv} \psi - \dot{w} - \dot{\tilde{w}} + \ddot{h}} , \quad (26.13a)$$

$$\boxed{\Phi \underset{\text{scalar}}{\equiv} } , \quad (26.13b)$$

$$\boxed{W_a \underset{\text{vector}}{\equiv} w_a + \tilde{w}_a - \dot{h}_a - \dot{\tilde{h}}_a} , \quad (26.13c)$$

$$\boxed{h_{ab} \underset{\text{tensor}}{\equiv} } . \quad (26.13d)$$

Since only the 6 tetrad and coordinate gauge-invariant potentials  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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 (§26.7), Newtonian gauge (§26.8), and synchronous gauge (§26.9). However, for the next several sections, no gauge will be chosen: the exposition will continue to be completely general.

## 26.2 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.

### 26.2.1 Metric

The unperturbed metric  $\overset{0}{g}_{\mu\nu}$  is the Minkowski metric, equation (26.1). The perturbation  $\overset{1}{g}_{\mu\nu}$  of the coordinate metric is, from equation (25.6),

$$\overset{1}{g}_{tt} = - \underset{\text{scalar}}{2\psi}, \quad (26.14a)$$

$$\overset{1}{g}_{ta} = - \underset{\text{scalar}}{\nabla_a(w + \tilde{w})} - \underset{\text{vector}}{(w_a + \tilde{w}_a)}, \quad (26.14b)$$

$$\overset{1}{g}_{ab} = - \underset{\text{scalar}}{\delta_{ab}} 2\Phi - \underset{\text{scalar}}{2\nabla_a\nabla_b h} - \underset{\text{vector}}{\nabla_a(h_b + \tilde{h}_b)} - \underset{\text{vector}}{\nabla_b(h_a + \tilde{h}_a)} - \underset{\text{tensor}}{2h_{ab}}. \quad (26.14c)$$

The coordinate metric is tetrad gauge-invariant, but not coordinate gauge-invariant.

### 26.2.2 Tetrad-frame connections

The tetrad-frame connections  $\Gamma_{kmn}$  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

$$\overset{1}{\Gamma}_{0a0} = - \underset{\text{scalar}}{\nabla_a(\psi - \dot{\tilde{w}})} + \underset{\text{vector}}{\dot{\tilde{w}}_a}, \quad (26.15a)$$

$$\overset{1}{\Gamma}_{0ab} = \underset{\text{scalar}}{\delta_{ab}\dot{\Phi}} - \underset{\text{scalar}}{\nabla_a\nabla_b(w - \dot{h})} - \frac{1}{2}\underset{\text{vector}}{(\nabla_aW_b + \nabla_bW_a)} + \underset{\text{vector}}{\nabla_b\tilde{w}_a} + \underset{\text{tensor}}{\dot{h}_{ab}}, \quad (26.15b)$$

$$\overset{1}{\Gamma}_{ab0} = \frac{1}{2}\underset{\text{vector}}{(\nabla_aW_b - \nabla_bW_a)} - \frac{\partial}{\partial t}\underset{\text{scalar}}{(\varepsilon_{abd}\nabla_d\tilde{h} - \nabla_a\tilde{h}_b + \nabla_b\tilde{h}_a)}, \quad (26.15c)$$

$$\overset{1}{\Gamma}_{abc} = \underset{\text{scalar}}{(\delta_{bc}\nabla_a - \delta_{ac}\nabla_b)\Phi} - \underset{\text{scalar}}{\nabla_k(\varepsilon_{abd}\nabla_d\tilde{h} - \nabla_a\tilde{h}_b + \nabla_b\tilde{h}_a)} + \underset{\text{vector}}{\nabla_a h_{bc}} - \underset{\text{tensor}}{\nabla_b h_{ac}}. \quad (26.15d)$$

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 (26.12). 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 §25.7, 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  $\epsilon$ . 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  $\varphi_{mn}$ .

### 26.2.3 Tetrad-frame Einstein tensor

The tetrad-frame Einstein tensor  $G_{mn}$  in perturbed Minkowski space follows from the usual formulae (11.61), (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

$$\boxed{\begin{aligned} \overset{1}{G}_{00} &= 2 \underset{\text{scalar}}{\nabla^2 \Phi}, \end{aligned}} \quad (26.16a)$$

$$\boxed{\begin{aligned} \overset{1}{G}_{0a} &= 2 \underset{\text{scalar}}{\nabla_a \dot{\Phi}} + \frac{1}{2} \underset{\text{vector}}{\nabla^2 W_a}, \end{aligned}} \quad (26.16b)$$

$$\boxed{\begin{aligned} \overset{1}{G}_{ab} &= 2 \underset{\text{scalar}}{\delta_{ab} \ddot{\Phi}} - (\underset{\text{scalar}}{\nabla_a \nabla_b} - \underset{\text{scalar}}{\delta_{ab} \nabla^2})(\Psi - \Phi) + \frac{1}{2} (\underset{\text{vector}}{\nabla_a \dot{W}_b + \nabla_b \dot{W}_a}) + \underset{\text{tensor}}{\square h_{ab}}, \end{aligned}} \quad (26.16c)$$

where  $\square$  is the d'Alembertian, the 4-dimensional wave operator

$$\square \equiv \nabla_m \nabla^m = - \frac{\partial^2}{\partial t^2} + \nabla^2. \quad (26.17)$$

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 (26.16) depend only on the coordinate and tetrad gauge-invariant potentials  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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

$$\overset{1}{G}_{mn} = \begin{pmatrix} 2 \nabla_z^2 \Phi & \frac{1}{2} \nabla_z^2 W_x & \frac{1}{2} \nabla_z^2 W_y & 2 \nabla_z \dot{\Phi} \\ \frac{1}{2} \nabla_z^2 W_x & 2 \ddot{\Phi} + \nabla_z^2(\Psi - \Phi) + \square h_+ & \square h_\times & \frac{1}{2} \nabla_z \dot{W}_x \\ \frac{1}{2} \nabla_z^2 W_y & \square h_\times & 2 \ddot{\Phi} + \nabla_z^2(\Psi - \Phi) - \square h_+ & \frac{1}{2} \nabla_z \dot{W}_y \\ 2 \nabla_z \dot{\Phi} & \frac{1}{2} \nabla_z \dot{W}_x & \frac{1}{2} \nabla_z \dot{W}_y & 2 \ddot{\Phi} \end{pmatrix}, \quad (26.18)$$

where  $h_+$  and  $h_\times$  are the two linear polarizations of gravitational waves, discussed further in §26.13,

$$h_+ \equiv h_{xx} = -h_{yy}, \quad h_\times \equiv h_{xy} = h_{yx}. \quad (26.19)$$

The tetrad-frame complexified Weyl tensor is

$$\begin{aligned} \tilde{C}_{0a0b} &= \frac{1}{4} (\underset{\text{scalar}}{\nabla_a \nabla_b} - \frac{1}{3} \underset{\text{scalar}}{\delta_{ab} \nabla^2})(\Psi + \Phi) \\ &\quad + \frac{1}{8} [ -(\underset{\text{vector}}{\nabla_a \dot{W}_b + \nabla_b \dot{W}_a}) + i(\varepsilon_{acd} \nabla_b + \varepsilon_{bcd} \nabla_a) \nabla_c W_d ] \\ &\quad + \frac{1}{4} [ \underset{\text{tensor}}{\ddot{h}_{ab} - \varepsilon_{acd} \varepsilon_{bef} \nabla_c \nabla_e h_{df} - i(\varepsilon_{acd} \nabla_c \dot{h}_{bd} + \varepsilon_{bcd} \nabla_c \dot{h}_{ad})} ]. \end{aligned} \quad (26.20)$$

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  $\Psi$ ,  $\Phi$ ,  $W_a$ , and  $h_{ab}$ .

### 26.3 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, §12.1.1.

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 are (compare equations (26.16))

$$\overset{1}{G}_{00} = 2 \underset{\text{spin-0}}{\nabla_z^2} \Phi , \quad \overset{1}{G}_{0z} = 2 \underset{\text{spin-0}}{\nabla_z} \dot{\Phi} , \quad \overset{1}{G}_{zz} = \underset{\text{spin-0}}{2 \ddot{\Phi}} , \quad (26.21a)$$

$$\overset{1}{G}_{+-} - \overset{1}{G}_{zz} = \underset{\text{spin-0}}{\nabla_z^2} (\Psi - \Phi) , \quad (26.21b)$$

$$\overset{1}{G}_{0\pm} = \frac{1}{2} \underset{\text{spin-}\pm 1}{\nabla_z^2} W_{\pm} , \quad \overset{1}{G}_{z\pm} = \frac{1}{2} \underset{\text{spin-}\pm 1}{\nabla_z} \dot{W}_{\pm} , \quad (26.21c)$$

$$\overset{1}{G}_{\pm\pm} = \underset{\text{spin-}\pm 2}{\square} h_{\pm\pm} , \quad (26.21d)$$

where  $W_{\pm}$  are the spin- $\pm 1$  components of the vector perturbation  $W_a$

$$W_{\pm} = \frac{1}{\sqrt{2}} (W_x \pm i W_y) , \quad (26.22)$$

and  $h_{\pm\pm}$  are the spin- $\pm 2$  components of the tensor perturbation  $h_{ab}$

$$h_{\pm\pm} = h_{xx} \pm i h_{xy} = h_+ \pm i h_\times . \quad (26.23)$$

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  $\chi$  about the  $z$ -axis, while the linear polarizations  $h_+$  and  $h_\times$  transform as  $\cos 2\chi$  and  $-\sin 2\chi$ .

### 26.4 Too many Einstein equations?

The Einstein equations are as usual (units  $c = G = 1$ ; in the remainder of this chapter, perturbation overscripts <sup>1</sup> on the Einstein and energy-momentum tensors are dropped for brevity, which is fine because the unperturbed tensors vanish identically in the Minkowski background)

$$G_{mn} = 8\pi T_{mn} . \quad (26.24)$$

There are 10 Einstein equations, but the Einstein tensor (26.16) depends on only 6 independent potentials: the two scalars  $\Psi$  and  $\Phi$ , 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 (26.16a) and (26.16b) for  $G_{00}$  and  $G_{0a}$  enforces conservation of energy

$$D^m G_{m0} = 0 , \quad (26.25)$$

while the form of equations (26.16b) and (26.16c) for  $G_{0a}$  and  $G_{ab}$  enforces conservation of momentum

$$D^m G_{ma} = 0 . \quad (26.26)$$

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 (26.25) and (26.26) can be regarded as redundant. Since equations (26.25) and (26.26) 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.

## 26.5 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

$$\square h_{ab} = 0 . \quad (26.27)$$

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 (26.16). Specifically, there are 4 components of the Einstein tensor (26.16) that apparently depend only on spatial derivatives, not on time derivatives. The 4 corresponding Einstein equations are

$$\nabla^2 \Phi = 4\pi T_{00} , \quad \text{scalar} \quad (26.28a)$$

$$\nabla^2 W_a = 16\pi T_{0a} , \quad \text{vector} \quad (26.28b)$$

$$\nabla^2 (\Psi - \Phi) = -8\pi Q_{ab} T_{ab} , \quad \text{scalar} \quad (26.28c)$$

where  $Q_{ab}$  in equation (26.28c) is the quadrupole operator defined below, equation (26.101). These conditions must be satisfied everywhere at every instant of time, giving the impression that signals are travelling instantaneously from place to place.

## 26.6 Comparison to electromagnetism

The previous two sections §26.4 and §26.5 brought up two issues:

1. There are 10 Einstein equations, but only 6 independent gauge-invariant potentials  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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  $\Psi$ ,  $\Phi$ , and  $W_a$  satisfy equations (26.28) 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  $\phi$  and the vector potential  $\mathbf{A}$ :

$$A^m \equiv \{\phi, \mathbf{A}\} . \quad (26.29)$$

In flat (Minkowski) space, the electric and magnetic fields  $\mathbf{E}$  and  $\mathbf{B}$  are defined in terms of the potentials  $\phi$  and  $\mathbf{A}$  by

$$\mathbf{E} \equiv -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t} , \quad (26.30a)$$

$$\mathbf{B} \equiv \nabla \times \mathbf{A} . \quad (26.30b)$$

Given their definition (26.30), the electric and magnetic fields automatically satisfy the two source-free Maxwell's equations

$$\nabla \cdot \mathbf{B} = 0 , \quad (26.31a)$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 . \quad (26.31b)$$

The remaining two Maxwell's equations, the sourced ones, are

$$\nabla \cdot \mathbf{E} = 4\pi q , \quad (26.32a)$$

$$\nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} = 4\pi \mathbf{j} , \quad (26.32b)$$

where  $q$  and  $\mathbf{j}$  are the electric charge and current density, the time and space components of the electric 4-current density  $j^m$

$$j^m \equiv \{q, \mathbf{j}\} . \quad (26.33)$$

The electromagnetic potentials  $\phi$  and  $\mathbf{A}$  are not unique, but rather are defined only up to a gauge transformation by some arbitrary gauge field  $\theta$

$$\phi \rightarrow \phi - \frac{\partial \theta}{\partial t} , \quad \mathbf{A} \rightarrow \mathbf{A} + \nabla \theta . \quad (26.34)$$

The gauge transformation (26.34) evidently leaves the electric and magnetic fields  $\mathbf{E}$  and  $\mathbf{B}$ , equations (26.30), invariant.

Following the path of previous sections, §26.1 and thereafter, decompose the vector potential  $\mathbf{A}$  into its scalar and vector parts

$$\mathbf{A} = \underset{\text{scalar}}{\nabla A_{\parallel}} + \underset{\text{vector}}{\mathbf{A}_{\perp}}, \quad (26.35)$$

in which the vector part by definition satisfies the transversality condition  $\nabla \cdot \mathbf{A}_{\perp} = 0$ . Under a gauge transformation (26.34), the potentials transform as

$$\phi \rightarrow \phi - \frac{\partial \theta}{\partial t}, \quad (26.36a)$$

$$A_{\parallel} \rightarrow A_{\parallel} + \theta, \quad (26.36b)$$

$$\mathbf{A}_{\perp} \rightarrow \mathbf{A}_{\perp}. \quad (26.36c)$$

Eliminating the gauge field  $\theta$  yields 3 gauge-invariant potentials, comprising 1 scalar  $\Phi$ , and 1 vector  $\mathbf{A}_{\perp}$  containing 2 degrees of freedom:

$$\boxed{\Phi_{\text{scalar}} \equiv \phi + \frac{\partial A_{\parallel}}{\partial t}}, \quad (26.37a)$$

$$\boxed{\mathbf{A}_{\perp}^{\text{vector}}}. \quad (26.37b)$$

This shows that the electromagnetic field contains 3 independent degrees of freedom, consisting of 1 scalar and 1 vector.

**Concept question 26.1 Are gauge-invariant potentials Lorentz-invariant?** The potentials  $\Phi$  and  $\mathbf{A}_{\perp}$ , equations (26.37), are by construction gauge-invariant, but is this construction Lorentz-invariant? Do  $\Phi$  and  $\mathbf{A}_{\perp}$  constitute the components of a 4-vector?

In terms of the gauge-invariant potentials  $\Phi$  and  $\mathbf{A}_{\perp}$ , equations (26.37), the electric and magnetic fields are

$$\mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}_{\perp}}{\partial t}, \quad (26.38a)$$

$$\mathbf{B} = \nabla \times \mathbf{A}_{\perp}. \quad (26.38b)$$

The sourced Maxwell's equations (26.32) thus become, in terms of  $\Phi$  and  $\mathbf{A}_{\perp}$ ,

$$\boxed{-\nabla^2\Phi_{\text{scalar}} = 4\pi q_{\text{scalar}}}, \quad (26.39a)$$

$$\boxed{\nabla \dot{\Phi}_{\text{scalar}} - \square \mathbf{A}_{\perp}^{\text{vector}} = 4\pi \nabla j_{\parallel} + 4\pi \mathbf{j}_{\perp}^{\text{vector}}}, \quad (26.39b)$$

where  $\nabla j_{\parallel}$  and  $\mathbf{j}_{\perp}$  are the scalar and vector parts of the current density  $\mathbf{j}$ . Equations (26.39) bear a striking similarity to the Einstein equations (26.16). Only the vector part  $\mathbf{A}_{\perp}$  satisfies a wave equation,

$$\square \mathbf{A}_{\perp} = -4\pi \mathbf{j}_{\perp}, \quad (26.40)$$

while the scalar part  $\Phi$  satisfies an instantaneous equation (26.39a) that seemingly violates causality. And just as Einstein's equations (26.16) enforce conservation of energy-momentum, so also Maxwell's equations (26.39) enforce conservation of electric charge

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{j} = 0, \quad (26.41)$$

or in 4-dimensional form

$$\nabla_m j^m = 0. \quad (26.42)$$

The fact that only the vector part  $\mathbf{A}_{\perp}$  satisfies a wave equation (26.40) reflects physically the fact that electromagnetic waves are transverse, and they contain only two propagating degrees of freedom, the vector, or spin  $\pm 1$ , components.

Why do Maxwell's equations (26.39) have this structure? Although equation (26.40) 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. (25.23)). 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 (26.39b) 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 (26.39b), the Maxwell equation (26.39a) then serves precisely to enforce conservation of electric charge, equation (26.41).

Just as it is possible to regard the Einstein equations (26.16a) and (26.16b) as constraint equations whose continued satisfaction is guaranteed by conservation of energy-momentum, so also the Maxwell equation (26.39a) for  $\Phi$  can be regarded as a constraint equation whose continued satisfaction is guaranteed by conservation of electric charge. For charge conservation (26.41) coupled with the spatial Maxwell equation (26.39b) ensures that

$$\frac{\partial}{\partial t} (4\pi q + \nabla^2 \Phi) = 0, \quad (26.43)$$

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 (26.39a).

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 (26.39a) determines the scalar potential  $\Phi$  on the initial hypersurface of constant time, but can be discarded thereafter as redundant.

**Concept question 26.2 What parts of Maxwell's equations can be discarded?** Is it possible to discard the scalar part of the spatial Maxwell equation (26.39b), rather than the scalar equation (26.39a) for  $\Phi$ ? Project out the scalar part of equation (26.39b) by taking its divergence,

$$\nabla^2 (4\pi j_{||} - \dot{\Phi}) = 0 . \quad (26.44)$$

Argue that the Maxwell equation (26.39a), coupled with charge conservation (26.41), ensures that equation (26.44) is true, subject to boundary condition that the current  $\mathbf{j}$  vanish sufficiently rapidly at spatial infinity, in accordance with the decomposition theorem of §25.8.1.

Since only gauge-invariant quantities have physical significance, it is legitimate to impose any condition on the gauge field  $\theta$ . A gauge in which the potentials  $\phi$  and  $\mathbf{A}$  individually satisfy wave equations is **Lorenz** (not Lorentz!) **gauge**, which consists of the Lorentz-invariant condition

$$\nabla_m A^m = 0 . \quad (26.45)$$

Under a gauge transformation (26.34), the left hand side of equation (26.45) transforms as

$$\nabla_m A^m \rightarrow \nabla_m A^m + \square \theta , \quad (26.46)$$

and the Lorenz gauge condition (26.45) can be accomplished as a particular solution of the wave equation for the gauge field  $\theta$ . In terms of the potentials  $\phi$  and  $A_{||}$ , the Lorenz gauge condition (26.45) is

$$\frac{\partial \phi}{\partial t} + \nabla^2 A_{||} = 0 . \quad (26.47)$$

In Lorenz gauge, Maxwell's equations (26.39) become

$$\square \phi = -4\pi q , \quad (26.48a)$$

$$\square \mathbf{A} = -4\pi \mathbf{j} , \quad (26.48b)$$

which are manifestly wave equations for the potentials  $\phi$  and  $\mathbf{A}$ .

Does the fact that the potentials  $\phi$  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 (26.30) of  $\mathbf{E}$  and  $\mathbf{B}$  that if the potentials  $\phi$  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  $\phi$  and  $\mathbf{A}$  coincide with the gauge-invariant potentials  $\Phi$  and  $\mathbf{A}_{\perp}$ , equations (26.37). This gauge, known as **Coulomb gauge**, is accomplished by setting

$$A_{||} = 0 , \quad (26.49)$$

or equivalently

$$\nabla \cdot \mathbf{A} = 0 . \quad (26.50)$$

The gravitational analogue of this gauge is the Newtonian gauge discussed in the next section but one, §26.8.

Does the fact that in Lorenz gauge the potentials  $\phi$  and  $\mathbf{A}$  propagate at the speed of light (in the absence of sources,  $j^m = 0$ ) imply that the gauge-invariant potentials  $\Phi$  and  $\mathbf{A}_\perp$  propagate at the speed of light? No. The gauge-invariant potentials  $\Phi$  and  $\mathbf{A}_\perp$ , equations (26.37), are related to the Lorenz gauge potentials  $\phi$  and  $\mathbf{A}$  by a non-local decomposition.

## 26.7 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 (26.51), which can be considered an analogue of the Lorenz gauge of electromagnetism, equation (26.45). 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 (26.56), 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  $\Psi$ ,  $\Phi$ ,  $W_a$ , and  $h_{ab}$  defined by equations (26.13), only the tensor potential  $h_{ab}$  propagates causally. The situation is entirely analogous to that of electromagnetism, §26.6, where in Lorenz gauge the potentials  $\phi$  and  $\mathbf{A}$  propagate causally, equations (26.48), yet of the gauge-invariant potentials  $\Phi$  and  $\mathbf{A}_\perp$  defined by equations (26.37), only the vector potential  $\mathbf{A}_\perp$  propagates causally.

Harmonic gauge is the set of 4 coordinate conditions

$$\nabla^m(\varphi_{mn} + \varphi_{nm}) - \nabla_n \varphi_m{}^m = 0. \quad (26.51)$$

The conditions (26.51) are arranged in a form that is tetrad gauge-invariant (the conditions depend only on the symmetric part of  $\varphi_{mn}$ ). The quantities on the left hand side of equations (26.51) transform under a coordinate gauge transformation, in accordance with (26.9), as

$$\nabla^m(\varphi_{mn} + \varphi_{nm}) - \nabla_n \varphi_m{}^m \rightarrow \nabla^m(\varphi_{mn} + \varphi_{nm}) - \nabla_n \varphi_m{}^m - \square \epsilon_n. \quad (26.52)$$

The change  $\square \epsilon_n$  resulting from the coordinate gauge transformation is the 4-dimensional wave operator  $\square$  acting on the coordinate shift  $\epsilon_n$ . Indeed, the harmonic gauge conditions (26.51) 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 (26.46), and (b) that the conditions be tetrad gauge-invariant. The harmonic gauge conditions (26.51) can be accomplished as a particular solution of the wave equation for the coordinate shift  $\epsilon_n$ . In terms of the potentials defined by equations (26.6) and (26.13), the 4 harmonic gauge conditions (26.51) are

$$\dot{\Psi} + 3\dot{\Phi} - \square(w + \tilde{w} - \dot{h}) = 0, \quad (26.53a)$$

$$\dot{W}_a - \square(h_a + \tilde{h}_a) = 0, \quad (26.53b)$$

$$-\Psi + \Phi - \square h = 0, \quad (26.53c)$$

or equivalently

$$\square(w + \tilde{w}) = 4\dot{\Phi} , \quad (26.54a)$$

$$\square(h_a + \tilde{h}_a) = \dot{W}_a , \quad (26.54b)$$

$$\square h = -\Psi + \Phi . \quad (26.54c)$$

Substituting equations (26.54) into the Einstein tensor  $G_{mn}$  leads, after some calculation, to the result that in harmonic gauge,

$$\frac{1}{2}\square(\varphi_{mn} + \varphi_{nm} - \eta_{mn}\varphi) = G_{mn} , \quad (26.55)$$

or equivalently

$$\boxed{\frac{1}{2}\square(\varphi_{mn} + \varphi_{nm}) = R_{mn}} , \quad (26.56)$$

where  $R_{mn}$  is the Ricci tensor. Equation (26.56) 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 (26.56) 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  $\varphi_{mn}$  and their derivatives (and are therefore local), must also propagate at the speed of light.

The 4 coordinate gauge conditions (26.51) 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  $\Gamma_{kmn}$ , equations (26.15), is the 6 tetrad gauge conditions

$$\tilde{w} = \tilde{w}_a = \tilde{h} = \tilde{h}_a = 0 . \quad (26.57)$$

**Exercise 26.3 Einstein tensor in harmonic gauge.** Confirm equation (26.56).

## 26.8 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 (26.13). 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 (§26.9), 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

$$w = \tilde{w} = \tilde{w}_a = h = \tilde{h} = h_a = \tilde{h}_a = 0 , \quad (26.58)$$

so that the retained perturbations are the 6 coordinate and tetrad gauge-invariant perturbations (26.13)

$$\begin{matrix} \Psi \\ \text{scalar} \end{matrix} = \psi , \quad (26.59a)$$

$$\begin{matrix} \Phi \\ \text{scalar} \end{matrix} , \quad (26.59b)$$

$$\begin{matrix} W_a \\ \text{vector} \end{matrix} = w_a , \quad (26.59c)$$

$$\begin{matrix} h_{ab} \\ \text{tensor} \end{matrix} . \quad (26.59d)$$

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 (26.7),

$$\varphi_{mn} = \begin{pmatrix} \Psi & W_x & W_y & 0 \\ 0 & \Phi + h_{xx} & h_{xy} & 0 \\ 0 & h_{xy} & \Phi - h_{xx} & 0 \\ 0 & 0 & 0 & \Phi \end{pmatrix} . \quad (26.60)$$

The Newtonian line-element is, in a form that keeps the Newtonian tetrad manifest,

$$ds^2 = -[(1 + \Psi) dt]^2 + \delta_{ab} [(1 - \Phi) dx^a - h_c^a dx^c - W^a dt] [(1 - \Phi) dx^b - h_d^b dx^d - W^b dt] , \quad (26.61)$$

which reduces to the Newtonian metric

$$ds^2 = -(1 + 2\Psi) dt^2 - 2W_a dt dx^a + [\delta_{ab}(1 - 2\Phi) - 2h_{ab}] dx^a dx^b . \quad (26.62)$$

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  $\Psi$  and  $\Phi$  non-zero. Furthermore, as discussed in §26.12, 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  $u^m \equiv dx^m/d\tau = \{1, 0, 0, 0\}$ , and the corresponding coordinate 4-velocity  $u^\mu$  is, in Newtonian gauge,

$$u^\mu = e_0^\mu = \{1 - \Psi, W_a\} . \quad (26.63)$$

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 §26.11. The proper acceleration experienced by a person at rest in the tetrad frame, with tetrad 4-velocity  $u^m = \{1, 0, 0, 0\}$ , is

$$\frac{Du^a}{D\tau} = u^0 D_0 u^a = u^0 (\partial_0 u^a + \Gamma_{00}^a u^0) = \Gamma_{00}^a = \nabla_a \Psi . \quad (26.64)$$

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  $\Psi$ .

---

**Concept question 26.4 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?

---

## 26.9 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 (26.71) and (26.72), 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  $\varphi_{mn}$  with  $m = 0$  or  $n = 0$  of the vierbein perturbations to zero

$$\psi = w = \tilde{w} = w_a = \tilde{w}_a = 0 , \quad (26.65)$$

and makes the additional tetrad gauge choices

$$\tilde{h} = \tilde{h}_a = 0 , \quad (26.66)$$

with the result that the retained perturbations are the spatial perturbations

$$\begin{array}{lll} \Phi & , & h \\ \text{scalar} & , & \text{scalar} \\ & & \end{array} , \quad (26.67)$$

In terms of these spatial perturbations, the gauge-invariant perturbations (26.13) are

$$\begin{array}{ll} \Psi & = \ddot{h} \\ \text{scalar} & \end{array} , \quad (26.68a)$$

$$\begin{array}{ll} \Phi & , \\ \text{scalar} & \end{array} , \quad (26.68b)$$

$$\begin{array}{ll} W_a & = -\dot{h}_a \\ \text{vector} & \end{array} , \quad (26.68c)$$

$$\begin{array}{ll} h_{ab} & . \\ \text{tensor} & \end{array} \quad (26.68d)$$

The synchronous line-element is, in a form that keeps the synchronous tetrad manifest,

$$ds^2 = -dt^2 + \delta_{ab} [(1 - \Phi)dx^a - (\nabla_c \nabla^a h + \nabla_c h^a + h_c^a)dx^c] [(1 - \Phi)dx^b - (\nabla_d \nabla^b h + \nabla_d h^b + h_d^b)dx^d] , \quad (26.69)$$

which reduces to the synchronous metric

$$ds^2 = -dt^2 + [(1 - 2\Phi)\delta_{ab} - 2\nabla_a \nabla_b h - \nabla_a h_b - \nabla_b h_a - 2h_{ab}] dx^a dx^b . \quad (26.70)$$

In synchronous gauge, a person at rest in the tetrad frame has coordinate 4-velocity

$$u^\mu = e_0^\mu = \{1, 0, 0, 0\} , \quad (26.71)$$

so that the tetrad rest frame coincides with the coordinate rest frame. 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,

$$\frac{Du^a}{D\tau} = u^0 (\partial_0 u^a + \Gamma_{00}^a u^0) = \Gamma_{00}^a = 0 , \quad (26.72)$$

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 (26.15a) with the synchronous gauge choices (26.65) and (26.66).

## 26.10 Newtonian potential

The next few sections examine the physical meaning of each of the gauge-invariant potentials  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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 (26.16a), imply Poisson's equation

$$\nabla^2 \Phi = 4\pi\rho , \quad (26.73)$$

where  $\rho$  is the mass-energy density

$$\rho \equiv T_{00} . \quad (26.74)$$

The solution of Poisson's equation (26.73) is

$$\Phi(\mathbf{x}) = - \int \frac{\rho(\mathbf{x}') d^3x'}{|\mathbf{x}' - \mathbf{x}|} . \quad (26.75)$$

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  $\rho$ ,

$$M \equiv \int \rho(\mathbf{x}') d^3x' . \quad (26.76)$$

Equation (26.76) agrees with what the definition of the mass  $M$  would be in the non-relativistic limit, and as seen below, equation (26.79), it is what a distant observer would infer the mass of the body to be based on its gravitational potential  $\Phi$  far away. Thus equation (26.76) 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

$$\int \mathbf{x}' \rho(\mathbf{x}') d^3x' = 0 . \quad (26.77)$$

Consider the potential  $\Phi$  at a point  $\mathbf{x}$  far outside the body. Expand the denominator of the integral on the right hand side of equation (26.75) as a Taylor series in  $1/x$  where  $x \equiv |\mathbf{x}|$

$$\frac{1}{|\mathbf{x}' - \mathbf{x}|} = \frac{1}{x} \sum_{\ell=0}^{\infty} \left( \frac{x'}{x} \right)^{\ell} P_{\ell}(\hat{\mathbf{x}} \cdot \hat{\mathbf{x}'}) = \frac{1}{x} + \frac{\hat{\mathbf{x}} \cdot \mathbf{x}'}{x^2} + \dots \quad (26.78)$$

where  $P_{\ell}(\mu)$  are Legendre polynomials. Then

$$\begin{aligned} \Phi(\mathbf{x}) &= -\frac{1}{x} \int \rho(\mathbf{x}') d^3 x' - \frac{1}{x^2} \hat{\mathbf{x}} \cdot \int \mathbf{x}' \rho(\mathbf{x}') d^3 x' - O(x^{-3}) \\ &= -\frac{M}{x} - O(x^{-3}) . \end{aligned} \quad (26.79)$$

Equation (26.79) shows that the potential far from a body goes as  $\Phi = -M/x$ , reproducing the usual Newtonian formula.

## 26.11 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 (26.63). 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 (26.16b), imply

$$\nabla^2 \mathbf{W} = -16\pi \mathbf{f} , \quad (26.80)$$

where  $\mathbf{W} \equiv W_a$  is the gauge-invariant vector potential, and  $\mathbf{f}$  is the vector part of the energy flux  $T^{0a}$

$$\mathbf{f} \equiv f_a = f^a \equiv \underset{\text{vector}}{T^{0a}} = -\underset{\text{vector}}{T_{0a}} . \quad (26.81)$$

The solution of equation (26.80) is

$$\mathbf{W}(\mathbf{x}) = 4 \int \frac{\mathbf{f}(\mathbf{x}') d^3 x'}{|\mathbf{x}' - \mathbf{x}|} . \quad (26.82)$$

As in the previous section, §26.10, 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,

$$\int \mathbf{f}(\mathbf{x}') d^3 x' = 0 . \quad (26.83)$$

Define the angular momentum  $\mathbf{L}$  of the body to be

$$\mathbf{L} \equiv \int \mathbf{x}' \times \mathbf{f}(\mathbf{x}') d^3 x' . \quad (26.84)$$

Equation (26.84) agrees with what the definition of angular momentum would be in the non-relativistic limit, where the mass-energy flux of a mass density  $\rho$  moving at velocity  $\mathbf{v}$  is  $\mathbf{f} = \rho \mathbf{v}$ . As will be seen below, the angular momentum (26.84) 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 (26.84) 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 (26.85), the integral  $\int x'_a f_b(\mathbf{x}') d^3x'$  is antisymmetric in  $ab$ . To show this, write  $f_b = \varepsilon_{bcd} \nabla_c \phi_d$  for some potential  $\phi_d$ , which is valid because  $f_b$  is the vector (curl) part of the energy flux. Then

$$\int x'_a f_b(\mathbf{x}') d^3x' = \int x'_a \varepsilon_{bcd} \nabla'_c \phi_d(\mathbf{x}') d^3x' = - \int \varepsilon_{bcd} \phi_d(\mathbf{x}') \nabla'_c x'_a d^3x' = \int \varepsilon_{abd} \phi_d(\mathbf{x}') d^3x' , \quad (26.85)$$

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 (26.82) using equation (26.78) gives

$$\begin{aligned} \mathbf{W}(\mathbf{x}) &= \frac{4}{x} \int \mathbf{f}(\mathbf{x}') d^3x' + \frac{4}{x^2} \int (\hat{\mathbf{x}} \cdot \mathbf{x}') \mathbf{f}(\mathbf{x}') d^3x' + O(x^{-3}) \\ &= \frac{2}{x^2} \int [(\hat{\mathbf{x}} \cdot \mathbf{x}') \mathbf{f}(\mathbf{x}') - (\hat{\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}')) \mathbf{x}'] d^3x' + O(x^{-3}) \\ &= \frac{2}{x^2} \mathbf{L} \times \hat{\mathbf{x}} + O(x^{-3}) , \end{aligned} \quad (26.86)$$

where the first integral on the right hand side of the first line of equation (26.86) vanishes because the frame is the rest frame of the body, equation (26.83), 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 (26.85). 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 (26.86) says that a body of angular momentum  $\mathbf{L}$  drags frames around it at angular velocity  $\boldsymbol{\Omega}$  at large distances  $x$

$$\mathbf{W} = \boldsymbol{\Omega} \times \mathbf{x} , \quad \boldsymbol{\Omega} = \frac{2\mathbf{L}}{x^3} . \quad (26.87)$$

**Exercise 26.5 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  $\sigma^m$  of the gyro has only spatial components

$$\sigma^m = \{0, \sigma^a\} . \quad (26.88)$$

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  $\sigma^m$  in the gyro frame by a Lorentz boost at 4-velocity  $-u^m$  (early alphabet indices  $a, b, \dots$  signify spatial components):

$$\{s^0, s^a\} = \left\{ \sigma_b u^b, \sigma^a + \frac{\sigma_b u^b u^a}{1 + u^0} \right\} . \quad (26.89)$$

Conversely, the components  $\sigma^m$  of the spin vector in the gyro frame are related to those  $s^m$  in the tetrad frame by

$$\sigma^a = s^a - \frac{s^0 u^a}{1 + u^0} . \quad (26.90)$$

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

$$\frac{du^k}{d\tau} + \Gamma_{mn}^k u^m u^n = 0 , \quad \frac{ds^k}{d\tau} + \Gamma_{mn}^k s^m u^n = 0 . \quad (26.91)$$

**1. Spin equation.** Show that

$$\frac{d\sigma^a}{d\tau} = \sigma^b \left[ -\Gamma_{abc} u^c - \Gamma_{ab0} u^0 + \frac{u^c}{1 + u^0} (\Gamma_{0ac} u^b - \Gamma_{0bc} u^a) + \frac{u^0}{1 + u^0} (\Gamma_{0a0} u^b - \Gamma_{0b0} u^a) \right] . \quad (26.92)$$

[Hint: The first step is to convert  $\sigma^a$  to  $s^k$  and  $u^k$ , using equation (26.90). Then apply the geodesic equations (26.91). Then convert  $s^k$  back to  $\sigma^a$  using equation (26.89).]

**2. Spin precession.** Gravitational fields in the solar system are weak, so perturbation theory in Minkowski space is valid. The tetrad connections  $\Gamma_{kmn}$  in Newtonian gauge are, from equations (26.15),

$$\Gamma_{0a0} = -\nabla_a \Psi , \quad (26.93a)$$

$$\Gamma_{0ab} = \delta_{ab} \dot{\Phi} - \frac{1}{2} (\nabla_a W_b + \nabla_b W_a) , \quad (26.93b)$$

$$\Gamma_{ab0} = \frac{1}{2} (\nabla_a W_b - \nabla_b W_a) , \quad (26.93c)$$

$$\Gamma_{abc} = (\delta_{bc} \nabla_a - \delta_{ac} \nabla_b) \Phi + \nabla_a h_{bc} - \nabla_b h_{ac} . \quad (26.93d)$$

Show from equation (26.92) that the spin  $\boldsymbol{\sigma} \equiv \sigma^a$  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 (26.92) can be converted to the coordinate time derivative  $d/dt$  by dividing by  $u^0 = dt/d\tau$ )

$$\frac{d\boldsymbol{\sigma}}{dt} = \boldsymbol{\sigma} \times \left[ \mathbf{v} \times \nabla \Phi + \frac{u^0}{1 + u^0} \mathbf{v} \times \nabla \Psi - \frac{1}{2} \nabla \times \mathbf{W} + \frac{v^c}{2(1 + u^0)} (\nabla W_c + \nabla_c \mathbf{W}) - v^c \nabla \times \mathbf{h}_c \right] . \quad (26.94)$$

where the vector of vectors  $\mathbf{h}_c$  is shorthand for the tensor potential,  $\mathbf{h}_c \equiv h_{ac}$ . Conclude that at non-relativistic velocities,  $|\mathbf{u}| \ll u^0 \approx 1$ , and for  $\Psi = \Phi$  and  $h_{ab} = 0$ , equation (26.94) reduces to

$$\frac{d\boldsymbol{\sigma}}{dt} = \boldsymbol{\sigma} \times \left( \frac{3}{2} \mathbf{v} \times \nabla \Phi - \frac{1}{2} \nabla \times \mathbf{W} \right) . \quad (26.95)$$

By comparing your equation (26.95) to the equation of motion of a 3-vector rotating at angular velocity  $\boldsymbol{\omega}$ ,

$$\frac{d\boldsymbol{\sigma}}{dt} = \boldsymbol{\omega} \times \boldsymbol{\sigma} , \quad (26.96)$$

deduce the angular velocity  $\boldsymbol{\omega}$  with which the spin  $\mathbf{s}$  precesses. The term depending on  $\Phi$  is the geodetic,

or de Sitter (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 non-relativistic, 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  $\Phi$  and  $\mathbf{W}$  at spatial position  $\mathbf{x}$  of, equations (26.79) and (26.86),

$$\Phi(\mathbf{x}) = -\frac{M}{x}, \quad \mathbf{W}(\mathbf{x}) = \frac{2}{x^2} \mathbf{L} \times \hat{\mathbf{x}}. \quad (26.97)$$

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,

$$\boldsymbol{\omega}_{\text{dS}} = \frac{3(GM)^{3/2}}{2c^2 x^{5/2}} \mathbf{n}, \quad (26.98)$$

while the frame-dragging/Lense-Thirring precession is

$$\boldsymbol{\omega}_{\text{LT}} = \frac{G}{c^2 x^3} [-\mathbf{L} + 3\hat{\mathbf{x}}(\hat{\mathbf{x}} \cdot \mathbf{L})]. \quad (26.99)$$

[Hint: You will need to use the relation between velocity  $v$  and potential  $\Phi$  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 [http://einstein.stanford.edu/content/fact\\_sheet/GPB\\_FactSheet-0405.pdf](http://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 <http://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?
- 

## 26.12 Quadrupole pressure

Einstein's equations applied to the part of the Einstein tensor (26.16c) involving  $\Psi - \Phi$  imply

$$\nabla^2(\Psi - \Phi) = -8\pi Q_{ab}T_{ab}, \quad (26.100)$$

where  $Q_{ab}$  is the quadrupole operator (an integro-differential operator) defined by

$$Q_{ab} \equiv \frac{3}{2} \nabla_a \nabla_b \nabla^{-2} - \frac{1}{2} \delta_{ab}, \quad (26.101)$$

with  $\nabla^{-2}$  the inverse spatial Laplacian operator. In Fourier space, the quadrupole operator is

$$Q_{ab} = \frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} . \quad (26.102)$$

The quadrupole operator  $Q_{ab}$  yields zero when acting on  $\delta_{ab}$  (that is,  $Q_{ab}$  is traceless), and the Laplacian operator  $\nabla^2$  when acting on  $\nabla_a \nabla_b$

$$Q_{ab} \delta_{ab} = 0 , \quad Q_{ab} \nabla_a \nabla_b = \nabla^2 . \quad (26.103)$$

The solution of equation (26.100) is

$$\Psi - \Phi = - \int \left[ \frac{3}{2} \frac{(x_a - x'_a)(x_b - x'_b)}{|x - x'|^2} - \frac{1}{2} \delta_{ab} \right] \frac{T_{ab}(x') d^3 x'}{|x - x'|} . \quad (26.104)$$

Taylor expanding equation (26.104) using equation (26.78) yields  $\Psi - \Phi$  at large distance in the  $z$ -direction from a finite body,

$$\Psi - \Phi = - \frac{1}{x} \int [T_{zz} - \frac{1}{2}(T_{xx} + T_{yy})] d^3 x' + O(x^{-2}) . \quad (26.105)$$

Equation (26.100) 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. However, this is still not enough to ensure that  $\Psi$  differs from  $\Phi$ : as found in Exercise 26.6, if the energy-momentum of a body is spherically symmetric, then  $\Psi - \Phi$  vanishes outside (but not inside) the body.

One situation where the difference between  $\Psi$  and  $\Phi$  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 26.6 Equality of the two scalar potentials outside a spherical body.** Argue that the traceless part of the energy-momentum tensor of a spherically symmetric distribution must take the form

$$T_{ab}(r) = (\hat{r}_a \hat{r}_b - \frac{1}{3} \delta_{ab})(p(r) - p_\perp(r)) , \quad (26.106)$$

where  $p(r)$  and  $p_\perp(r)$  are the radial and transverse pressures at radius  $r$ . From equation (26.104), show that  $\Psi - \Phi$  at radial distance  $x$  from the centre of a spherically symmetric distribution is

$$\Psi(x) - \Phi(x) = - \int_x^\infty (r^2 - x^2)(p(r) - p_\perp(r)) \frac{4\pi dr}{r} . \quad (26.107)$$

Notice that the integral is over  $r > x$ , that is, only energy-momentum outside radius  $x$  produces non-vanishing  $\Psi - \Phi$ . In particular, if the spherical body has finite extent, then  $\Psi - \Phi$  vanishes outside the body.

### 26.13 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 (26.19). 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 by Figure 26.1. The  $h_+$  polarization varies as  $\cos 2\chi$  under a right-handed rotation by angle  $\chi$  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 (26.16c) imply that gravitational waves are sourced by the tensor component of the energy-momentum

$$\square h_{ab} = 8\pi T_{ab} \text{ .} \quad (26.108)$$

The solution of the wave equation (26.108) can be obtained from the Green's function of the d'Alembertian wave operator  $\square$  defined by equation (26.17). 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.

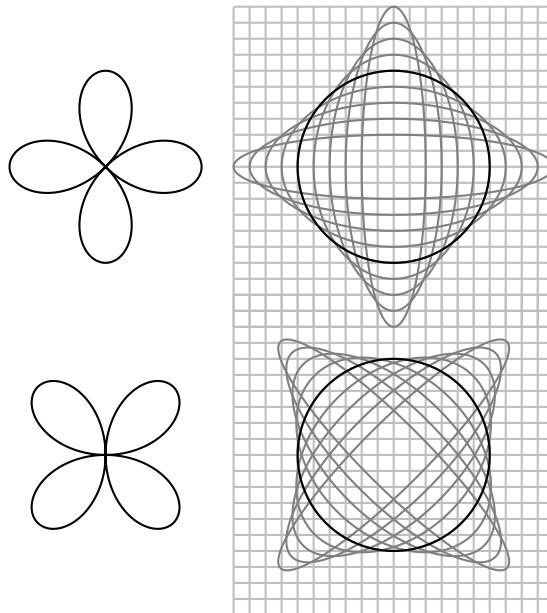


Figure 26.1 The two polarizations of gravitational waves. The (top) polarization  $h_+$  varies as  $\cos 2\chi$  under a right-handed rotation by angle  $\chi$  about the direction of propagation (into the paper), while the (bottom) polarization  $h_\times$  varies as  $-\sin 2\chi$ . A gravitational wave causes a system of freely falling test masses to oscillate relative to a grid of points a fixed proper distance apart.

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

$$\square F = \delta^4(x) , \quad (26.109)$$

where  $\delta^4(x) \equiv \delta(t)\delta^3(\mathbf{x})$  is the 4-dimensional Dirac delta-function. The solution of equation (26.109) subject to retarded boundary conditions is (a standard exercise in mathematics) the retarded Green's function

$$F = \frac{\delta(t - |\mathbf{x}|)\Theta(t)}{4\pi|\mathbf{x}|} , \quad (26.110)$$

where and  $\Theta(t)$  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 (26.108) is thus

$$h_{ab}(t, \mathbf{x}) = -2 \int \frac{T_{ab}(t', \mathbf{x}') d^3x'}{\frac{\text{tensor}}{|\mathbf{x}' - \mathbf{x}|}} , \quad (26.111)$$

where  $t'$  is the retarded time

$$t' \equiv t - |\mathbf{x}' - \mathbf{x}| , \quad (26.112)$$

which lies on the past light cone of the observer, and is the time at which the source emitted the signal. The solution (26.111) resembles the solution of Poisson's equation, except that the source is evaluated along the past light cone of the observer.

As in §§26.10 and 26.11, 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 26.8, is a fine example. Far from the body, the leading order contribution to the tensor potential  $h_{ab}$  is, from the multipole expansion (26.78),

$$h_{ab}(t, \mathbf{x}) = -\frac{2}{x} \int T_{ab}(t', \mathbf{x}') d^3x' . \quad (26.113)$$

The integral (26.113) 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

$$\frac{\partial^2 T^{00}}{\partial t^2} - \nabla_a \nabla_b T^{ba} = \frac{\partial}{\partial t} \left( \frac{\partial T^{00}}{\partial t} + \nabla_a T^{0a} \right) - \nabla_a \left( \frac{\partial T^{0a}}{\partial t} + \nabla_b T^{ba} \right) = 0 . \quad (26.114)$$

Multiply by  $x^a x^b$  and integrate

$$\int x^a x^b \frac{\partial^2 T^{00}}{\partial t^2} d^3x = \int x^a x^b \nabla_c \nabla_d T^{cd} d^3x = \int T^{cd} \nabla_c \nabla_d (x^a x^b) d^3x = 2 \int T^{ab} d^3x , \quad (26.115)$$

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 (26.115) is

$$\int x_a x_b \frac{\partial^2 T^{00}}{\partial t^2} d^3x \approx \frac{\partial^2}{\partial t^2} \int x_a x_b T^{00} d^3x = \frac{\partial^2 I_{ab}}{\partial t^2}, \quad (26.116)$$

where  $I_{ab}$  is the second moment of the mass

$$I_{ab} \equiv \int x_a x_b T^{00} d^3x. \quad (26.117)$$

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)

$$\mathcal{I}_{ab} \equiv I_{ab} - \frac{1}{3} \delta_{ab} I_c^c = \int (x_a x_b - \frac{1}{3} \delta_{ab} x^2) T^{00} d^3x. \quad (26.118)$$

Substituting the last expression of equations (26.115) into equation (26.113) gives the quadrupole formula for gravitational radiation at wavelengths long compared to the size of the emitting body

$$h_{ab}(t, \mathbf{x}) = -\frac{1}{x} \ddot{\mathcal{I}}_{ab}(t-x) \Big|_{\text{tensor}}.$$

(26.119)

Equation (26.119) 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

$$\mathcal{I}_+ = \frac{1}{2}(I_{xx} - I_{yy}), \quad \mathcal{I}_\times = \frac{1}{2}(I_{xy} + I_{yx}). \quad (26.120)$$

**Concept question 26.7 Units of the gravitational quadrupole radiation formula.** Restore units to the quadrupole formula (26.119) for gravitational radiation. **Answer:**

$$h_{ab}(t, \mathbf{x}) = -\frac{G}{c^4 x} \ddot{\mathcal{I}}_{ab}(t-x) \Big|_{\text{tensor}}. \quad (26.121)$$

## 26.14 Energy-momentum carried by gravitational waves

The gravitational wave equation (26.27) 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

$$\overset{0}{G}_{mn} = \overset{1}{G}_{mn} = 0. \quad (26.122)$$

The parts of the second-order perturbation that depend on the tensor perturbation  $h_{ab}$  are, in a frame where the wavevector  $\mathbf{k}$  is along the  $z$ -axis,

$$\overset{2}{G}_{00} = -(\dot{h}_{ab})(\dot{h}^{ab}) + \frac{1}{4}\left(\frac{\partial^2}{\partial t^2} + \nabla_z^2\right)h^2, \quad (26.123a)$$

$$\overset{2}{G}_{0z} = -(\dot{h}_{ab})(\nabla_z h^{ab}) + \frac{1}{2}\frac{\partial}{\partial t}\nabla_z h^2, \quad (26.123b)$$

$$\overset{2}{G}_{zz} = -(\nabla_z h_{ab})(\nabla_z h^{ab}) + \frac{1}{4}\left(\frac{\partial^2}{\partial t^2} + \nabla_z^2\right)h^2, \quad (26.123c)$$

where

$$h^2 \equiv h_{ab}h^{ab} = 2(h_+^2 + h_\times^2) = 2h_{++}h_{--}. \quad (26.124)$$

Since the Einstein tensor vanishes to linear order, equations (26.122), the Lie derivative of the linear order Einstein tensor is zero, and consequently the quadratic order expressions (26.123) are coordinate gauge-invariant. They are also tetrad gauge-invariant since they depend only on the (coordinate and) tetrad gauge-invariant perturbation  $h_{ab}$ . The rightmost set of terms on the right hand side of each of equations (26.123) are total derivatives (with respect to either time  $t$  or space  $z$ ). 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 (26.123) do not average to zero; for example, the terms for  $G_{00}$  and  $G_{zz}$  are negative everywhere, being minus a sum of squares. 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  $T_{mn}^{\text{gw}}$  in gravitational waves

$$T_{00}^{\text{gw}} \equiv \frac{1}{8\pi} \left[ (\dot{h}_{ab})(\dot{h}^{ab}) - \frac{1}{4}\left(\frac{\partial^2}{\partial t^2} + \nabla_z^2\right)h^2 \right], \quad (26.125a)$$

$$T_{0z}^{\text{gw}} \equiv \frac{1}{8\pi} \left[ (\dot{h}_{ab})(\nabla_z h^{ab}) - \frac{1}{2}\frac{\partial}{\partial t}\nabla_z h^2 \right], \quad (26.125b)$$

$$T_{zz}^{\text{gw}} \equiv \frac{1}{8\pi} \left[ (\nabla_z h_{ab})(\nabla_z h^{ab}) - \frac{1}{4}\left(\frac{\partial^2}{\partial t^2} + \nabla_z^2\right)h^2 \right]. \quad (26.125c)$$

The terms involving total derivatives, although they vanish when averaged over a region larger than many wavelengths, ensure that the energy-momentum  $T_{mn}^{\text{gw}}$  in gravitational waves satisfies conservation of energy-momentum in the flat background space

$$\nabla^m T_{mn}^{\text{gw}} = 0. \quad (26.126)$$

Averaged over a region larger than many wavelengths, the energy-momentum in gravitational waves is

$$\boxed{\langle T_{mn}^{\text{gw}} \rangle = \frac{1}{8\pi}(\nabla_m h_{ab})(\nabla_n h^{ab})}. \quad (26.127)$$

Equation (26.127) may also be written explicitly as a sum over the two linear or circular polarizations

$$\begin{aligned}\langle T_{mn}^{\text{gw}} \rangle &= \frac{1}{4\pi} [(\nabla_m h_+)(\nabla_n h_+) + (\nabla_m h_\times)(\nabla_n h_\times)] \\ &= \frac{1}{8\pi} [(\nabla_m h_{++})(\nabla_n h_{--}) + (\nabla_n h_{++})(\nabla_m h_{--})] .\end{aligned}\quad (26.128)$$

### Exercise 26.8 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  $I_{ab}$  of the mass distribution defined by

$$I_{ab} \equiv \sum_{\text{masses } X} M_X (r_{X,a} r_{X,b} - \frac{1}{3} \delta_{ab} r_X^2) \quad (26.129)$$

is

$$I_{ab} = mr^2 (\hat{r}_a \hat{r}_b - \frac{1}{3} \delta_{ab}) , \quad (26.130)$$

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

$$m \equiv \frac{M_1 M_2}{M} , \quad M \equiv M_1 + M_2 . \quad (26.131)$$

[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  $\iota$  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

$$\mathbf{r} = r [(\hat{\mathbf{z}} \cos \iota - \hat{\mathbf{y}} \sin \iota) \cos \omega t + \hat{\mathbf{x}} \sin \omega t] \quad (26.132)$$

where  $\omega$  is the orbital frequency. Deduce that the tensor components of the quadrupole moment are

$$I_+ \equiv \frac{1}{2} (I_{xx} - I_{yy}) = \frac{1}{4} mr^2 [\cos^2 \iota - (1 + \sin^2 \iota) \cos 2\omega t] , \quad (26.133a)$$

$$I_\times \equiv I_{xy} = -\frac{1}{2} mr^2 \sin \iota \sin 2\omega t . \quad (26.133b)$$

[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

$$h_{ab} = -\frac{1}{z} \ddot{I}_{ab}(t-z) . \quad (26.134)$$

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  $T_{mn}^{\text{gw}}$  in gravitational waves is given by the quadrupole formula

$$4\pi T_{mn}^{\text{gw}} = (\nabla_m h_+)(\nabla_n h_+) + (\nabla_m h_\times)(\nabla_n h_\times) , \quad (26.135)$$

where  $\nabla_m = \{\partial/\partial t, \partial/\partial x^a\}$ . Show that the non-vanishing components of the gravitational wave energy-momentum tensor are

$$T_{00}^{\text{gw}} = -T_{0z}^{\text{gw}} = T_{zz}^{\text{gw}} = \frac{m^2 r^4 \omega^6}{2\pi z^2} (1 + 6 \sin^2 \iota + \sin^4 \iota - \cos^4 \iota \cos 4\omega(t-z)) . \quad (26.136)$$

[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  $T^{0z} = -T_{0z}$ ),

$$\dot{E} = - \int_{-\pi/2}^{\pi/2} T_{0z}^{\text{gw}} 2\pi z^2 \cos \iota dt . \quad (26.137)$$

Show that (with units of  $c$  and  $G$  restored)

$$\dot{E} = \frac{32Gm^2r^4\omega^6}{5c^5} . \quad (26.138)$$

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

$$E = -\frac{GM}{2r} , \quad (26.139)$$

and the radius  $r$  and angular frequency  $\omega$  are related by Kepler's third law

$$r^3 = \frac{GM}{\omega^2} . \quad (26.140)$$

The orbital period  $P$  is related to the angular frequency  $\omega$  by

$$P \equiv \frac{2\pi}{\omega} . \quad (26.141)$$

Conclude that

$$\frac{\dot{P}}{P} = -\frac{\dot{\omega}}{\omega} = \frac{3}{2} \frac{\dot{E}}{E} = -\frac{96(Gm)(GM)^{2/3}\omega^{8/3}}{5c^5} , \quad (26.142)$$

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)

$$M_1 = 1.4414 \text{ M}_\odot , \quad M_2 = 1.3867 \text{ M}_\odot . \quad (26.143)$$

The orbital period is

$$P = 0.322997448930 \text{ day} . \quad (26.144)$$

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 <http://link.springer.com/article/10.1007%2Fs10569-011-9352-4>.]

8. **Eccentricity correction.** Actually PSR B1913+16 has a substantial eccentricity,

$$e = 0.6171338 . \quad (26.145)$$

The correct general relativistic formula including the effects of eccentricity is equation (26.142) multiplied by a function  $f(e)$  of the eccentricity

$$\frac{\dot{P}}{P} = -\frac{96(Gm)(GM)^{2/3}\omega^{8/3}}{5c^5} f(e) , \quad (26.146)$$

with

$$f(e) = \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right) (1 - e^2)^{-7/2} . \quad (26.147)$$

Compare the eccentricity-corrected predicted numerical result for  $\dot{P}$  with the measured value

$$\dot{P} = -2.4184 \times 10^{-12} . \quad (26.148)$$

**Exercise 26.9 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

$$R_{0x0x} = -R_{0y0y} = -R_{0xzx} = R_{0yzy} = R_{zxzx} = -R_{zyzy} = \ddot{h}_+ , \quad (26.149a)$$

$$R_{0x0y} = -R_{0xzy} = -R_{0yzx} = R_{zxzy} = \ddot{h}_\times . \quad (26.149b)$$

From the expression (26.134) for  $h_{ab}$  that you derived in Exercise 26.8, and from the equation of geodesic deviation

$$\frac{D^2 \delta \xi_m}{D\tau^2} + R_{klmn} \delta \xi^k u^l u^n = 0 \quad (26.150)$$

deduce the tidal forces on a person moving non-relativistically. [Hint: If a person is moving non-relativistically, 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 meter

$$g = \frac{1}{M_g^2} \quad (26.151)$$

which you figured out in Exercise 11.9.]

---

---

## 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  $\Omega_x$  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  $\Omega_k$  affect the CMB power spectrum?
17. How does changing the dark energy  $\Omega_\Lambda$  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 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 centimeter 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  $P(k)$

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = \mathbf{1}_{\mathbf{k}\mathbf{k}'} P(k) , \quad (27.1)$$

where  $\mathbf{1}_{\mathbf{k}\mathbf{k}'}$  is the unit matrix in the Hilbert space of Fourier modes

$$\mathbf{1}_{\mathbf{k}\mathbf{k}'} \equiv (2\pi)^3 \delta_D^3(\mathbf{k} + \mathbf{k}') . \quad (27.2)$$

The “momentum-conserving” Dirac delta-function in equation (27.2) 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  $C_{\ell}$

$$\langle \Theta_{\ell m} \Theta_{\ell' m'} \rangle = \mathbf{1}_{\ell m, \ell' m'} C_{\ell} \quad (27.3)$$

where  $\mathbf{1}_{\ell m, \ell' m'}$  is the unit matrix in the Hilbert space of spherical harmonics (distinguish the three usages of  $\delta$  in this paragraph:  $\delta$  meaning fluctuation,  $\delta_D$  meaning Dirac delta-function, and  $\delta$  meaning Kronecker delta, as in the following equation)

$$\mathbf{1}_{\ell m, \ell' m'} \equiv \delta_{\ell\ell'} \delta_{m, -m'} . \quad (27.4)$$

Again, the “angular momentum-preserving” condition (27.4) that  $\ell = \ell'$  and  $m + m' = 0$  is a consequence of rotational symmetry. The same rotational symmetry implies that the power spectrum  $C_{\ell}$  is a function only of the harmonic number  $\ell$ , 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. 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. The relativistic suppression of the potential at small scales is responsible for the fact that the power spectrum of matter declines at small scales.
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  $\eta_s$  is the comoving sound horizon. 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  $\eta_s$  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. 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 non-relativistic ( $p = 0$ ) 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. 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 non-relativistic 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  $z_{\text{eq}} \approx 3400$ , 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 2.5'th 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. 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  $\rho$ , which was produced by both photons and baryons. The sound speed consequently decreased below  $\sqrt{\frac{1}{3}}$ . 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 dipole relative to the radiation monopole, 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. **Recombination.** As the temperature cooled below about 3,000 K, electrons combined with hydrogen and helium nuclei into neutral atoms. 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.
12. **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 (10.108). 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. 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. 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.

13. **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, 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.
14. **Matter fluctuations.** After recombination, perturbations in the non-baryonic and baryonic matter grew by gravity, essentially unaffected any longer by photon pressure. 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.
15. **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. If matter dominates the background, then the gravitational potential  $\Phi$  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.

# 28

---

## Cosmological perturbations in a flat Friedmann-Lemaître-Robertson-Walker 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  $\Omega$  was much closer to 1 at recombination than it is today. Since observations indicate that the Universe today is within 1% of being spatially flat (Ade, 2015), it is an excellent approximation to treat the Universe at the time of recombination as being spatially flat.

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 26.

The strategy is to start in a completely general gauge, and to discover how the conformal Newtonian (Copernican) gauge, which is used in subsequent chapters, emerges naturally as that gauge in which the perturbations are precisely the physical perturbations.

### 28.1 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  $\eta$  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

$$ds^2 = a(\eta)^2 (-d\eta^2 + dx^2 + dy^2 + dz^2) , \quad (28.1)$$

where  $a(\eta)$  is the cosmic scale factor. The unperturbed coordinate metric is thus the conformal Minkowski metric

$$\overset{0}{g}_{\mu\nu} = a(\eta)^2 \eta_{\mu\nu} . \quad (28.2)$$

The tetrad is taken to be orthonormal, with the unperturbed tetrad axes  $\gamma_m \equiv \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$  being aligned with the unperturbed coordinate axes  $\overset{0}{e}_\mu \equiv \{\overset{0}{e}_0, \overset{0}{e}_1, \overset{0}{e}_2, \overset{0}{e}_3\}$  so that the unperturbed vierbein and inverse vierbein are respectively  $1/a$  and  $a$  times the unit matrix

$$\overset{0}{e}_m{}^\mu = \frac{1}{a} \delta_m^\mu , \quad \overset{0}{e}^m{}_\mu = a \delta_\mu^m . \quad (28.3)$$

Let  $\nabla_a$  denote spatial derivatives with respect to comoving spatial coordinates,

$$\nabla_a \equiv \delta_a^{\alpha} \frac{\partial}{\partial x^{\alpha}} = \frac{\partial}{\partial x^a} , \quad (28.4)$$

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  $\nabla_a$  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  $\eta$ ,

$$\text{overdot} \equiv \frac{\partial}{\partial \eta} , \quad (28.5)$$

so that for example  $\dot{a} \equiv da/d\eta$ . The Hubble parameter  $H$  in the unperturbed background is

$$H \equiv \frac{\dot{a}}{a^2} . \quad (28.6)$$

## 28.2 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  $\nabla_a$  by  $-i$  times the comoving wavevector  $k_a$  (the choice of sign is the standard convention in cosmology)

$$\nabla_a \rightarrow -ik_a . \quad (28.7)$$

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 gradient  $\nabla_a$  will be used interchangeably with  $-ik_a$ , whichever is most convenient.

## 28.3 Classification of vierbein perturbations

The definition (25.1) of the vierbein perturbations  $\varphi_{mn}$  implies that the perturbed vierbein in the perturbed FLRW spacetime is

$$e_m^{\mu} = \frac{1}{a} (\delta_m^n + \varphi_m{}^n) \delta_n^{\mu} = (\eta_{mn} + \varphi_{mn}) {}^0 e_n^{\mu} , \quad (28.8)$$

while the perturbed inverse vierbein is

$$e^m_{\mu} = a (\delta_n^m - \varphi_n{}^m) \delta_n^{\mu} = (\eta^{mn} - \varphi^{nm}) {}^0 e_{n\mu} . \quad (28.9)$$

The covariant tetrad-frame components  $\varphi_{mn}$  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 (26.6) for the flat Minkowski background),

$$\varphi_{00} = \begin{matrix} \psi \\ \text{scalar} \end{matrix}, \quad (28.10a)$$

$$\varphi_{0a} = \begin{matrix} \nabla_a w \\ \text{scalar} \end{matrix} + \begin{matrix} w_a \\ \text{vector} \end{matrix}, \quad (28.10b)$$

$$\varphi_{a0} = \begin{matrix} \nabla_a \tilde{w} \\ \text{scalar} \end{matrix} + \begin{matrix} \tilde{w}_a \\ \text{vector} \end{matrix}, \quad (28.10c)$$

$$\varphi_{ab} = \begin{matrix} \delta_{ab} \phi \\ \text{scalar} \end{matrix} + \begin{matrix} \nabla_a \nabla_b h \\ \text{scalar} \end{matrix} + \varepsilon_{abc} \nabla_c \tilde{h} + \begin{matrix} \nabla_a h_b \\ \text{vector} \end{matrix} + \begin{matrix} \nabla_b \tilde{h}_a \\ \text{vector} \end{matrix} + \begin{matrix} h_{ab} \\ \text{tensor} \end{matrix}. \quad (28.10d)$$

The 4 covariant tetrad-frame components  $\epsilon_m$  of the coordinate shift of the coordinate gauge transformation (25.9) similarly decompose into 2 scalars and 1 vector (2 degrees of freedom) (the following equation is essentially the same as that (26.8) for the flat Minkowski background)

$$\epsilon_m = \left\{ \begin{matrix} \epsilon_0 \\ \text{scalar} \end{matrix}, \quad \begin{matrix} \nabla_a \epsilon \\ \text{scalar} \end{matrix} + \begin{matrix} \epsilon_a \\ \text{vector} \end{matrix} \right\}. \quad (28.11)$$

The vierbein perturbations  $\varphi_{mn}$  transform under a coordinate gauge transformation (25.9) as, equation (25.20),

$$\varphi_{mn} \rightarrow \varphi'_{mn} = \varphi_{mn} + \partial_m \epsilon_n = \varphi_{mn} + \frac{1}{a} \nabla_m \epsilon_n, \quad (28.12)$$

with vanishing contribution from the unperturbed tetrad-frame connection, equation (28.23), since the latter is symmetric whereas equation (25.20) depends on an antisymmetric combination of connections. The individual components of the vierbein perturbations transform under a coordinate gauge transformation as

$$\varphi_{00} \rightarrow \psi + \begin{matrix} \frac{1}{a} \frac{\partial \epsilon_0}{\partial \eta} \\ \text{scalar} \end{matrix}, \quad (28.13a)$$

$$\varphi_{0a} \rightarrow \nabla_a \left( w + \begin{matrix} \frac{1}{a} \left( \frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon \\ \text{scalar} \end{matrix} \right) + \left( w_a + \begin{matrix} \frac{1}{a} \left( \frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon_a \\ \text{vector} \end{matrix} \right), \quad (28.13b)$$

$$\varphi_{a0} \rightarrow \nabla_a \left( \tilde{w} + \begin{matrix} \frac{1}{a} \epsilon_0 \\ \text{scalar} \end{matrix} \right) + \begin{matrix} \tilde{w}_a \\ \text{vector} \end{matrix}, \quad (28.13c)$$

$$\varphi_{ab} \rightarrow \delta_{ab} \left( \phi - \frac{\dot{a}}{a^2} \epsilon_0 \right) + \nabla_a \nabla_b \left( h + \begin{matrix} \frac{1}{a} \epsilon \\ \text{scalar} \end{matrix} \right) + \varepsilon_{abc} \nabla_c \tilde{h} + \nabla_a \left( h_b + \begin{matrix} \frac{1}{a} \epsilon_b \\ \text{vector} \end{matrix} \right) + \nabla_b \tilde{h}_a + \begin{matrix} h_{ab} \\ \text{tensor} \end{matrix}, \quad (28.13d)$$

or equivalently

$$\psi \rightarrow \psi + \frac{1}{a} \frac{\partial \epsilon_0}{\partial \eta}, \quad (28.14a)$$

$$w \rightarrow w + \frac{1}{a} \left( \frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon, \quad w_a \rightarrow w_a + \frac{1}{a} \left( \frac{\partial}{\partial \eta} - \frac{\dot{a}}{a} \right) \epsilon_a, \quad (28.14b)$$

$$\tilde{w} \rightarrow \tilde{w} + \frac{1}{a} \epsilon_0, \quad \tilde{w}_a \rightarrow \tilde{w}_a, \quad (28.14c)$$

$$\phi \rightarrow \phi - \frac{\dot{a}}{a^2} \epsilon_0, \quad h \rightarrow h + \frac{1}{a} \epsilon, \quad \tilde{h} \rightarrow \tilde{h}, \quad h_a \rightarrow h_a + \frac{1}{a} \epsilon_a, \quad \tilde{h}_a \rightarrow \tilde{h}_a, \quad h_{ab} \rightarrow h_{ab}. \quad (28.14d)$$

Eliminating the coordinate shift  $\epsilon_m$  from the transformations (28.14) yields 12 coordinate gauge-invariant combinations of the perturbations

$$\begin{array}{ccccccccc} \psi - \left( \frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) \tilde{w}, & w - \dot{h}, & w_a - \dot{h}_a, & \tilde{w}_a, & \phi + \frac{\dot{a}}{a} \tilde{w}, & \tilde{h}, & \tilde{h}_a, & h_{ab} \\ \text{scalar} & \text{scalar} & \text{vector} & \text{vector} & \text{scalar} & \text{scalar} & \text{vector} & \text{tensor} \end{array}. \quad (28.15)$$

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

$$\boxed{\Psi_{\text{scalar}} \equiv \psi - \left( \frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) (w + \tilde{w} - \dot{h})}, \quad (28.16a)$$

$$\boxed{\Phi_{\text{scalar}} \equiv \phi + \frac{\dot{a}}{a} (w + \tilde{w} - \dot{h})}, \quad (28.16b)$$

$$\boxed{W_a_{\text{vector}} \equiv w_a + \tilde{w}_a - \dot{h}_a - \dot{\tilde{h}}_a}, \quad (28.16c)$$

$$\boxed{h_{ab}_{\text{tensor}}}. \quad (28.16d)$$

The coordinate and tetrad gauge-invariant perturbations (28.16) reduce to those (26.13) in Minkowski space when the cosmic scale factor does not change,  $\dot{a} = 0$ .

## 28.4 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 (28.13), 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  $\eta$ . The global gauge freedoms mean that the scalar and vector perturbations  $\Psi$ ,  $\Phi$ , 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.

To illustrate the global gauge freedoms, consider the line-element

$$ds^2 = a(\eta)^2 \left\{ -[1 + \Psi(\eta)]^2 d\eta^2 + [1 - \Phi(\eta)]^2 \delta_{ab} dx^a dx^b \right\}, \quad (28.17)$$

in which  $\Psi(\eta)$  and  $\Phi(\eta)$  are functions only of conformal time  $\eta$ . A rescaling of the cosmic scale factor  $a$ , together with a coordinate transformation of conformal time  $\eta$ ,

$$a \rightarrow a' = a(1 - \Phi), \quad (28.18a)$$

$$d\eta \rightarrow d\eta' = \left( \frac{1 + \Psi}{1 - \Phi} \right) d\eta, \quad (28.18b)$$

brings the line-element (28.19) to FLRW form,

$$ds^2 = a'(\eta')^2 (-d\eta'^2 + \delta_{ab} dx^a dx^b). \quad (28.19)$$

The rescaling (28.18a) 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 §28.3. The transformation (28.18b) of the time coordinate is allowed because  $\Psi$  and  $\Phi$  are functions only of time. The argument in §28.3 that  $\Psi$  and  $\Phi$  are gauge-invariant is spoiled because in the particular case that the time coordinate shift  $\epsilon_0$  is a function only of time  $\eta$ , the change in the perturbation  $\tilde{w}$  is decoupled from the change in  $\epsilon_0$ , because  $\tilde{w}$  and  $\epsilon_0$  appear only inside a spatial gradient in the transformation (28.13c). The freedom to adjust  $\tilde{w}$  by an amount depending only on time propagates into a freedom to adjust  $\Psi$  and  $\Phi$ , equations (28.16a) and (28.16b). More generally, the combination  $w + \tilde{w} - h$  upon which both  $\Psi$  and  $\Phi$  depend can be adjusted by adjusting any of  $w$ ,  $\tilde{w}$ , or  $h$  by an amount depending only on time, since all these perturbations appear inside spatial gradients in equations (28.13). Similarly, the vector perturbation  $W_a$ , equation (28.16c), can be adjusted by an amount depending only on time by adjusting either of  $h_a$  or  $\tilde{h}_a$ .

Physically, the residual global gauge freedom in the scalar perturbations  $\Psi$  and  $\Phi$  reflects the impossibility of distinguishing a perturbation of the mean from the mean. Any perturbation of the mean can be absorbed into an adjustment of the parameters of the unperturbed background.

To what does the residual global gauge freedom in the vector perturbation  $W_a$  correspond? Physically,  $W_a$  represents the velocity of dragging of the tetrad frame through the coordinates. A spatially uniform  $W_a$  corresponds to a uniform velocity of the entire Universe, which is observationally undetectable.

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  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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 28.1 Global curvature as a perturbation?** The usual FLRW metric contains a curvature constant  $\kappa$  in addition to a cosmic scale factor  $a$ . Can curvature  $\kappa$ , 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,  $\kappa$ , if small, can be treated as a perturbation. The isotropic (Poincaré) form of the FLRW line-element, equation (10.26), takes the form

$$ds^2 = a(\eta)^2 \left( -d\eta^2 + \frac{1}{1 + \frac{1}{4}\kappa x^2} \delta_{ab} dx^a dx^b \right) , \quad (28.20)$$

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,

$$\Phi(x) = \frac{1}{8}\kappa x^2 . \quad (28.21)$$

Is this a residual global gauge freedom? Equation (28.21) 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  $\kappa$  and  $\Phi$ . In pre-1998 days when astronomers were measuring  $\Omega_m \approx 0.3$  and only the reckless contemplated non-zero  $\Omega_\Lambda$ , it was necessary to consider that Nature might have chosen a substantial curvature  $\Omega_k \approx 0.7$ , in which case  $\kappa$  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  $\kappa = 0$ .

**Concept question 28.2 Can the Universe at large rotate?** Is it possible for a Universe to rotate globally? What would be the observable signature, if any?

---

## 28.5 Metric, tetrad connections, and Einstein tensor

This section gives expressions in a completely general gauge for perturbed quantities in the flat Friedmann-Lemaître-Robertson-Walker background geometry.

### 28.5.1 Metric

The unperturbed metric is the FLRW metric (28.2). The perturbation  $\overset{1}{g}_{\mu\nu}$  to the coordinate metric is, equation (25.6),

$$\overset{1}{g}_{\eta\eta} = -a^2 \underset{\text{scalar}}{2\psi}, \quad (28.22a)$$

$$\overset{1}{g}_{\eta a} = -a^2 \underset{\text{scalar}}{[\nabla_a(w + \tilde{w}) + (w_a + \tilde{w}_a)]}, \quad (28.22b)$$

$$\overset{1}{g}_{ab} = -a^2 \underset{\text{scalar}}{[2\phi\delta_{ab} + 2\underset{\text{scalar}}{\nabla_a\nabla_b h} + \underset{\text{vector}}{\nabla_a(h_b + \tilde{h}_b)} + \underset{\text{vector}}{\nabla_b(h_a + \tilde{h}_a)} + 2\underset{\text{tensor}}{h_{ab}}]}. \quad (28.22c)$$

The coordinate metric is tetrad gauge-invariant, but not coordinate gauge-invariant.

### 28.5.2 Tetrad-frame connections

The tetrad-frame connections  $\Gamma_{kmn}$  are obtained from the usual formula (11.54). The non-vanishing unperturbed tetrad-frame connections are

$$\overset{0}{\Gamma}_{0ab} = -\frac{\dot{a}}{a^2}\delta_{ab}. \quad (28.23)$$

The perturbations  $\overset{1}{\Gamma}_{kmn}$  to the tetrad-frame connections are

$$\overset{1}{\Gamma}_{0a0} = \frac{1}{a} \left[ -\underset{\text{scalar}}{\nabla_a} \left( \psi - \left( \frac{\partial}{\partial\eta} + \frac{\dot{a}}{a} \right) \tilde{w} \right) + \left( \frac{\partial}{\partial\eta} + \frac{\dot{a}}{a} \right) \tilde{w}_a \right], \quad (28.24a)$$

$$\overset{1}{\Gamma}_{0ab} = \frac{1}{a} \left[ \underset{\text{scalar}}{F\delta_{ab}} - \underset{\text{scalar}}{\nabla_a\nabla_b(w - \dot{h})} - \frac{1}{2}(\underset{\text{vector}}{\nabla_aW_b + \nabla_bW_a}) + \underset{\text{vector}}{\nabla_b\tilde{w}_a + \dot{h}_{ab}} \right], \quad (28.24b)$$

$$\overset{1}{\Gamma}_{ab0} = \frac{1}{a} \left[ \frac{1}{2}(\underset{\text{vector}}{\nabla_aW_b - \nabla_bW_a}) - \frac{\partial}{\partial\eta} (\underset{\text{scalar}}{\varepsilon_{abd}\nabla_d\tilde{h}} - \underset{\text{vector}}{\nabla_a\tilde{h}_b + \nabla_b\tilde{h}_a}) \right], \quad (28.24c)$$

$$\begin{aligned} \overset{1}{\Gamma}_{abc} = \frac{1}{a} & \left[ (\underset{\text{scalar}}{\delta_{bc}\nabla_a - \delta_{ac}\nabla_b})(\phi + \frac{\dot{a}}{a}\tilde{w}) - \frac{\dot{a}}{a}(\underset{\text{vector}}{\delta_{ac}\delta_{bd} - \delta_{bc}\delta_{ad}})\tilde{w}_d \right. \\ & \left. - \underset{\text{scalar}}{\nabla_c(\varepsilon_{abd}\nabla_d\tilde{h} - \nabla_a\tilde{h}_b + \nabla_b\tilde{h}_a)} + \underset{\text{vector}}{\nabla_a h_{bc} - \nabla_b h_{ac}} \right], \end{aligned} \quad (28.24d)$$

where  $F$  is defined by

$$F \equiv \frac{\dot{a}}{a}\psi + \dot{\phi}. \quad (28.25)$$

Equations (28.24) show that the perturbations  $\overset{1}{\Gamma}_{klm}$  of the tetrad-frame connections depend on all 12 of the coordinate gauge-invariant potentials (28.15). The only non-coordinate-gauge-invariant dependence of the

tetrad-frame connections is on  $F$  defined by equation (28.25). The quantity  $F$  transforms under a coordinate gauge transformation (25.9) as, from equations (28.14),

$$F \rightarrow F - \epsilon_0 \frac{d}{d\eta} \frac{\dot{a}}{a^2}. \quad (28.26)$$

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 (28.26) of  $F$  implies that  $\Gamma_{0ab}$  with  $a = b$  transforms under an infinitesimal coordinate transformation (25.9) as

$$\overset{1}{\Gamma}_{0ab} \rightarrow \overset{1}{\Gamma}_{0ab} - \frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{\dot{a}}{a^2} \delta_{ab}. \quad (28.27)$$

The transformation of the tetrad-frame connections under coordinate transformations can be checked another way. According to the rule established in §25.7, 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  $\epsilon$ . 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  $\Gamma_{0ab}$ . 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 (28.23), is a tetrad-frame tensor, because the spatial unit matrix  $\delta_{ab}$  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  $\Gamma_{0ab}$  transform as

$$\overset{1}{\Gamma}_{0ab} \rightarrow \overset{1}{\Gamma}_{0ab} - \mathcal{L}_\epsilon \Gamma_{0ab}, \quad \mathcal{L}_\epsilon \Gamma_{0ab} = \epsilon^k \partial_k \Gamma_{0ab} = \frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{\dot{a}}{a^2} \delta_{ab}, \quad (28.28)$$

in agreement with the transformation (28.27).

### 28.5.3 Tetrad-frame Einstein tensor

The tetrad-frame Einstein tensor  $G_{mn}$  follows from the usual formulae (11.61), (11.79), and (11.81). The unperturbed tetrad-frame Einstein tensor  $\overset{0}{G}_{mn}$  is (equations (28.29) differ from equations (10.29) because the time coordinate here is the conformal time  $\eta$ , not the cosmic time  $t$ )

$$\overset{0}{G}_{00} = 3 \frac{\dot{a}^2}{a^4}, \quad (28.29a)$$

$$\overset{0}{G}_{0a} = 0, \quad (28.29b)$$

$$\overset{0}{G}_{ab} = \left( -2 \frac{\ddot{a}}{a^3} + \frac{\dot{a}^2}{a^4} \right) \delta_{ab}. \quad (28.29c)$$

The perturbation  $\dot{G}_{mn}$  of the tetrad-frame Einstein tensor is

$$\dot{G}_{00} = \frac{1}{a^2} \left[ -6 \frac{\dot{a}}{a} F + 2 \nabla^2 \Phi \right], \quad (28.30a)$$

$$\dot{G}_{0a} = \frac{1}{a^2} \left[ 2 \nabla_a \left( F + \left( \frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \tilde{w} \right) + \frac{1}{2} \nabla^2 W_a + 2 \left( \frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \tilde{w}_a \right], \quad (28.30b)$$

$$\begin{aligned} \dot{G}_{ab} = \frac{1}{a^2} & \left[ \left( 2 \left( \frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) F + 2 \left( \frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \psi \right) \delta_{ab} - (\nabla_a \nabla_b - \delta_{ab} \nabla^2) (\Psi - \Phi) \right. \\ & \left. + \frac{1}{2} \left( \frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) (\nabla_a W_b + \nabla_b W_a) - \left( \frac{\partial^2}{\partial \eta^2} + 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} - \nabla^2 \right) h_{ab} \right]. \end{aligned} \quad (28.30c)$$

According to the rule established in §25.7, the variation of the Einstein tensor under a coordinate transformation equals minus its Lie derivative,

$$\dot{G}_{mn} \rightarrow \dot{G}_{mn} - \mathcal{L}_\epsilon G_{mn}. \quad (28.31)$$

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 (25.9) are

$$\mathcal{L}_\epsilon G_{00} = \epsilon^k \partial_k G_{00} = -\frac{\epsilon_0}{a} \frac{d}{d\eta} \frac{3\dot{a}^2}{a^4}, \quad (28.32a)$$

$$\mathcal{L}_\epsilon G_{ab} = \epsilon^k \partial_k G_{ab} = \frac{\epsilon_0}{a} \frac{d}{d\eta} \left( \frac{2\ddot{a}}{a^3} - \frac{\dot{a}^2}{a^4} \right) \delta_{ab}. \quad (28.32b)$$

It can be checked that the same transformations of the tetrad-frame Einstein components under a coordinate transformation follow from the expressions (28.30) for the perturbed Einstein components and the coordinate transformations (28.14) of the potentials.

The time-time and space-space perturbations  $\dot{G}_{00}$  and  $\dot{G}_{ab}$  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  $\dot{G}_{0a}$  are not tetrad gauge-invariant, as is evident from the fact that equation (28.30b) 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  $\tilde{h}_a$ ).

## 28.6 Gauge choices

Since only the 6 tetrad and coordinate gauge-invariant potentials  $\Psi$ ,  $\Phi$ ,  $W_a$ , 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, §28.8, which is constructed so as to retain only physical perturbations.

## 28.7 ADM gauge choices

The ADM (3+1) formalism, Chapter 17, chooses the tetrad time axis  $\gamma_0$  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

$$\tilde{w} = \tilde{w}_a = 0 . \quad (28.33)$$

The ADM choice simplifies the tetrad-frame connections (28.24) and the time-space component  $G_{0a}$  of the tetrad-frame frame Einstein tensor, equation (28.30b). The ADM lapse  $\alpha$  and shift  $\beta^\alpha$  are

$$\alpha = a(1 + \psi) , \quad \beta^\alpha = \nabla_\alpha w + w_\alpha . \quad (28.34)$$

Another gauge choice that significantly simplifies the tetrad connections (28.24), though does not affect the Einstein tensor (28.30), is

$$\tilde{h} = \tilde{h}_a = 0 . \quad (28.35)$$

If the wavevector  $\mathbf{k}$  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)  $\gamma_3$  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  $\gamma_1 \cdot e_y = \gamma_2 \cdot e_x$ .

## 28.8 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 (28.16). This gauge is called **conformal Newtonian gauge**, analogously to the Newtonian gauge of Minkowski space, §26.8. 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, §26.8. 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 (26.58),

$$w = \tilde{w} = \tilde{w}_a = h = \tilde{h} = h_a = \tilde{h}_a = 0 , \quad (28.36)$$

so that the retained perturbations are the 6 coordinate and tetrad gauge-invariant perturbations (28.16)

$$\begin{matrix} \Psi \\ \text{scalar} \end{matrix} = \psi , \quad (28.37\text{a})$$

$$\begin{matrix} \Phi \\ \text{scalar} \end{matrix} = \phi , \quad (28.37\text{b})$$

$$\begin{matrix} W_a \\ \text{vector} \end{matrix} = w_a , \quad (28.37\text{c})$$

$$\begin{matrix} h_{ab} \\ \text{tensor} \end{matrix} . \quad (28.37\text{d})$$

In conformal Newtonian gauge, the quantity  $F$  defined by equation (28.25) becomes the coordinate and tetrad gauge-invariant quantity

$$F \equiv \frac{\dot{a}}{a} \Psi + \dot{\Phi} . \quad (28.38)$$

The conformal Newtonian metric is

$$ds^2 = a^2 \left\{ - (1 + 2\Psi) d\eta^2 - 2W_a d\eta dx^a + [\delta_{ab}(1 - 2\Phi) - 2h_{ab}] dx^a dx^b \right\} . \quad (28.39)$$

### 28.8.1 Conformal Newtonian gauge: scalar perturbations

In conformal Newtonian gauge, the scalar perturbations of the Einstein equations are, from the expressions (28.30) for the Einstein tensor, the energy density, energy flux, monopole pressure, and quadrupole pressure equations

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 \overset{1}{T}{}^{00} , \quad (28.40\text{a})$$

$$ikF = 4\pi G a^2 \hat{k}_a T^{0a} , \quad (28.40\text{b})$$

$$\dot{F} + 2 \frac{\dot{a}}{a} F + \left( \frac{\ddot{a}}{a} - 2 \frac{\dot{a}^2}{a^2} \right) \Psi - \frac{k^2}{3} (\Psi - \Phi) = \frac{4}{3} G \pi a^2 \delta_{ab} \overset{1}{T}{}^{ab} , \quad (28.40\text{c})$$

$$k^2 (\Psi - \Phi) = 8\pi G a^2 \left( \frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) T^{ab} . \quad (28.40\text{d})$$

The perturbation overscript 1 has been omitted from the right hand sides of equations (28.40b) and (28.40d) since the unperturbed energy-momentum vanishes for these components. All 4 of the scalar Einstein equations (28.40) 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 (28.40a)–(28.40d) are redundant, since they serve simply to enforce conservation of energy and scalar momentum. Thus it suffices to take, as the equations

governing the potentials  $\Psi$  and  $\Phi$ , any 2 of the equations (28.40a)–(28.40d). One is free to retain whichever 2 of the equations is convenient. Usually the 1st equation, the energy equation (28.40a), and the 4th equation, the quadrupole pressure equation (28.40d), are most convenient to retain. But sometimes the 2nd equation, the scalar momentum equation (28.40b), is more convenient in place of the energy equation (28.40a).

### 28.8.2 Conformal Newtonian gauge: vector perturbations

The vector (spin-1) Einstein equations in conformal Newtonian gauge are, from the expressions (28.30) for the Einstein tensor,

$$\nabla^2 W_a = -16\pi G a^2 \underset{\text{vector}}{T^{0a}}, \quad (28.41a)$$

$$\left( \frac{\partial}{\partial \eta} + 2 \frac{\dot{a}}{a} \right) (\nabla_a W_b + \nabla_b W_a) = 16\pi G a^2 \underset{\text{vector}}{T^{ab}}. \quad (28.41b)$$

If the overall matter energy-momentum is conserved, as it must be, then either equation (28.41a) or equation (28.41b) 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  $T^{ab}$  of energy-momentum, equation (28.41b) ensures that the vector perturbation redshifts as  $a^{-2}$ ,

$$W_a \propto a^{-2} \quad \text{if } \underset{\text{vector}}{T^{ab}} = 0. \quad (28.42)$$

The tendency of vector perturbations to redshift away has the consequence that vector perturbations are usually negligible in standard cosmological models.

**Concept question 28.3 Evolution of vector perturbations in FLRW spacetimes.** For vanishing vector source, equation (28.41a) requires  $W_a = 0$ , yet equation (28.41b) for vanishing vector source has the more general solution  $W_a \propto a^{-2}$ . How are these compatible? **Answer.** Equations (28.41a) are constraint equations (comprising two of the momentum constraints). The constraint equations must be arranged to be satisfied on the initial hypersurface, and then equations (28.41b), which enforce conservation of energy-momentum, ensure that the constraint equations are satisfied thereafter. If there is no initial vector source of energy-momentum, then the constraints must vanish on the initial hypersurface, requiring  $W_a = 0$ , whereafter equations (28.41b) ensure that  $W_a = 0$ .

### 28.8.3 Conformal Newtonian gauge: tensor perturbations

The tensor (spin-2) Einstein equations in conformal Newtonian gauge are, from the expressions (28.30) for the Einstein tensor,

$$\left( \frac{\partial^2}{\partial \eta^2} + 2 \frac{\dot{a}}{a} \frac{\partial}{\partial \eta} - \nabla^2 \right) h_{ab} = -8\pi G a^2 \underset{\text{tensor}}{T^{ab}}. \quad (28.43)$$

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,

$$h_{ab} = \text{constant} \quad \text{for } k\eta \ll 1 \quad \text{if } T_{\text{tensor}}^{ab} = 0 . \quad (28.44)$$

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, 2014), but the signal may have been from aligned galactic dust rather primordial (Ade, 2015). The cosmic gravitational wave background could potentially be observed directly in the future.

**Exercise 28.4 Evolution of tensor perturbations (gravitational waves) in FLRW spacetimes.**  
Show that equation (28.43) can be rewritten in Fourier space

$$\left( \frac{\partial^2}{\partial \eta^2} - \frac{\ddot{a}}{a} + k^2 \right) (a h_{ab}) = -8\pi G a^3 T_{\text{tensor}}^{ab} . \quad (28.45)$$

What is the solution of equation (28.45) 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.** From equation (10.81) it follows that, for background energy-momentum dominated by a single species with  $p/\rho = w = \text{constant}$ ,

$$\frac{\dot{a}}{a} = \frac{2}{(1+3w)\eta}, \quad \frac{\ddot{a}}{a} = \frac{2(1-3w)}{(1+3w)^2\eta^2} . \quad (28.46)$$

The tensor evolution equation (28.45) in the absence of sources becomes

$$\left[ \frac{\partial^2}{\partial \eta^2} - \frac{2(1-3w)}{(1+3w)^2\eta^2} + k^2 \right] (a h_{ab}) = 0 . \quad (28.47)$$

The solution of equation (28.47) is a linear combination of Bessel functions  $J_{\pm n}$  (for  $w < -1/3$ , replace  $\eta$  with its magnitude  $|\eta|$ ),

$$h_{ab} = (k\eta)^{-n} [A_+ J_n(k\eta) + A_- J_{-n}(k\eta)] \quad (28.48)$$

of argument

$$n = \frac{3(1-w)}{2(1+3w)} . \quad (28.49)$$

Special cases are

$$n = \begin{cases} \frac{1}{2} & w = \frac{1}{3} , \\ \frac{3}{2} & w = 0 , \\ -\frac{3}{2} & w = -1 , \end{cases} \quad (28.50)$$

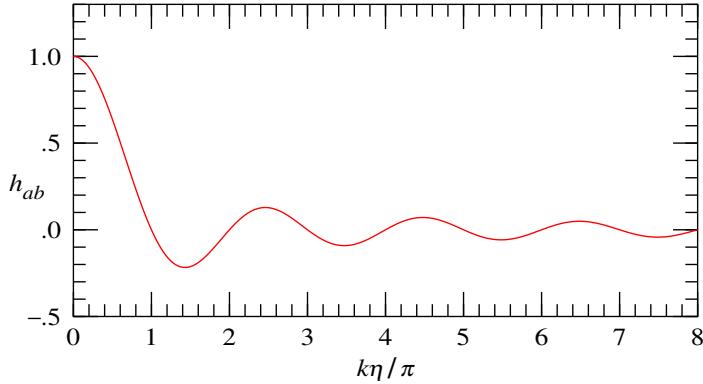


Figure 28.1 Evolution of the tensor potential  $h_{ab}$  in the radiation-dominated regime, where  $w = 1/3$ ,  $n = 1/2$ , and  $h_{ab} \propto \sin(k\eta)/(k\eta)$ .

in which case the solution reduces to spherical Bessel functions. The solution that is finite at  $\eta \rightarrow 0$  is, for  $n > 0$ , the  $A_+$  component. Normalized to 1 at  $\eta = 0$ , the finite solution is

$$h_{ab} = \Gamma(1+n) \left(\frac{k\eta}{2}\right)^{-n} J_n(k\eta) \rightarrow \begin{cases} 1 & k\eta \ll 1, \\ \frac{\Gamma(1+n)}{\sqrt{\pi}} \left(\frac{k\eta}{2}\right)^{-(n+1/2)} \cos[k\eta - (n + \frac{1}{2})\pi/2] & k\eta \gg 1. \end{cases} \quad (28.51)$$

Since

$$\eta^{n+1/2} \propto a, \quad (28.52)$$

the solution goes as  $h_{ab} \sim a^{-1} \cos(k\eta + \text{constant})$  at large  $k\eta$ . 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 28.1 illustrates the evolution of the tensor potential  $h_{ab}$  in the radiation-dominated regime.

#### 28.8.4 Conformal Newtonian gauge: energy-momentum conservation

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,

$$\overset{0}{T}{}^{00} \equiv \bar{\rho}, \quad (28.53a)$$

$$\overset{0}{T}{}^{0a} \equiv 0, \quad (28.53b)$$

$$\overset{0}{T}{}^{ab} \equiv \bar{p} \delta_{ab}. \quad (28.53c)$$

The perturbed components  $T^{mn}$  of the tetrad-frame energy-momentum are the energy density  $\rho$ , the energy flux  $f_a$ , and the pressure  $p_{ab}$ ,

$$T^{00} \equiv \rho = \bar{\rho} + \delta\rho , \quad (28.54a)$$

$$T^{0a} \equiv f_a , \quad (28.54b)$$

$$T^{ab} \equiv p_{ab} = \bar{p}\delta_{ab} + \delta p_{ab} . \quad (28.54c)$$

In perturbation theory, the perturbations  $\delta\rho$ ,  $f_a$ , and  $\delta p_{ab}$  are treated as of linear order. The trace of the spatial energy-momentum defines the isotropic pressure  $p$ ,

$$\frac{1}{3}T_a^a = p = \bar{p} + \delta p . \quad (28.55)$$

In conformal Newtonian gauge, the equations of conservation of energy and momentum are to linear order

$$D_m T^{m0} = \frac{1 - \Psi}{a} \left[ \frac{\partial \rho}{\partial \eta} + \nabla_a f_a + 3(\rho + p) \left( \frac{\dot{a}}{a} - \dot{\Phi} \right) \right] = 0 , \quad (28.56a)$$

$$D_m T^{ma} = \frac{1}{a} \left[ \frac{\partial f_a}{\partial \eta} + 4 \frac{\dot{a}}{a} f_a + \nabla_b p_{ab} + (\rho + p) \nabla_a \Psi \right] = 0 . \quad (28.56b)$$

Notice that the energy-momentum conservation equations (28.56) involve only the scalar potentials  $\Psi$  and  $\Phi$ , not the vector or tensor potentials  $W_a$  or  $h_{ab}$ . The energy equation (28.56a) has only a scalar component, while the momentum equation (28.56b) has both scalar and vector components, Exercise 28.5. The energy conservation equation (28.56a) has an unperturbed part,

$$D_m T^{\rho m0} = \frac{1}{a} \left[ \frac{\partial \bar{\rho}}{\partial \eta} + 3(\bar{\rho} + \bar{p}) \frac{\dot{a}}{a} \right] = 0 . \quad (28.57)$$

Any fluid component that conserves energy-momentum satisfies equations similar to (28.56). For a fluid component with equation of state  $p/\rho = w = \text{constant}$ , the unperturbed energy conservation equation (28.57) recovers the usual result that  $\bar{\rho} \propto a^{-3(1+w)}$ .

#### Concept question 28.5 Scalar, vector, tensor components of energy-momentum conservation.

What are the scalar, vector, and tensor components of the energy-momentum conservation equations (28.56)?

**Answer.** The energy conservation equation (28.56a) contains only scalar components. The momentum conservation equation (28.56b) 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 §31.6. The vector component of the pressure takes the form

$$p_{ab} = \underset{\text{vector}}{\nabla_a p_{\perp,b} + \nabla_b p_{\perp,a}} , \quad (28.58)$$

where  $p_{\perp,a}$  is transverse,  $\nabla_a p_{\perp,a} = 0$ . The vector part of the momentum conservation equation (28.56b) is

$$D_m T^{\rho ma} = \frac{1}{a} \left( \frac{\partial f_{\perp,a}}{\partial \eta} + 4 \frac{\dot{a}}{a} f_{\perp,a} + \nabla^2 p_{\perp,a} \right) = 0 . \quad (28.59)$$

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 (28.56) contain no tensor component.

---

## 28.9 Conformal synchronous gauge

One gauge that remains in common use in cosmology, but is not used here, is conformal synchronous gauge, discussed in the case of Minkowski background space in §26.9. The cosmological synchronous gauge choices are the same as for the Minkowski background, equations (26.65) and (26.66):

$$\psi = w = \tilde{w} = w_a = \tilde{w}_a = \tilde{h} = \tilde{h}_a = 0 . \quad (28.60)$$

The gauge-invariant perturbations (28.16) in synchronous gauge are

$$\begin{aligned} \Psi_{\text{scalar}} &= \left( \frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \right) \dot{h} , \end{aligned} \quad (28.61a)$$

$$\Phi_{\text{scalar}} = \phi - \frac{\dot{a}}{a} \dot{h} , \quad (28.61b)$$

$$W_a_{\text{vector}} = -\dot{h}_a , \quad (28.61c)$$

$$h_{ab} . \quad (28.61d)$$

Like synchronous gauge, §26.9, 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 28.6 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 32.1.

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 32.1. Thus at subhorizon scales conformal Newtonian gauge provides a good approximation to the frame of CMB photons.

**Concept question 28.7 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, §18.6? **Answer.** Common cosmological gauges, including conformal Newtonian or conformal synchronous, impose the ADM gauge condition that the time axis  $\gamma_0$  is orthogonal to hypersurfaces of constant time  $t$ , §28.7. 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, §18.6. 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.

---

## 29

---

### 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  $\rho$  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 (28.40d). 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 34.

## 29.1 Perturbed FLRW line-element

The perturbed FLRW line-element in conformal Newtonian gauge, equation (28.39), including only scalar perturbations, is

$$ds^2 = a^2 [-(1+2\Psi)d\eta^2 + \delta_{ab}(1-2\Phi)dx^adx^b] , \quad (29.1)$$

where  $a(\eta)$  is the cosmic scale factor, a function only of conformal time  $\eta$ .

## 29.2 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  $\rho$  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

$$T^{mn} = (\rho + p)u^mu^n + p\eta^{mn} . \quad (29.2)$$

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  $\rho$  (such a fluid is called **barotropic**),

$$p = p(\rho) . \quad (29.3)$$

Define  $w$  to be the derivative

$$w \equiv \frac{dp}{d\rho} , \quad (29.4)$$

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 31 considers the more realistic situation of a photon-baryonic fluid with non-constant  $w$ .

Each fluid moves with non-relativistic 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

$$u^m = \{1, v_a\} , \quad (29.5)$$

where  $v_a$  is its non-relativistic 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  $\rho$  can be written as a sum of an unperturbed density  $\bar{\rho}$  and a linear order fluctuation  $\delta\rho$ ,

$$\rho = \bar{\rho} + \delta\rho . \quad (29.6)$$

It proves advantageous, because it simplifies the resulting perturbation equations (29.13), to characterize the density fluctuation  $\delta\rho$  in terms of a fluctuation  $\delta$

$$\delta\rho = (\bar{\rho} + \bar{p})\delta , \quad (29.7)$$

where  $\bar{p} = p(\bar{\rho})$  in the unperturbed pressure. As you will discover in Exercise 29.1, the fluctuation  $\delta$  can be interpreted physically as the entropy fluctuation,

$$\delta \equiv \frac{\delta\rho}{\bar{\rho} + \bar{p}} = \frac{\delta s}{\bar{s}} . \quad (29.8)$$

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 (29.2) of the perfect fluid is then

$$T^{00} \equiv \rho = \bar{\rho} + \delta\rho , \quad (29.9a)$$

$$T^{0a} \equiv f_a = (\bar{\rho} + \bar{p})v_a , \quad (29.9b)$$

$$T^{ab} \equiv p\delta_{ab} = (\bar{p} + \delta p)\delta_{ab} , \quad (29.9c)$$

where the pressure fluctuation  $\delta p$  is, from equation (29.4),

$$\delta p = w\delta\rho . \quad (29.10)$$

If a species does not exchange energy or momentum with other species, then it satisfies the energy-momentum conservation equations (28.56) in conformal Newtonian gauge. Subtracting appropriate amounts of the unperturbed energy conservation equation (28.57) from the perturbed energy-momentum conservation equations (28.56) yields equations for the entropy fluctuation  $\delta$  and bulk velocity  $v_a$  of the fluid (recall that overdot denotes partial differentiation with respect to conformal time  $\eta$ , equation (28.5), so for example  $\dot{\delta} = \partial\delta/\partial\eta$ ),

$$\dot{\delta} + \nabla_a v_a = 3\dot{\Phi} , \quad (29.11a)$$

$$\dot{v}_a + (1 - 3w)\frac{\dot{a}}{a}v_a + w\nabla_a\delta = -\nabla_a\Psi . \quad (29.11b)$$

Physically, equation (29.11a) represents conservation of entropy, while equation (29.11b) 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

$$v_a = -i\hat{k}_a v + v_{\perp,a} . \quad (29.12)$$

The advantage of this choice is that  $v$  is dimensionless, as are  $\delta$  and  $\Psi$  and  $\Phi$ . Note that the comoving

wavenumber  $k$  (a constant for any given mode) has units of  $\eta^{-1}$ . The scalar parts of the perturbation equations (29.11) are then

$$\dot{\delta} - kv = 3\dot{\Phi}, \quad (29.13a)$$

$$\dot{v} + (1 - 3w)\frac{\dot{a}}{a}v + wk\delta = -k\Psi. \quad (29.13b)$$

Combining the two equations (29.13) for the scalar fluctuation  $\delta$  and scalar bulk velocity  $v$  yields a second-order differential equation for  $\delta - 3\Phi$ ,

$$\left[ \frac{d^2}{d\eta^2} + (1 - 3w)\frac{\dot{a}}{a}\frac{d}{d\eta} + wk^2 \right] (\delta - 3\Phi) = -k^2(\Psi + 3w\Phi). \quad (29.14)$$

Equation (29.14) holds for any perfect fluid that conserves energy-momentum and that has equation of state (29.4), with  $w$  not necessarily constant. For positive  $w$ , equation (29.14) 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 §29.5 and §29.6 below.

A more careful treatment, deferred to Chapter 32, accounts for the complete momentum distribution of radiation by expanding the temperature perturbation  $\Theta \equiv \delta T/\bar{T}$  in multipole moments, equation (32.46). The radiation fluctuation  $\delta_r$  and scalar bulk velocity  $v_r$  are related to the first two multipole moments of the temperature perturbation, the monopole  $\Theta_0$  and the dipole  $\Theta_1$ , by

$$\delta_r = 3\Theta_0, \quad (29.15a)$$

$$v_r = 3\Theta_1. \quad (29.15b)$$

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 29.1 Entropy perturbation.** The purpose of this exercise is to discover that the fluctuation  $\delta$  defined by equation (29.6) can be interpreted as the entropy fluctuation. According to the first law of thermodynamics, the entropy density  $s$  of a fluid of energy density  $\rho$ , pressure  $p$ , and temperature  $T$  in a volume  $V$  satisfies

$$d(\rho V) + pdV = Td(sV). \quad (29.16)$$

If the fluid is ideal, so that  $\rho$ ,  $p$ ,  $T$ , and  $s$  are independent of volume  $V$ , then integrating the first law (29.16) implies that

$$\rho V + pV = TsV. \quad (29.17)$$

This implies that the entropy density  $s$  is related to the other variables by

$$s = \frac{\rho + p}{T}. \quad (29.18)$$

Show that, for a perfect, barotropic fluid (one in which pressure is a prescribed function  $p(\rho)$  of density  $\rho$ ),

small variations of the density and entropy are related by

$$\frac{\delta\rho}{\rho+p} = \frac{\delta s}{s}, \quad (29.19)$$

confirming equation (29.8). [Hint: Do not confuse what is being asked here with adiabatic expansion. The result (29.19) 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 (29.18) to eliminate the temperature  $T$  from the first law (29.16), obtaining

$$\frac{d\rho}{\rho+p} = \frac{ds}{s}. \quad (29.20)$$

In the situation being considered, where pressure is a prescribed function  $p(\rho)$  of density, equation (29.20) implies equation (29.19).

**Concept question 29.2 Entropy perturbation when number is conserved.** The derivation of the entropy perturbation (29.19) in Exercise 29.1 was based on the first law of thermodynamics (29.16) without any term  $\mu dN$  representing number conservation. Should not such a term be included? **Answer.** This question was addressed in Exercise 10.14. Each chemical potential  $\mu$  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,

$$\mu_e dN_e + \mu_{\bar{e}} dN_{\bar{e}} = \mu_e d(N_e - N_{\bar{e}}) = 0. \quad (29.21)$$

Thus the entropy perturbation equation (29.19) does not hold individually for electrons and positrons, but it does hold for the combined electron-positron fluid.

---

### 29.3 Entropy conservation at superhorizon scales

At superhorizon scales, where  $k\eta \ll 1$ , the bulk velocity term in equation (29.13a) for the entropy fluctuation  $\delta$  is negligible, and the equation reduces to

$$\dot{\delta} = 3\dot{\Phi}. \quad (29.22)$$

It is conventional to define a quantity  $\zeta$  by

$$\zeta \equiv \frac{1}{3}\delta - \Phi, \quad (29.23)$$

which has the property that it is constant at large scales in any fluid component that does not exchange energy with other components,

$$\zeta = \text{constant} \quad \text{if } k\eta \ll 1. \quad (29.24)$$

Since both  $\delta$  and  $\Phi$  are gauge-invariant (all quantities in Newtonian gauge being gauge-invariant), so also is  $\zeta$ .

Physically, the constancy of  $\zeta$  at superhorizon scales is associated with a conservation law that has the appearance of a law of conservation of entropy. Recall that in a FLRW universe, the Einstein equations enforce a conservation law (10.56) that looks like the first law of thermodynamics with conserved entropy. The constancy of  $\zeta$  is a generalization of this law to superhorizon perturbations of a 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 (28.18) of the cosmic scale factor  $a$  (and time coordinate  $\eta$ ) so as to set the large-scale  $\Phi$  (and  $\Psi$ ) to zero in their patch. Then equation (29.22) becomes  $\dot{\delta} = 0$ , expressing the first law of thermodynamics (10.56) in the FLRW background of the patch.

The  $-3\Phi$  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 29.3 Relation between entropy and  $\zeta$ .** Assume that the proper pressure  $p(\rho)$  is a definite function of proper density  $\rho$ . Define entropy  $s$  per unit volume by (see Exercise 29.1)

$$\ln s \equiv \int \frac{d\rho}{\rho + p}. \quad (29.25)$$

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 (28.56) in Newtonian gauge reduces to

$$d\ln s + 3d\ln[a(1 - \Phi)] = 0, \quad (29.26)$$

whose unperturbed part is

$$d\ln \bar{s} + 3d\ln a = 0. \quad (29.27)$$

Conclude that energy conservation implies the conservation of  $\zeta$  defined by

$$\zeta = \frac{1}{3}\ln(s/\bar{s}) - \Phi. \quad (29.28)$$

**Concept question 29.4 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 (10.56). The constancy of  $\zeta$  is a generalization of this law to evolution at superhorizon scales, Exercise 29.3. 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  $\zeta$  is constant. How can these viewpoints be reconciled? **Answer.** The first law (10.56) 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 (10.56)

enforced by the Friedmann equations simply represents conservation of energy in a general relativistic context. The proof in Exercise 29.3 that  $\zeta$  represents a fluctuation in entropy depended on the proposition that the proper pressure  $p(\rho)$  is a definite function of proper density  $\rho$ . 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.

---

### 29.3.1 Primordial curvature fluctuation

It was remarked above, §29.3, that an observer inside a horizon patch can perform a global gauge transformation (28.18) so as to set the large-scale  $\Phi$  to zero in their patch. Alternatively, the observer has the gauge freedom to set the large scale fluctuation  $\delta$  in their patch to zero, in which case  $\zeta = -\Phi$ . For this reason, the total conserved fluctuation  $\zeta$  is commonly called the **primordial curvature fluctuation**.

The constancy of the primordial curvature fluctuation  $\zeta$  at superhorizon scales makes it useful for characterizing fluctuations during inflation. At the end of inflation, the “vacuum” energy-momentum of the inflaton field converts to the energy-momentum of matter and radiation. The details of this event, called reheating, are not well understood. However, since  $\zeta$  is constant, its value when a fluctuation first exits the horizon during inflation equals its value when the fluctuation reenters the horizon some time later. The constancy of  $\zeta$  makes the details of reheating largely inconsequential to the evolution of perturbations.

### 29.3.2 Adiabatic and isocurvature initial conditions

The conserved fluctuation in any particular species  $x$  that does not exchange energy with other species is denoted  $\zeta_x$  with a subscript  $x$ . The conserved fluctuation over all species is denoted  $\zeta$  with no subscript. A generic prediction of inflation is that the conserved fluctuation  $\zeta_x$  is the same for all species  $x$ ,

$$\zeta_x = \zeta . \quad (29.29)$$

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  $\delta_x$  of different species oppose each other so as to make zero contribution to the curvature potential  $\Phi$ . Among  $N$  species, there are 1 adiabatic and  $N - 1$  isocurvature modes subject to the condition that the initial fluctuations are finite.

## 29.4 Unperturbed background

The evolution of the cosmic scale factor  $a$  as a function of conformal time  $\eta$  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

$$\bar{\rho}_c \propto a^{-3}, \quad \bar{\rho}_r \propto a^{-4}. \quad (29.30)$$

The Hubble parameter  $H$  is defined in the usual way to be

$$H \equiv \frac{1}{a} \frac{da}{d\eta} = \frac{\dot{a}}{a^2}, \quad (29.31)$$

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

$$H^2 = \frac{8\pi G}{3} (\bar{\rho}_c + \bar{\rho}_r) = \frac{H_{\text{eq}}^2}{2} \left( \frac{a_{\text{eq}}^3}{a^3} + \frac{a_{\text{eq}}^4}{a^4} \right), \quad (29.32)$$

where  $a_{\text{eq}}$  and  $H_{\text{eq}}$  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  $\eta$  is defined to be the comoving distance that light travels starting from zero expansion. This is

$$\eta = \int_0^a \frac{da}{a^2 H} = \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left( \sqrt{1 + \frac{a}{a_{\text{eq}}}} - 1 \right) = \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left( \frac{a/a_{\text{eq}}}{1 + \sqrt{1 + a/a_{\text{eq}}}} \right). \quad (29.33)$$

The horizon distance  $\eta_{\text{eq}}$  at matter-radiation equality  $a = a_{\text{eq}}$  is

$$\eta_{\text{eq}} = \frac{2\sqrt{2}}{(1 + \sqrt{2}) a_{\text{eq}} H_{\text{eq}}}. \quad (29.34)$$

Equation (29.33) inverts to give the cosmic factor  $a$  as a function of the horizon distance  $\eta$ ,

$$\frac{a}{a_{\text{eq}}} = \frac{\eta}{8\eta_{\text{eq}}} \left( \frac{\eta}{\eta_{\text{eq}}} + 4\sqrt{2} \right). \quad (29.35)$$

In the radiation- and matter-dominated epochs respectively, the comoving horizon distance  $\eta$  is

$$\eta = \begin{cases} \frac{\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left( \frac{a}{a_{\text{eq}}} \right) = \frac{(1 + \sqrt{2})\eta_{\text{eq}}}{2} \left( \frac{a}{a_{\text{eq}}} \right) & \propto a \quad \text{radiation-dominated}, \\ \frac{2\sqrt{2}}{a_{\text{eq}} H_{\text{eq}}} \left( \frac{a}{a_{\text{eq}}} \right)^{1/2} = (1 + \sqrt{2})\eta_{\text{eq}} \left( \frac{a}{a_{\text{eq}}} \right)^{1/2} & \propto a^{1/2} \quad \text{matter-dominated}. \end{cases} \quad (29.36)$$

The ratio of the comoving horizon distance  $\eta$  to the comoving Hubble distance  $1/(aH)$  is

$$\eta aH = \frac{2\sqrt{1+a/a_{\text{eq}}}}{1+\sqrt{1+a/a_{\text{eq}}}} , \quad (29.37)$$

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 29.5** What is meant by the horizon in cosmology? See §10.21.

**Exercise 29.6 Redshift of matter-radiation equality.**

1. Argue that the redshift  $z_{\text{eq}}$  of matter-radiation equality is given by

$$1 + z_{\text{eq}} = \frac{a_0}{a_{\text{eq}}} = ? \Omega_m h^2 , \quad (29.38)$$

where  $\Omega_m$  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  $g_\rho$ , equation (10.149b). Should this  $g_\rho$  be that now, or that at matter-radiation equality?

2. Show that the ratio  $H_{\text{eq}}/H_0$  of the Hubble parameter at matter-radiation equality to that today is

$$\frac{H_{\text{eq}}}{H_0} = \sqrt{2\Omega_m} (1 + z_{\text{eq}})^{3/2} . \quad (29.39)$$

**Solution.** The redshift  $z_{\text{eq}}$  of matter-radiation equality is given by

$$1 + z_{\text{eq}} = \frac{\Omega_m}{\Omega_r} = \frac{45c^5\hbar^3\Omega_m H_0^2}{4\pi^3 G g_\rho (kT_0)^4} = 8.093 \times 10^4 \frac{\Omega_m h^2}{g_\rho} = 3400 \left( \frac{g_\rho}{3.36} \right)^{-1} \left( \frac{\Omega_m h^2}{0.142} \right) , \quad (29.40)$$

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 (10.149b). The value  $\Omega_m h^2 = 0.142 \pm 0.001$  is from Ade (2015).

## 29.5 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 (29.14), which expresses energy-momentum conservation of a fluid, reduces for  $w = 0$  to

$$\left( \frac{d^2}{d\eta^2} + \frac{\dot{a}}{a} \frac{d}{d\eta} \right) (\delta_c - 3\Phi) = -k^2 \Psi . \quad (29.41)$$

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  $v_c \propto 1/a$ , equation (29.13b), and the dark matter density would then evolve as  $\dot{\delta}_c = kv_c \propto a^{-1}$ , equation (29.13a).

In the radiation-dominated epoch, where  $\eta \propto a$ , this leads to a logarithmic growth in the overdensity  $\delta_c$ , 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  $\delta_c$  would freeze out at a constant value, in the absence of a driving potential.

More generally, equation (29.41) is a linear differential equation for  $\delta_c - 3\Phi$  driven by a potential  $\Psi$ . You will find the solution to this equation for a prescribed potential  $\Psi$  in Exercise 29.7.

**Exercise 29.7 Generic behaviour of dark matter.** Find the homogeneous solutions of equation (29.41) for  $\delta_c - 3\Phi$  with horizon distance  $\eta$  related to cosmic scale factor  $a$  by equation (29.33). Hence find the retarded Green's function of the equation. Write down the general solution of equation (29.41) as an integral over the Green's function. Solve for the case of constant potential  $\Psi$ .

**Solution.** The general solution of equation (29.41) is, in units  $a_{\text{eq}} = 1$ ,

$$\delta_c(a) - 3\Phi(a) = A_0 + A_1 \ln x + 2k^2 \int_0^x \Psi(a') \ln \left( \frac{x'}{x} \right) a'^2 \frac{dx'}{x'} , \quad (29.42)$$

where  $A_0$  and  $A_1$  are constants, and

$$x \equiv \exp \left( \frac{1}{\sqrt{2}} \int \frac{d\eta}{a} \right) = \frac{a}{(1 + \sqrt{1 + a})^2} = \frac{\eta}{\eta + 4\sqrt{2}} , \quad (29.43)$$

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 (29.42) reduces to

$$\delta_c(a) - 3\Phi(a) \rightarrow \begin{cases} A_0 + A_1 \ln(a/4) + 2k^2 \int_0^a \Psi(a') \ln \left( \frac{a'}{a} \right) a' da' & (a \ll 1) , \\ B_0 - 2B_1 a^{-1/2} + 4k^2 \int_0^a \Psi(a') \left( 1 - \sqrt{\frac{a'}{a}} \right) da' & (a \gg 1) , \end{cases} \quad (29.44)$$

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.

## 29.6 Generic behaviour of radiation

Before recombination, photons are tightly coupled to baryons through non-relativistic 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 (29.14) then reduces to

$$\left( 3 \frac{d^2}{d\eta^2} + k^2 \right) (\Theta_0 - \Phi) = -k^2 (\Psi + \Phi) . \quad (29.45)$$

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

$$c_s = \frac{\omega}{k} = \sqrt{\frac{1}{3}} . \quad (29.46)$$

Define the sound horizon distance  $\eta_s$  by

$$\eta_s \equiv c_s \eta = \frac{\eta}{\sqrt{3}} . \quad (29.47)$$

In terms of the sound horizon distance  $\eta_s$ , the differential equation (29.45) becomes

$$\left( \frac{d^2}{d\eta_s^2} + k^2 \right) (\Theta_0 - \Phi) = -k^2(\Psi + \Phi) . \quad (29.48)$$

Equation (29.48) 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 29.8.

**Exercise 29.8 Generic behaviour of radiation.** Find the homogeneous solutions of equation (29.48). Hence find the retarded Green's function of the equation. Write down the general solution of equation (29.48) as an integral over the Green's function. Convince yourself that  $\Theta_0 - \Phi$  oscillates about  $-(\Psi + \Phi)$ .

**Solution.** The general solution of equation (29.48) is, with  $y \equiv k\eta_s$ ,

$$\Theta_0(y) - \Phi(y) = A_0 \cos y + A_1 \sin y - \int_0^y [\Psi(y') + \Phi(y')] \sin(y - y') dy' , \quad (29.49)$$

where  $A_0$  and  $A_1$  are constants.

**Concept question 29.9 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 (32.89) 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 31.7 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.

## 29.7 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):

$$\dot{\delta}_c - k v_c = 3 \dot{\Phi} , \quad (29.50a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c = -k \Psi . \quad (29.50b)$$

The next two equations express conservation of energy and momentum for radiation (subscript r), which includes both photons and neutrinos:

$$\dot{\Theta}_0 - k \Theta_1 = \dot{\Phi} , \quad (29.51a)$$

$$\dot{\Theta}_1 + \frac{k}{3} \Theta_0 = -\frac{k}{3} \Psi . \quad (29.51b)$$

The final equation is the Einstein energy equation:

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 (\bar{\rho}_c \delta_c + 4\bar{\rho}_r \Theta_0) , \quad (29.52)$$

where

$$F \equiv \frac{\dot{a}}{a} \Psi + \dot{\Phi} . \quad (29.53)$$

In place of one of the equations (29.50)–(29.52) it is sometimes convenient to use the Einstein momentum equation

$$-kF = 4\pi G a^2 (\bar{\rho}_c v_c + 4\bar{\rho}_r \Theta_1) , \quad (29.54)$$

which, because the matter and radiation equations (29.50) and (29.51) 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 (28.40d) implies that the scalar potentials  $\Psi$  and  $\Phi$  are equal,

$$\Psi = \Phi . \quad (29.55)$$

**Exercise 29.10 Program the equations for the simplest set of cosmological assumptions.** Write computer code that integrates numerically the evolution equations (29.50)–(29.52). You will be generalizing this code later 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  $\delta_c$  and  $\Theta_0$ . 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 §29.10), 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  $-2\Phi$  together to bring out the fact that the

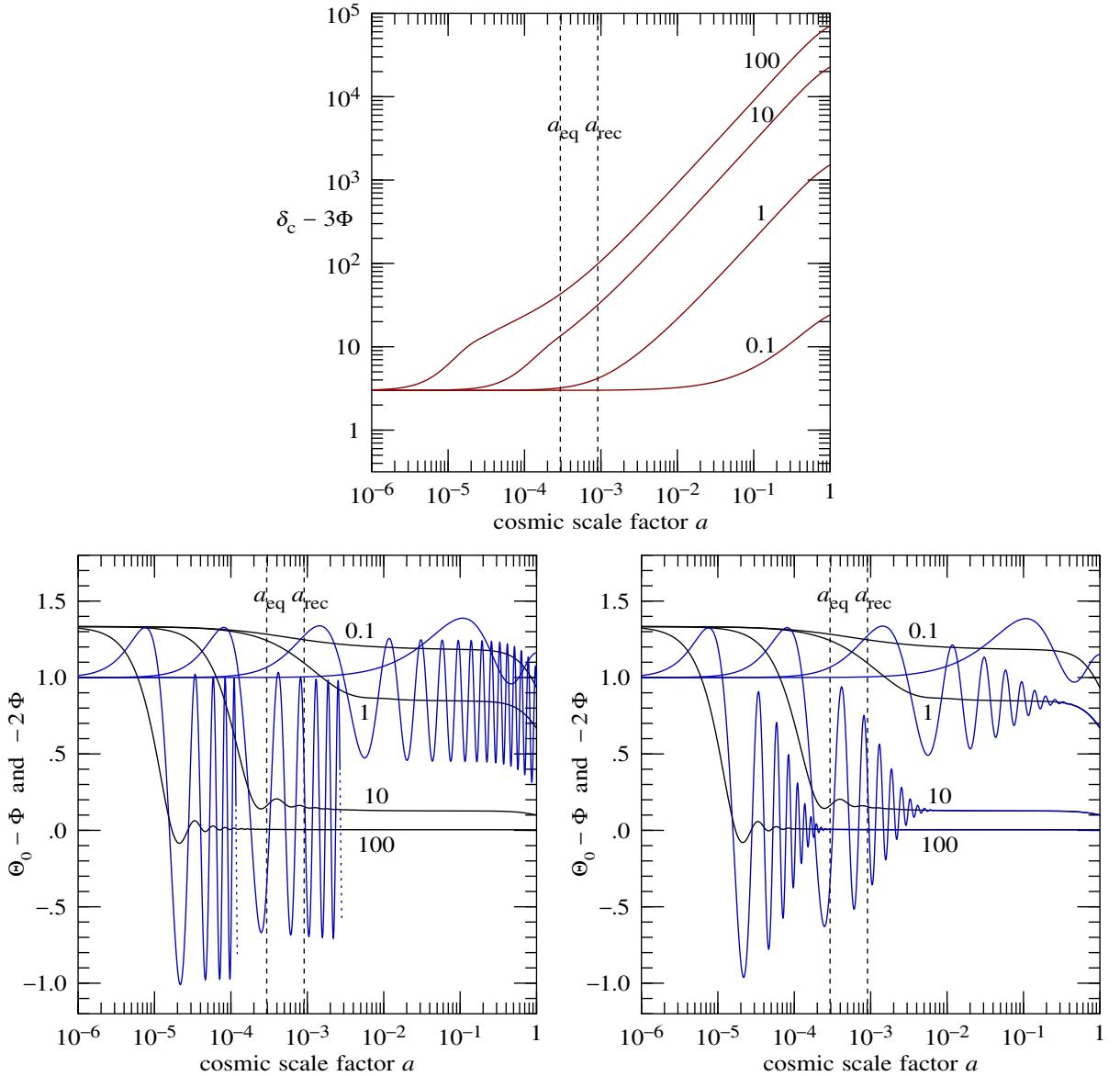


Figure 29.1 (Top) dark matter overdensity  $\delta_c - 3\Phi$ , (bottom left) radiation monopole  $\Theta_0 - \Phi$ , and (bottom right) radiation monopole  $\Theta_0 - \Phi$  with artificial damping, as a function of cosmic scale factor  $a$  in units  $a_0 = 1$ , in the simple approximation, for several wavenumbers  $k$ . The cosmological model is a flat  $\Lambda$ CDM model with concordance parameters  $\Omega_\Lambda = 0.69$  and  $\Omega_m = 0.31$ , and adiabatic initial conditions (see §31.3). The radiation monopole  $\Theta_0 - \Phi$  (blue) is plotted along with minus twice the gravitational potential,  $-2\Phi$  (black), to bring out the fact that the former oscillates about the latter, as expected from equation (29.45). 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 31.3, and using a Boltzmann computation, Figure 32.3.

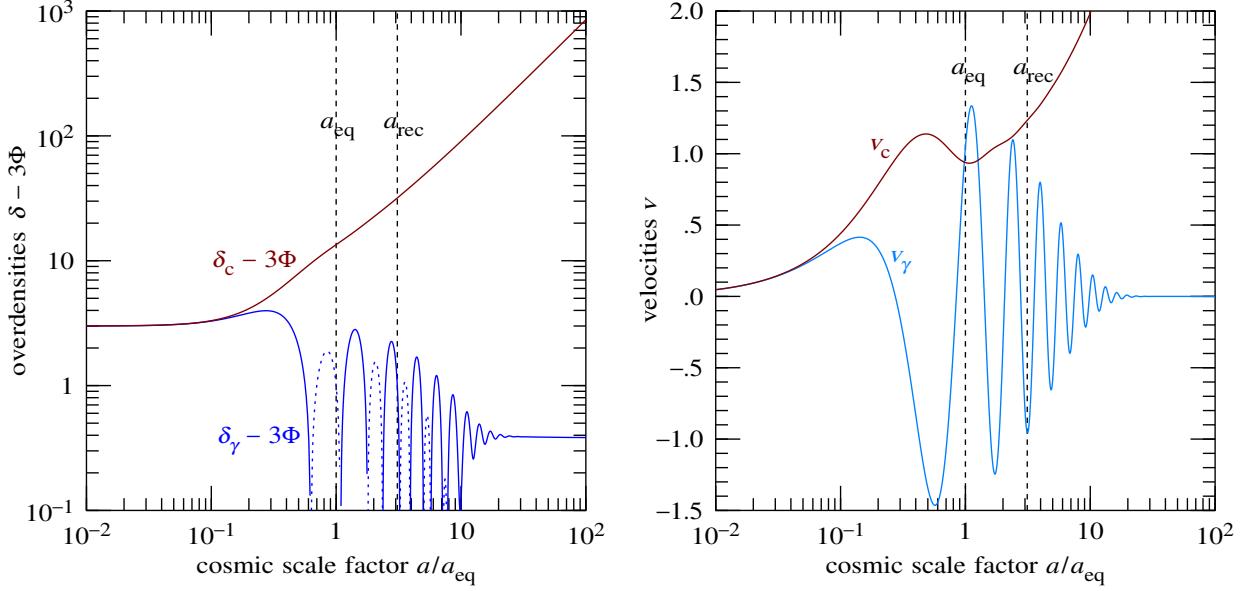


Figure 29.2 (Left) Overdensities  $\delta - 3\Phi$ , and (right) bulk velocities  $v$  in the simple approximation with artificial damping, as a function of cosmic scale factor  $a/a_{\text{eq}}$ , at wavenumber  $k/(a_{\text{eq}} H_{\text{eq}}) = 10$ , for non-baryonic dark matter (c) and radiation ( $\gamma$ ). 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 (29.15). The results may be compared to those from the hydrodynamic approximation, Figure 31.1, and a Boltzmann computation, Figure 32.1.

former oscillates about the latter, as expected from Exercise 29.8. 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 (29.51b) to

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 + \Psi) = -2k\kappa\Theta_1 , \quad (29.56)$$

where  $\kappa$  is a dimensionless damping coefficient that becomes large when the fluctuation is well inside the horizon,  $k\eta \gg 1$ ,

$$\kappa = \epsilon k\eta , \quad (29.57)$$

with  $\epsilon$  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 §31.5. The introduction of damping anticipates, but is not an adequate substitute for, the physical processes of damping addressed in Chapter 31.

**Solution.** Figure 29.1 shows the dark matter overdensity, radiation monopole, and potential for a flat  $\Lambda$ CDM

model with  $\Omega_\Lambda = 0.69$  and  $\Omega_c = 0.31$ , and adiabatic initial conditions. The radiation monopole is shown both without (bottom left panel) and with (bottom right panel) artificial damping. Figure 29.2 illustrates the overdensity and bulk velocity of each of matter and radiation for the same model at an illustrative wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ .

---

## 29.8 On the numerical computation of cosmological power spectra

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 30);
3. uses a Boltzmann code (Chapter 32) to calculate the evolution of non-baryonic cold dark matter, baryonic matter, photons, and neutrinos at each of  $\sim 200$  wavenumbers  $k$ ;
4. pre-calculates tables of spherical Bessel functions  $j_\ell$ ;
5. computes CMB transfer functions  $T_\ell(\eta_0, k)$ , equation (33.18), by integrating source functions over Bessel functions at each of  $\sim 2000$  wavenumbers  $k$  and  $\sim 100$  harmonics  $\ell$  (Chapter 33);
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 (33.33).

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  $j_\ell$  are inexplicably slow for large  $\ell$ , 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.

## 29.9 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 (29.50)–(29.55).

Figure 29.3 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  $\eta$  increases monotonically.

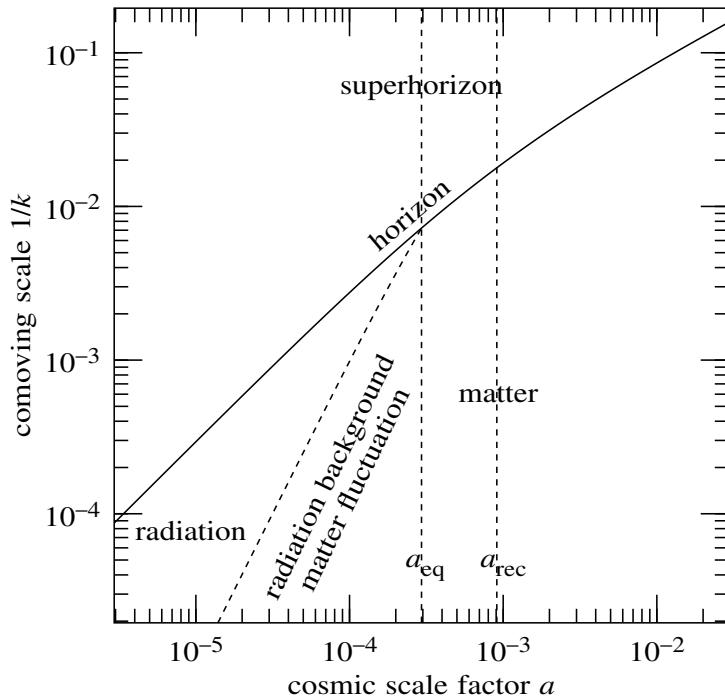


Figure 29.3 Various regimes in the evolution of fluctuations. The line increasing diagonally from bottom left to top right is the comoving horizon distance  $\eta$ . Above this line are superhorizon fluctuations, whose comoving wavelengths exceed the horizon distance, while below the line are subhorizon 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 tend 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 (29.50)–(29.52).

The regimes covered below are:

1. Superhorizon scales, §29.10.
2. Radiation-dominated:
  - a. adiabatic initial conditions, §29.11;
  - b. isocurvature initial conditions, §29.12.
3. Subhorizon scales, §29.13.
4. Fluctuations that enter the horizon in the matter-dominated epoch, §29.14.
5. Matter-dominated regime, §29.15.
6. Baryons post-recombination, §29.16.
7. Matter with dark energy, §29.17.
8. Matter with dark energy and curvature, §29.18.

## 29.10 Superhorizon scales

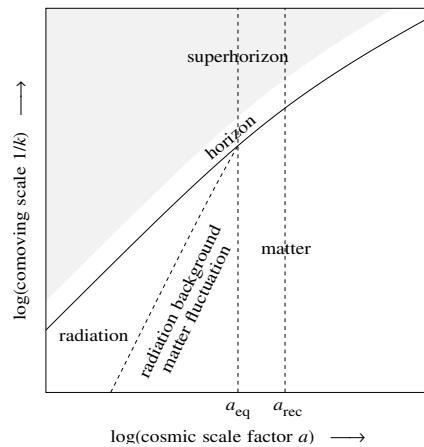


Figure 29.4 Superhorizon scales.

At sufficiently early times, any mode is outside the horizon,  $k\eta < 1$ . In the superhorizon limit  $k\eta \ll 1$ , the

evolution equations (29.50a)–(29.52) reduce to

$$\dot{\delta}_c = 3\dot{\Phi} , \quad (29.58a)$$

$$\dot{\Theta}_0 = \dot{\Phi} , \quad (29.58b)$$

$$-3\frac{\dot{a}}{a}F = 4\pi G a^2 (\bar{\rho}_c \delta_c + 4\bar{\rho}_r \Theta_0) . \quad (29.58c)$$

In effect, the dark matter velocity  $v_c$  and radiation dipole  $\Theta_1$  are negligibly small at superhorizon scales,

$$v_c = \Theta_1 = 0 . \quad (29.59)$$

The first two of equations (29.58) imply that the dark matter overdensity  $\delta_c$  and radiation monopole  $\Theta_0$  are related to the potential  $\Phi$  by

$$\frac{1}{3}\delta_c - \Phi = \zeta_c , \quad (29.60a)$$

$$\Theta_0 - \Phi = \zeta_r , \quad (29.60b)$$

where  $\zeta_c$  and  $\zeta_r$  are constants set by initial conditions. Plugging the solutions (29.60) into the Einstein energy equation (29.58c), and replacing derivatives with respect to horizon distance  $\eta$  with derivatives with respect to cosmic scale factor  $a$ ,

$$\frac{\partial}{\partial\eta} = \dot{a}\frac{\partial}{\partial a} = a^2 H \frac{\partial}{\partial a} , \quad (29.61)$$

with the Hubble parameter  $H$  from equation (29.32) gives the first order differential equation, in units  $a_{\text{eq}} = 1$ ,

$$2a(1+a)\Phi' + (6+5a)\Phi + 4\zeta_r + 3\zeta_c a = 0 , \quad (29.62)$$

where prime ' denotes differentiation with respect to cosmic scale factor,  $d/da$ . The solution to equation (29.62) that is finite at  $a = 0$  is

$$\Phi = -\frac{2}{3}\zeta_r + \left(\frac{2}{3}\zeta_r - \frac{3}{5}\zeta_c\right)f , \quad (29.63)$$

where  $f(a)$  is the function

$$f \equiv 1 - \frac{2}{a} + \frac{8}{a^2} + \frac{16}{a^3} - \frac{16\sqrt{1+a}}{a^3} = \frac{a(6+a+4\sqrt{1+a})}{(1+\sqrt{1+a})^4} , \quad (29.64)$$

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

$$\Phi(0) = -\frac{2}{3}\zeta_r , \quad \Phi(\text{late}) = -\frac{3}{5}\zeta_c . \quad (29.65)$$

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.

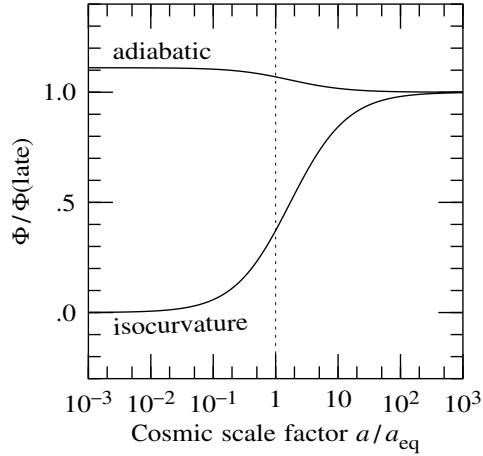


Figure 29.5 Evolution of the scalar potential  $\Phi$  at superhorizon scales, equations (29.69), from radiation-dominated to matter-dominated. The scale for the potential is normalized to its value  $\Phi(\text{late})$  at late times  $a \gg a_{\text{eq}}$ .

There are adiabatic and isocurvature initial conditions. Inflation generically produces adiabatic fluctuations, in which matter and radiation fluctuate together,

$$\zeta_c = \zeta_r \quad \text{adiabatic} , \quad (29.66)$$

so that

$$\delta_c(0) = 3\Theta_0(0) = -\frac{3}{2}\Phi(0) = \zeta_c \quad \text{adiabatic} . \quad (29.67)$$

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

$$\zeta_r = 0 \quad \text{isocurvature} . \quad (29.68)$$

The adiabatic and isocurvature solutions for the superhorizon potential  $\Phi$ , equation (29.63), are

$$\Phi_{\text{ad}} = \zeta_c \left( -\frac{2}{3} + \frac{1}{15}f \right) , \quad (29.69a)$$

$$\Phi_{\text{iso}} = -\frac{3}{5}\zeta_c f , \quad (29.69b)$$

with  $f$  given by equation (29.64). Figure 29.5 shows the evolution of the potential  $\Phi$  from equations (29.69), 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.

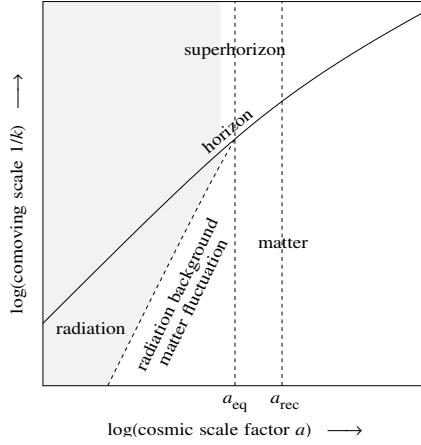


Figure 29.6 Radiation-dominated regime.

## 29.11 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 (29.51), (29.52), and (29.54),

$$\dot{\Theta}_0 - k\Theta_1 = \dot{\Phi} , \quad (29.70a)$$

$$-3\frac{\dot{a}}{a}F - k^2\Phi = 16\pi G a^2 \bar{\rho}_r \Theta_0 , \quad (29.70b)$$

$$-kF = 16\pi G a^2 \bar{\rho}_r \Theta_1 , \quad (29.70c)$$

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 (29.36). Inserting  $\Theta_0$  and  $\Theta_1$  from the Einstein energy and momentum equations (29.70b) and (29.70c) into the radiation monopole equation (29.70a) gives a second order differential equation for the potential  $\Phi$ ,

$$\ddot{\Phi} + \frac{4}{\eta}\dot{\Phi} + \frac{k^2}{3}\Phi = 0 . \quad (29.71)$$

Equation (29.71) 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  $\eta$  multiplied

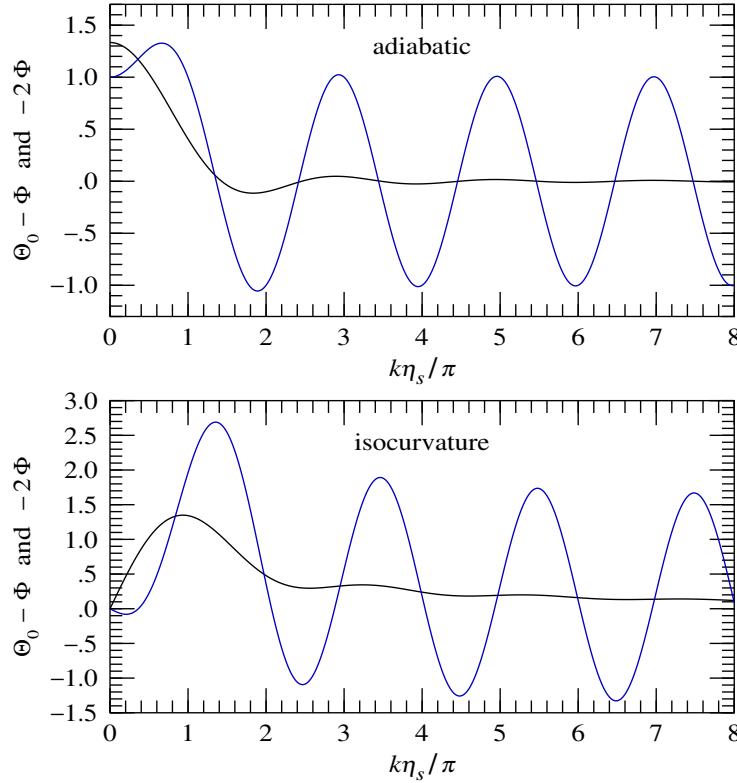


Figure 29.7 The potential  $\Phi$  and radiation monopole  $\Theta_0$  for modes that enter the sound horizon  $k\eta_s = 1$  during the radiation-dominated regime well before matter-radiation equality, for (top) adiabatic initial conditions, equations (29.74) and (29.75), and (bottom) isocurvature initial conditions, equations (29.81) and (29.82). The quantities shown are (blue)  $\Theta_0 - \Phi$  and (black)  $-2\Phi$ , to illustrate that the former oscillates about the latter as expected from equation (29.45). The difference,  $(\Theta_0 - \Phi) - (-2\Phi) = \Theta_0 + \Phi$ , which is the temperature  $\Theta_0$  redshifted by the potential  $\Phi$ , is (for  $\Psi = \Phi$ ) the monopole contribution to the temperature fluctuation of the CMB, equation (33.15). The units of  $\Phi$  and  $\Theta_0$  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 (29.71) are

$$\Phi_{\text{grow}} = \frac{3j_1(y)}{y} = \frac{3(\sin y - y \cos y)}{y^3}, \quad (29.72a)$$

$$\Phi_{\text{decay}} = -\frac{j_{-2}(y)}{y} = \frac{\cos y + y \sin y}{y^3}, \quad (29.72b)$$

where the dimensionless parameter  $y$  is the wavenumber  $k$  multiplied by the sound horizon distance  $\eta_s$ ,

$$y \equiv k\eta_s = \frac{k\eta}{\sqrt{3}} = \sqrt{\frac{2}{3}} \frac{k}{a_{\text{eq}} H_{\text{eq}}} \frac{a}{a_{\text{eq}}}, \quad (29.73)$$

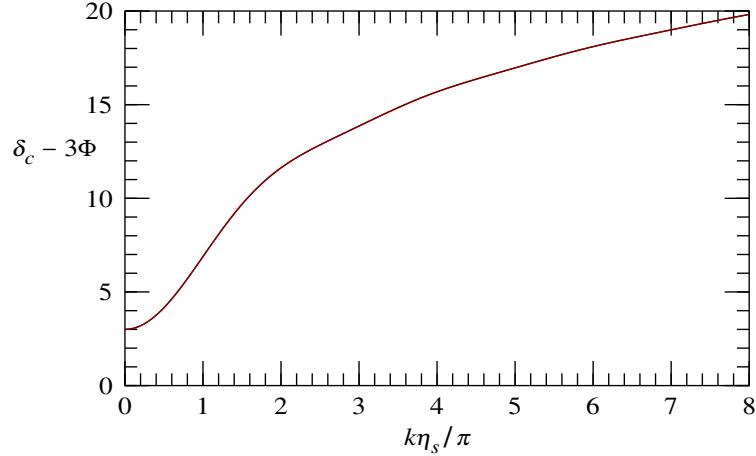


Figure 29.8 Evolution of the dark matter overdensity  $\delta_c - 3\Phi$  for a mode that enters the horizon during the radiation-dominated regime, for adiabatic initial conditions. Like the radiation fluctuation  $\Theta_0$  illustrated in the top panel of Figure 29.7, the matter fluctuation  $\delta_c$  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

$$\Phi = \Phi(0) \Phi_{\text{grow}}. \quad (29.74)$$

The growing solution (29.72a) shows that, after a mode enters the sound horizon the scalar potential  $\Phi$  oscillates with an envelope that decays as  $y^{-2}$ . Physically, relativistically propagating sound waves tend to suppress the gravitational potential  $\Phi$ . The suppression of the potential is responsible for the turnover in the observed power spectrum of matter fluctuations today from large to small scales.

The radiation monopole  $\Theta_0$  can be inferred either from the Einstein equation (29.70b) with the solutions (29.72) for the potential  $\Phi$ , or from the Green's function solution (29.49) 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 (29.74) is

$$\Theta_0 - \Phi = \zeta_r \frac{(2 \sin y - y \cos y)}{y}. \quad (29.75)$$

The top panel of Figure 29.7 shows the growing mode potential  $\Phi$ , equation (29.74), and the radiation monopole  $\Theta_0$ , equation (29.75). The Figure plots these two quantities in the form  $-2\Phi$  and  $\Theta_0 - \Phi$  in order to bring out the fact that  $\Theta_0 - \Phi$  oscillates about  $-2\Phi$ , in accordance with equation (29.45). After a mode

is well inside the sound horizon,  $y \gg 1$ , the radiation monopole oscillates with constant amplitude,

$$\Theta_0 = -\zeta_r \cos y \quad \text{for } y \gg 1 . \quad (29.76)$$

Fluctuations in the dark matter are driven by the gravitational potential of the radiation. The radiation-dominated Green's function solution (29.44) for the dark matter fluctuation  $\delta_c$  driven by the growing mode potential (29.74) and satisfying adiabatic initial conditions (29.67) is

$$\delta_c - 3\Phi = 6\zeta_r \left( \frac{\sin y}{y} - \frac{1}{2} + \text{Cin } y \right) , \quad (29.77)$$

where  $\text{Cin } y \equiv \int_0^y (1 - \cos x) dx/x$  is the cosine integral. Figure 29.8 shows the density fluctuation (29.77). Once the mode is well inside the sound horizon,  $y \gg 1$ , the dark matter density  $\delta_c$ , equation (29.77), evolves as, from the asymptotic behaviour  $\text{Cin } y \sim \ln y + \gamma$  with  $\gamma \equiv 0.5772\dots$  Euler's constant,

$$\delta_c - 3\Phi = 6\zeta_r \left( \ln y + \gamma - \frac{1}{2} \right) \quad \text{for } y \gg 1 , \quad (29.78)$$

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 29.11 Radiation-dominated fluctuations.

1. Confirm equation (29.71). You might like to start by seeking a solution using the monopole and dipole radiation equations (29.51) along with the Einstein energy equation (29.52) including only radiation, namely equation (29.70b). Then try the solution advocated in the text, namely use the Einstein momentum equation (29.70c) 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 (29.71) are as given in the text, equations (29.72).
3. The initial condition for the temperature monopole is determined by equation (29.60b),  $\Theta_0(0) - \Phi(0) = \zeta_r$ , where  $\zeta_r$  is some constant, the initial radiation entropy fluctuation set up during inflation. Find the initial conditions for the scalar potentials  $\Psi$  and  $\Phi$  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 (29.49) for  $\Theta_0 - \Phi$  satisfying the requisite boundary conditions is equation (29.75). Plot the solution for  $\Theta_0 - \Phi$ , along with  $-2\Phi$ . Confirm that  $\Theta_0 - \Phi$  oscillates around  $-2\Phi$ .
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?

## 29.12 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 (29.52) as giving the radiation monopole  $\Theta_0$ , and the Einstein momentum equation (29.54) as giving the radiation dipole  $\Theta_1$ . Insert these into the radiation monopole equation (29.51a), and eliminate the  $\dot{\delta}_c$  terms using the dark matter density equation (29.50a). The result is, in units  $a_{\text{eq}} = 1$ ,

$$2a(1+a)\Phi'' + (8+9a)\Phi' + 2\left(1 + \frac{2k^2 a}{3}\right)\Phi + \delta_c = 0 , \quad (29.79)$$

where prime ' denotes differentiation with respect to cosmic scale factor  $a$ . Equation (29.79) 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  $\Phi$  is small compared to  $\delta_c$ , while  $\delta_c$  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  $k^2 a \gg 1$ , equation (29.79) simplifies to

$$2a\Phi'' + 8\Phi' + \frac{4k^2 a}{3}\Phi + \delta_c = 0 . \quad (29.80)$$

For constant  $\delta_c = \delta_c(0) = 3\zeta_c$ , the solution of equation (29.80) vanishing at  $a = 0$  is, with  $y$  given by equation (29.73),

$$\Phi = -\frac{3\sqrt{3}\zeta_c}{\sqrt{2}k} \frac{1 - \cos y - y \sin y + \frac{1}{2}y^2}{y^3} . \quad (29.81)$$

With units restored,  $k$  is  $k/(a_{\text{eq}} H_{\text{eq}})$ . The Green's function solution (29.49) for the difference  $\Theta_0 - \Phi$  between the radiation monopole and potential driven by the potential (29.81) is

$$\Theta_0 - \Phi = \frac{3\sqrt{3}\zeta_c}{\sqrt{2}k} \frac{(1 - \cos y - \frac{1}{2}y \sin y)}{y} . \quad (29.82)$$

Equations (29.81) and (29.82) 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 (29.82) oscillates with constant amplitude,

$$\Theta_0 = -\frac{3\sqrt{3}\zeta_c}{2\sqrt{2}k} \sin y \quad y \gg 1 . \quad (29.83)$$

The lower panel of Figure 29.7 shows the potential  $\Phi$ , equation (29.81), and the radiation monopole  $\Theta_0$  from equation (29.82), again plotted as  $\Theta_0 - \Phi$  and  $-2\Phi$  to bring out the fact that  $\Theta_0 - \Phi$  oscillates about  $-2\Phi$ . Whereas for adiabatic initial conditions the radiation monopole oscillated as  $\cos y$  well inside the sound

horizon, equation (29.76), for isocurvature initial conditions it oscillates as  $\sin y$  well inside the sound horizon, equation (29.83).

The solution (29.81) for the potential  $\Phi$  was derived from equation (29.80) on the assumption of constant  $\delta_c$ . The accuracy of the approximation may be checked by calculating the radiation-dominated Green's function solution (29.44) for  $\delta_c$  driven by this potential, which is

$$\delta_c - 3\Phi = 3\zeta_c \left( 1 + a \frac{3 - 3 \cos y - 3y \operatorname{Si} y + \frac{3}{2}y^2}{y^2} \right), \quad (29.84)$$

where  $\operatorname{Si} y \equiv \int_0^y \sin x dx / x$  is the sine integral. Equation (29.84) 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  $\delta_c$  itself is not constant, because  $\Phi$  varies. However,  $\Phi$  from equation (29.81) is of order  $a\delta_c(0)$  for any  $y$ , and the small order  $a$  correction to  $\delta_c$  leads to corrections of next order  $a^2$  to  $\Phi$ ,  $\Theta_0 - \Phi$ , and  $\delta_c - 3\Phi$ , and can be neglected.

## 29.13 Subhorizon scales

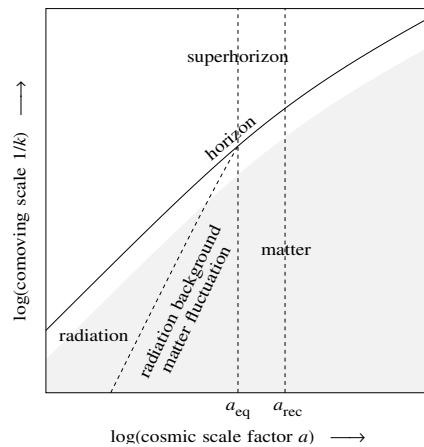


Figure 29.9 Subhorizon scales.

After a mode enters the horizon, the radiation fluctuation  $\Theta_0$  oscillates, but the non-baryonic cold dark matter fluctuation  $\delta_c$  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 §29.12, 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  $\Phi$  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 (29.52) as giving the dark matter overdensity  $\delta_c$ , and the Einstein momentum equation (29.54) as giving the dark matter velocity  $v_c$ . Insert these into the dark matter density equation (29.50a) and eliminate the  $\dot{\Theta}_0$  terms using the radiation monopole equation (29.51a). The result is, in units  $a_{\text{eq}} = 1$ ,

$$2a^2(1+a)\Phi'' + a(6+7a)\Phi' - 2\Phi - 4\Theta_0 = 0 , \quad (29.85)$$

where prime ' denotes differentiation with respect to cosmic scale factor  $a$ . Equation (29.85) 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  $\Theta_0$  oscillates about an average value of  $-\Phi$  (since  $\Theta_0 - \Phi$  oscillates about  $-2\Phi$ , as noted in §29.6):

$$\langle \Theta_0 \rangle = -\Phi . \quad (29.86)$$

Inserting this cycle-averaged value of  $\Theta_0$  into equation (29.85) gives the Meszaros differential equation

$$2(1+a)a^2\Phi'' + (6+7a)a\Phi' + 2\Phi = 0 . \quad (29.87)$$

The solutions of Meszaros' differential equation (29.87) are a linear combination of growing and decaying solutions

$$\Phi_{\text{grow}} = -\frac{3}{4k^2a} \left( 1 + \frac{3a}{2} \right) , \quad (29.88a)$$

$$\Phi_{\text{decay}} = -\frac{3}{4k^2a} \left\{ \left( 1 + \frac{3a}{2} \right) \ln \left[ \frac{(\sqrt{1+a}+1)^2}{a} \right] - 3\sqrt{1+a} \right\} . \quad (29.88b)$$

A constant factor of  $-3/(4k^2)$  has been included in the potential, arbitrarily, to simplify the overall factor in the resulting solution for the dark matter overdensity  $\delta_c$ , equations (29.90). The solutions for  $\delta_c$  driven by the growing and decaying potentials (29.88) are, from the Green's function solution (29.42), in units  $a_{\text{eq}} = 1$ ,

$$\delta_c - 3\Phi = -\frac{4k^2a}{3}\Phi , \quad (29.89)$$

which holds for both growing and decaying modes. The solutions (29.89) omit possible additional contributions from the homogeneous solutions in equation (29.42), but the regime of interest is modes well inside the horizon,  $ka \gg 1$ , and the omitted contributions become dominated by the solutions (29.89) as the cosmic scale factor  $a$  increases. Explicitly, the growing and decaying modes for  $\delta_c - 3\Phi$  are

$$(\delta_c - 3\Phi)_{\text{grow}} = 1 + \frac{3}{2}a , \quad (29.90a)$$

$$(\delta_c - 3\Phi)_{\text{decay}} = \left( 1 + \frac{3}{2}a \right) \ln \left[ \frac{(\sqrt{1+a}+1)^2}{a} \right] - 3\sqrt{1+a} . \quad (29.90b)$$

The desired solution for the dark matter overdensity  $\delta_c$  is a linear combination of growing and decaying modes,

$$\delta_c - 3\Phi = C_{\text{grow}}(\delta_c - 3\Phi)_{\text{grow}} + C_{\text{decay}}(\delta_c - 3\Phi)_{\text{decay}} . \quad (29.91)$$

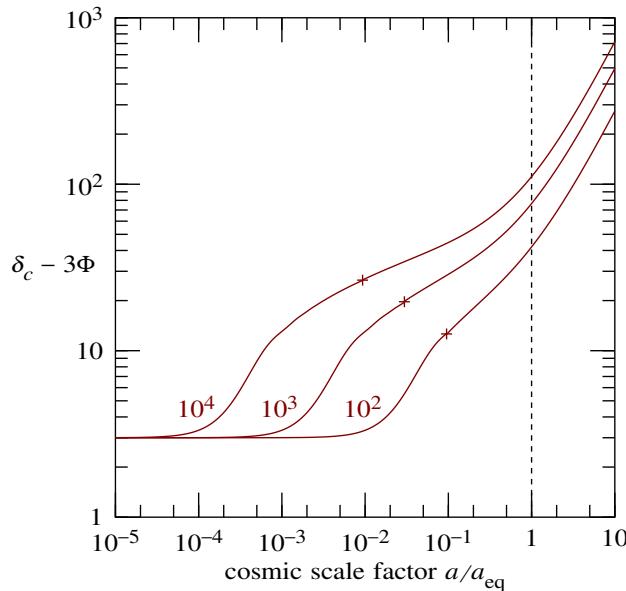


Figure 29.10 After an initial boost on entering the sound horizon, the dark matter overdensity  $\delta_c$  grows logarithmically with cosmic scale factor  $a$  during the radiation-dominated regime, but then linearly with  $a$  after matter-radiation equality  $a = a_{\text{eq}}$ . The curves are labelled 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 (29.77) at small  $a$ , and by the Meszaros solution (29.91) 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  $C_{\text{grow}}$  and  $C_{\text{decay}}$  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 (29.90) simplify to

$$(\delta_c - 3\Phi)_{\text{grow}} = 1 , \quad (\delta_c - 3\Phi)_{\text{decay}} = -\ln(a/4) - 3 \quad \text{for } a \ll 1 . \quad (29.92)$$

It was found in §29.11 that the potential  $\Phi$  in the radiation-dominated regime oscillated with an envelope that decayed as  $\sim a^{-2}$ , equation (29.72a), driving a dark matter overdensity that grew as a combination of linear and logarithmic parts, equation (29.78). The result (29.92) 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  $\delta_c$  is the one that matches smoothly on to the logarithmically growing solution given by equation (29.78). Matching to the adiabatic solution (29.78) for  $\delta_c - 3\Phi$  well inside the horizon determines the constants

$$C_{\text{grow}} = 6\zeta_r \left[ \gamma - \frac{7}{2} + \ln \left( 4\sqrt{\frac{2}{3}} k \right) \right] , \quad C_{\text{decay}} = -6\zeta_r \quad \text{adiabatic} . \quad (29.93)$$

For isocurvature initial conditions,  $\delta_c - 3\Phi$  is sensibly constant in the radiation-dominated regime  $a \ll 1$ , equation (29.84), and only the growing mode is present,

$$C_{\text{grow}} = 3\zeta_c, \quad C_{\text{decay}} = 0 \quad \text{isocurvature}. \quad (29.94)$$

At late times well into the matter-dominated epoch,  $a \gg 1$ , the growing mode of the Meszaros solution dominates,

$$(\delta_c - 3\Phi)_{\text{grow}} = \frac{3}{2}a, \quad (\delta_c - 3\Phi)_{\text{decay}} = \frac{4}{15}a^{-3/2} \quad \text{for } a \gg 1, \quad (29.95)$$

so that the dark matter overdensity  $\delta_c$  at late times is

$$\delta_c - 3\Phi = \frac{3}{2}C_{\text{grow}}a \quad \text{for } a \gg 1. \quad (29.96)$$

The potential  $\Phi$ , equation (29.88a), at late times is constant,

$$\Phi = -\frac{9}{8k^2}C_{\text{grow}} \quad \text{for } a \gg 1. \quad (29.97)$$

The constancy of the potential  $\Phi$ , and the linear growth of the dark matter density  $\delta_c$ , is characteristic of the matter-dominated regime.

Figure 29.10 shows the dark matter overdensity  $\delta_c$  calculated for adiabatic conditions from the radiation-dominated solution (29.77) at small  $a$ , and the Meszaros solution (29.91) at larger  $a$ . The overdensity  $\delta_c$  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  $\Theta_0$  at late times  $a \gg 1$  is, with  $y \equiv k\eta/\sqrt{3}$ ,

$$\Theta_0 = -\Phi - \zeta_r \cos y \quad \text{adiabatic}, \quad (29.98a)$$

$$\Theta_0 = -\Phi - \frac{3\sqrt{3}\zeta_c}{2\sqrt{2}k} \sin y \quad \text{isocurvature}. \quad (29.98b)$$

## 29.14 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  $\Phi$  before entering the horizon is given by the superhorizon solution (29.63), while after entering the horizon the evolution of the potential is dominated by the dark matter density fluctuation. A satisfactory solution for the potential  $\Phi$  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 (29.60b), and inserting this value into the differential equation (29.85). This solution remains an adequate approximation inside the horizon because after horizon crossing the radiation fluctuation  $\Theta_0$  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  $\Phi''$  and  $\Phi'$  dominate the  $\Phi$  and  $\Theta_0$  terms in equation (29.85).

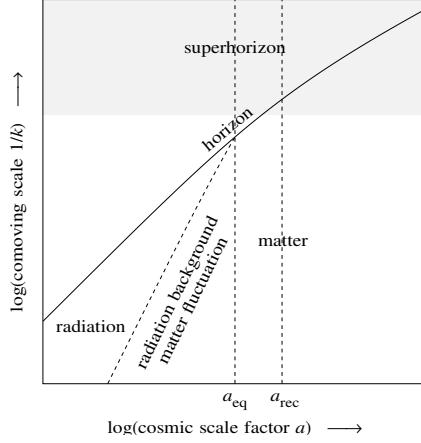


Figure 29.11 Fluctuations that enter the horizon in the matter-dominated regime.

Inserting the superhorizon solution  $\Theta_0 = \Phi + \zeta_r$  into the differential equation (29.85) recovers the superhorizon solution (29.63) for  $\Phi$ , which therefore remains a satisfactory approximation not only outside but also inside the horizon. But inside the horizon, the dark matter overdensity  $\delta_c$  and radiation monopole  $\Theta_0$  driven by this potential are no longer their superhorizon solutions (29.60). Rather, the solution for the dark matter overdensity  $\delta_c$  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 (29.42),

$$\delta_c - 3\Phi = 3\zeta_c + k^2 \left\{ -2a^2 \left( \frac{3}{5}\zeta_c + \Phi \right) + \left( \frac{8}{3}\zeta_r - \frac{16}{5}\zeta_c \right) \left[ 4 \ln \left( \frac{1 + \sqrt{1+a}}{2} \right) - a \right] \right\}. \quad (29.99)$$

Well after matter-radiation equality,  $a \gg 1$ , the dark matter overdensity (29.99) is (note that for large scale modes  $k^2a$  can be small even when  $a \gg 1$ )

$$\delta_c - 3\Phi = 3\zeta_c \left( 1 + \frac{4}{15}k^2a \right) = 3\zeta_c \left( 1 + \frac{1}{30}(k\eta)^2 \right) \quad \text{for } a \gg 1. \quad (29.100)$$

Since  $\Phi_{\text{super(late)}} = -\frac{3}{5}\zeta_c$  for both adiabatic and isocurvature modes, equation (29.65), the overdensity  $\delta_c$  from equation (29.100) is

$$\delta_c = \frac{6}{5}\zeta_c \left( 1 + \frac{2}{3}k^2a \right) = \frac{6}{5}\zeta_c \left( 1 + \frac{1}{12}(k\eta)^2 \right) \quad \text{for } a \gg 1. \quad (29.101)$$

The solution for  $\Theta_0 - \Phi$  driven by the superhorizon potential is, from the Green's function solution (29.49),

$$\begin{aligned} \Theta_0 - \Phi &= \frac{1}{3}\zeta_r(4 - \cos y) + \left( \frac{1}{3}\zeta_r - \frac{3}{10}\zeta_c \right) \left\{ -4 + (4 - y_k^2) \cos y + 3y_k \sin y + y_k^2 \left[ \frac{y_k}{y + y_k} \right. \right. \\ &\quad \left. \left. - y_k f(y_k + y) + 2g(y_k + y) + (y_k \cos y - 2 \sin y)f(y_k) - (2 \cos y + y_k \sin y)g(y_k) \right] \right\}, \end{aligned} \quad (29.102)$$

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

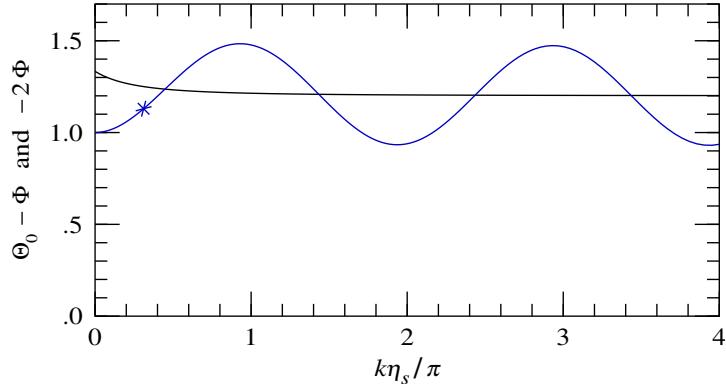


Figure 29.12 Similar to Figure 29.7, but for a mode that enters the sound horizon  $k\eta_s = 1$  during the matter-dominated regime, for adiabatic initial conditions. The mode shown has  $k = \sqrt{3/8} = 0.61$ , which enters the horizon at  $a = 3$ , approximately the time of recombination, marked by a star.

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

$$f(y) \equiv (\pi/2 - \text{Si } y) \cos y + \text{Ci } y \sin y , \quad (29.103a)$$

$$g(y) \equiv (\pi/2 - \text{Si } y) \sin y - \text{Ci } y \cos y . \quad (29.103b)$$

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 29.12 shows  $\Theta_0 - \Phi$  and  $-2\Phi$  from equations (29.102) and (29.63), 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 29.12 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 29.12, 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 31.7 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, §33.1.

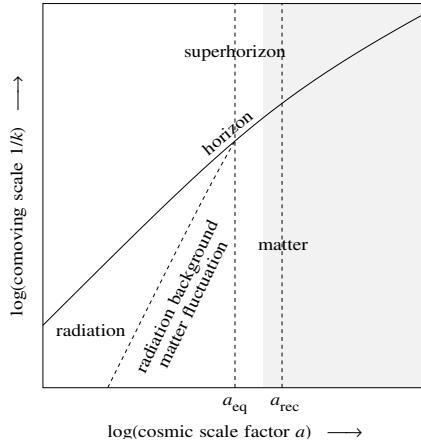


Figure 29.13 Matter-dominated regime.

## 29.15 Matter-dominated regime

After matter-radiation equality, but before curvature or dark energy become important, non-relativistic matter dominates the mass-energy density of the Universe.

In the matter-dominated epoch, the relevant equations are, from equations (29.50), (29.52), and (29.54),

$$\dot{\delta}_c - k v_c = 3 \dot{\Phi} , \quad (29.104a)$$

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 \bar{\rho}_c \delta_c , \quad (29.104b)$$

$$-kF = 4\pi G a^2 \bar{\rho}_c v_c , \quad (29.104c)$$

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 (29.36). Inserting  $\delta_c$  and  $v_c$  from the Einstein energy and momentum equations (29.104b) and (29.104c) into the matter density equation (29.104a) yields a second order differential equation for the potential  $\Phi$

$$\ddot{\Phi} + \frac{6}{\eta} \dot{\Phi} = 0 . \quad (29.105)$$

The general solution of equation (29.105) is a linear combination

$$\Phi = C_{\text{grow}} \Phi_{\text{grow}} + C_{\text{decay}} \Phi_{\text{decay}} \quad (29.106)$$

of growing and decaying solutions

$$\Phi_{\text{grow}} = 1 , \quad \Phi_{\text{decay}} = y^{-5} , \quad (29.107)$$

where the dimensionless parameter  $y$  is, as previously, the wavenumber  $k$  multiplied by the sound horizon distance  $\eta/\sqrt{3}$ . In the matter-dominated regime  $y$  is, in units  $a_{\text{eq}} = H_{\text{eq}} = 1$ ,

$$y \equiv \frac{k\eta}{\sqrt{3}} = 2\sqrt{\frac{2}{3}} ka^{1/2} . \quad (29.108)$$

The constants  $C_{\text{grow}}$  and  $C_{\text{decay}}$  in the solution (29.106) depend on conditions established before the matter-dominated epoch. The corresponding growing and decaying modes for the dark matter overdensity  $\delta_c$  are, from the Einstein energy equation (29.104b),

$$(\delta_c - 3\Phi)_{\text{grow}} = -\left(5 + \frac{1}{2}y^2\right)\Phi_{\text{grow}} = -\left(5 + \frac{4}{3}k^2a\right)\Phi_{\text{grow}} , \quad (29.109a)$$

$$(\delta_c - 3\Phi)_{\text{decay}} = -\frac{1}{2}y^2\Phi_{\text{decay}} = -\frac{4}{3}k^2a\Phi_{\text{decay}} . \quad (29.109b)$$

The behaviour of the growing and decaying modes (29.109) agrees with both the subhorizon Meszaros solution (29.95) and the superhorizon solution (29.100) 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.

## 29.16 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 29.13 you will figure out the behaviour of baryons and non-baryonic cold matter in the approximation that the Universe is matter-dominated.

### Exercise 29.13 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  $\delta_m$ ?
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,

$$\delta_m = f_c\delta_c + f_b\delta_b , \quad (29.110)$$

where the constants  $f_c$  and  $f_b$  are the dark matter and baryon fractions

$$f_c \equiv \frac{\rho_c}{\rho_m} = 1 - f_b, \quad f_b \equiv \frac{\rho_b}{\rho_m}. \quad (29.111)$$

Use the Green's function with the chosen initial conditions to derive solutions for the dark matter and baryon overdensities  $\delta_c$  and  $\delta_b$  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.
- 

## 29.17 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 (29.50) governing the matter overdensity  $\delta_m$  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 (29.52) and (29.54) sourced only by matter,

$$\dot{\delta}_m - k v_m = 3 \dot{\Phi}, \quad (29.112a)$$

$$\dot{v}_m + \frac{\dot{a}}{a} v_m = -k \Phi, \quad (29.112b)$$

$$-3 \frac{\dot{a}}{a} F - k^2 \Phi = 4\pi G a^2 \bar{\rho}_m \delta_m, \quad (29.112c)$$

$$-kF = 4\pi G a^2 \bar{\rho}_m v_m. \quad (29.112d)$$

The factor  $4\pi G a^2 \bar{\rho}_m$  on the right hand side of the two Einstein equations can be written

$$4\pi G a^2 \bar{\rho}_m = \frac{3a_0^3 H_0^2 \Omega_m}{2a}, \quad (29.113)$$

where  $a_0$  and  $H_0$  are the present-day cosmic scale factor and Hubble parameter, and  $\Omega_m$  is the present-day matter density (a constant). Allow the Hubble parameter  $H(a) \equiv \dot{a}/a^2$  to be an arbitrary function

of cosmic scale factor  $a$ . Inserting  $\delta_m$  and velocity  $v_m$  from the Einstein energy and momentum equations (29.112c) and (29.112d) into the matter equations (29.112a) and (29.112b), and taking the overdensity equation (29.112a) minus  $3\dot{a}/a$  times the velocity equation (29.112b), yields the condition

$$a^4 \frac{dH^2}{da} + 3a_0^3 H_0^2 \Omega_m = 0 , \quad (29.114)$$

whose solution is

$$\frac{H^2}{H_0^2} = \frac{\Omega_m}{(a/a_0)^3} + \Omega_\Lambda \quad (29.115)$$

for some constant  $\Omega_\Lambda$ . 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 (29.115), the matter and Einstein equations (29.112) yield a second order differential equation for the potential  $\Phi$ , in units  $a_0 = 1$ :

$$2a(\Omega_m + a^3 \Omega_\Lambda)\Phi'' + (7\Omega_m + 10a^3 \Omega_\Lambda)\Phi' + 6a^2 \Omega_\Lambda \Phi = 0 . \quad (29.116)$$

The growing and decaying solutions to equation (29.116) are, in units  $a_0 = 1$ ,

$$\Phi_{\text{grow}} = \frac{5\Omega_m H_0^2}{2} \frac{H(a)}{a} \int_0^a \frac{da'}{a'^3 H(a')^3} , \quad (29.117a)$$

$$\Phi_{\text{decay}} = \frac{H}{a} . \quad (29.117b)$$

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  $\Phi_{\text{grow}}$  can be expressed as an elliptic integral. The corresponding growing and decaying solutions for the matter overdensity  $\delta_m$  are, again in units  $a_0 = 1$ ,

$$(\delta_m - 3\Phi)_{\text{grow}} = -\frac{2k^2 a}{3\Omega_m H_0^2} \Phi_{\text{grow}} - 5 , \quad (\delta_m - 3\Phi)_{\text{decay}} = -\frac{2k^2 a}{3\Omega_m H_0^2} \Phi_{\text{decay}} . \quad (29.118)$$

For modes well inside the horizon,  $k\eta \sim ka^{1/2}/H_0 \gg 1$ , the relation (29.118) agrees with that (29.124) below.

## 29.18 Matter with dark energy and curvature

Curvature may also play a role after recombination. Observational evidence as of 2014 is 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.

**Concept question 29.14 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 (29.112) then go over to the Newtonian limit,

$$\dot{\delta}_m - k v_m = 0 , \quad (29.119a)$$

$$\dot{v}_m + \frac{\dot{a}}{a} v_m = -k\Phi , \quad (29.119b)$$

$$-k^2\Phi = 4\pi G a^2 \bar{\rho}_m \delta_m . \quad (29.119c)$$

The factor  $4\pi G a^2 \bar{\rho}_m$  in the Einstein equation can be written as equation (29.113). The matter and Einstein equations (29.119) yield a second order equation for the matter overdensity  $\delta_m$ , in units  $a_0 = 1$ :

$$\ddot{\delta}_m + \frac{\dot{a}}{a} \dot{\delta}_m - \frac{3\Omega_m H_0^2}{2} \frac{\delta_m}{a} = 0 . \quad (29.120)$$

Equation (29.120) can be recast as a differential equation with respect to cosmic scale factor  $a$ :

$$\delta''_m + \left( \frac{H'}{H} + \frac{3}{a} \right) \delta'_m - \frac{3\Omega_m H_0^2}{2} \frac{\delta_m}{a^5 H^2} = 0 , \quad (29.121)$$

where  $H \equiv \dot{a}/a^2$  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  $a_0 = 1$ ,

$$\frac{H^2}{H_0^2} = \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda , \quad (29.122)$$

where  $\Omega_m$ ,  $\Omega_k$ , and  $\Omega_\Lambda$  are the (constant) present-day values of the matter, curvature, and dark energy densities. The growing and decaying solutions to equation (29.121) are

$$\delta_{m,\text{grow}} \equiv a g(a) = \frac{5\Omega_m H_0^2}{2} H(a) \int_0^a \frac{da'}{a'^3 H(a')^3} , \quad \delta_{m,\text{decay}} = \frac{H}{H_0} . \quad (29.123)$$

The potential  $\Phi$  is related to the matter overdensity  $\delta_m$  by, again in units  $a_0 = 1$ , equation (29.119c),

$$\Phi = -\frac{3\Omega_m H_0^2}{2k^2} \frac{\delta_m}{a} . \quad (29.124)$$

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,

$$\delta \propto a g(a) , \quad \Phi \propto g(a) . \quad (29.125)$$

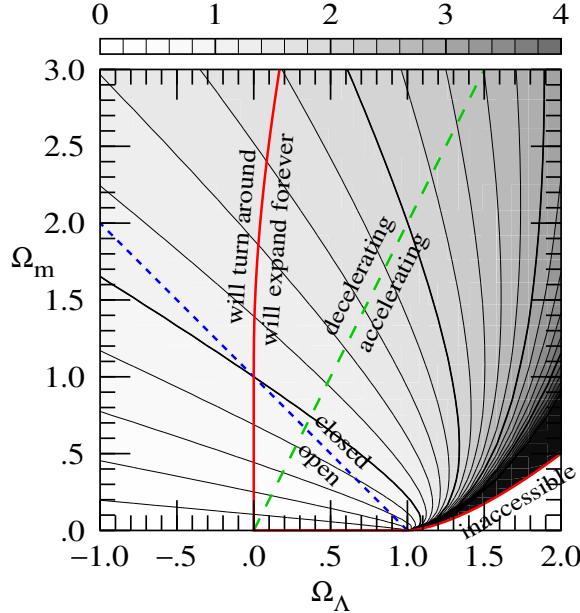


Figure 29.14 Contour plot of the growth factor  $g(a)$  in a universe containing matter, curvature, and a cosmological constant. If the Universe is flat,  $\Omega_k = 0$ , then the Universe evolves from matter-dominated ( $\Omega_m = 1$ ,  $\Omega_\Lambda = 0$ ) to  $\Lambda$ -dominated ( $\Omega_m = 0$ ,  $\Omega_\Lambda = 1$ ) along the (blue) dashed line.

The normalization factor  $\frac{5}{2}\Omega_m H_0^2$  in equation (29.123) 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,

$$g(a) = 1 \quad (a_{\text{rec}} \ll a \ll 1). \quad (29.126)$$

Thus as long as the Universe remains matter-dominated, the potential  $\Phi$  remains constant. Curvature or dark energy causes the potential  $\Phi$  to decrease. Figure 29.14 illustrates the growth factor  $g(a)$  as a function of  $\Omega_m$  and  $\Omega_\Lambda$ .

It should be emphasized that the growing and decaying solutions (29.123) are valid *only* for the case of matter plus curvature plus constant density dark energy, where the Hubble parameter takes the form (29.122). 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)

$$g \approx \frac{5\Omega_m}{2 \left[ \Omega_m^{4/7} - \Omega_\Lambda + \left(1 + \frac{1}{2}\Omega_m\right) \left(1 + \frac{1}{70}\Omega_\Lambda\right) \right]}, \quad (29.127)$$

where  $\Omega_x$  are densities at the epoch being considered (such as the present,  $a = a_0$ ).

## 29.19 Primordial power spectrum

Initial conditions from inflation are conveniently characterized in terms of the gauge-invariant fluctuation  $\zeta$  defined by equation (29.23), which has the property that it remains constant during evolution at superhorizon scales. The fluctuation  $\zeta$  is commonly called the primordial curvature fluctuation. According to the inflationary paradigm, fluctuations in  $\zeta$  are generated by quantum fluctuations in the inflaton field that drives inflation. The amplitude  $\zeta$  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  $\zeta$  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, §29.22.6, are ubiquitous in statistics as a consequence of the Central Limit Theorem (CLT), §29.22.5. 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  $\xi_\zeta$  of the dimensionless fluctuation  $\zeta$  is the same on all scales,

$$\langle \zeta(\mathbf{x}')\zeta(\mathbf{x}) \rangle \equiv \xi_\zeta(|\mathbf{x}' - \mathbf{x}|) = \text{constant} , \quad (29.128)$$

independent of spatial separation  $|\mathbf{x}' - \mathbf{x}|$ . A scale-free primordial spectrum of fluctuations was originally proposed as a natural initial condition by Harrison (1970) and Zeldovich (1972) before the idea of inflation was conceived. 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  $\zeta$  flying over the horizon remains the same as time goes by.

The power spectrum  $P_\zeta(k)$  of fluctuations in  $\zeta$  is defined by

$$\langle \zeta(\mathbf{k}')\zeta(\mathbf{k}) \rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) P_\zeta(k) . \quad (29.129)$$

The “momentum-conserving” Dirac delta-function  $(2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k})$  in equation (29.129) 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  $2\pi$ )

$$P_\zeta(k) = \int e^{i\mathbf{k}\cdot\mathbf{x}} \xi_\zeta(x) d^3x , \quad \xi_\zeta(x) = \int e^{-i\mathbf{k}\cdot\mathbf{x}} P_\zeta(k) \frac{d^3k}{(2\pi)^3} . \quad (29.130)$$

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

$$\Delta_\zeta^2(k) \equiv P_\zeta(k) \frac{4\pi k^3}{(2\pi)^3} \quad (29.131)$$

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  $n$ , such that

$$\Delta_\zeta^2(k) \propto k^{n-1}. \quad (29.132)$$

Thus a scale-invariant power spectrum has

$$n = 1 \quad (\text{scale-invariant}). \quad (29.133)$$

Different inflationary models predict different tilts, mostly close to but slightly less than 1.

A common practice is to report the value of the dimensionless primordial power spectrum  $\Delta_\zeta^2(k)$  at some pivot scale  $k_p$ ,

$$\Delta_\zeta^2(k) = \Delta_\zeta^2(k_p) \left( \frac{k}{k_p} \right)^{n-1}. \quad (29.134)$$

The Planck collaboration (Ade, 2015) report

$$\Delta_\zeta^2(k_p = 0.05 \text{ Mpc}^{-1}) = (2.14 \pm 0.05) \times 10^{-9}, \quad n = 0.967 \pm 0.004. \quad (29.135)$$

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  $n$ .

## 29.20 Matter power spectrum

The **matter power spectrum**  $P_m(\eta, k)$  at time  $\eta$  is defined by

$$\langle \delta_m(\eta, \mathbf{k}') \delta_m(\eta, \mathbf{k}) \rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) P_m(\eta, k), \quad (29.136)$$

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

$$P_m(\eta, k) = T_m(\eta, k)^2 P_\zeta(k) = T_m(\eta, k)^2 \frac{(2\pi)^3}{4\pi k^3} \Delta_\zeta^2(k), \quad (29.137)$$

where  $T_m(\eta, k)$  is the **matter transfer function** defined by

$$T_m(\eta, k) \equiv \frac{\delta_m(\eta, \mathbf{k})}{\zeta(\mathbf{k})}. \quad (29.138)$$

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 29.15.

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  $\delta_m$  to the potential  $\Phi$ , which in the Newtonian regime at subhorizon scales well after recombination is given in units  $a_0 = 1$  by equation (29.124); (2) a growth factor  $g(a)$ , equation (29.123), relating the potential  $\Phi(\eta)$  at recent times  $\eta$  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  $\zeta$ :

$$\begin{aligned} T_m(\eta, k) &= \frac{\delta_m(\eta, \mathbf{k})}{\Phi(\eta, \mathbf{k})} \times \frac{\Phi(\eta, \mathbf{k})}{\Phi(\text{late}, \mathbf{k})} \times \frac{\Phi(\text{late}, \mathbf{k})}{\zeta(\mathbf{k})} \\ &= -\left(\frac{2ak^2}{3\Omega_m H_0^2}\right) \times g(a) \times T_{\Phi(\text{late})}(k) , \end{aligned} \quad (29.139)$$

where

$$T_{\Phi(\text{late})}(k) \equiv \frac{\Phi(\text{late}, \mathbf{k})}{\zeta(\mathbf{k})} . \quad (29.140)$$

The potential transfer function  $T_{\Phi(\text{late})}(k)$  is independent of time  $\eta$  because the potential  $\Phi(\text{late})$  is constant in the matter-dominated regime before dark energy (or curvature) becomes important. The factorization (29.139) of the matter transfer function  $T_m(\eta, k)$  separates the dependence on time  $\eta$  (or equivalently cosmic scale factor  $a$ ) and wavenumber  $k$ . The first factor  $\delta_m/\Phi$  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 (29.139) of the matter transfer function  $T_m(\eta, k)$  implies that the matter power spectrum  $P_m(\eta, k)$ , equation (29.137), is related to the primordial power spectrum  $P_\zeta(\eta, k)$  by

$$P_m(\eta, k) = \left(\frac{2ag(a)}{3\Omega_m H_0^2}\right)^2 k^4 T_{\Phi(\text{late})}(k)^2 P_\zeta(k) = \left(\frac{2ag(a)}{3\Omega_m H_0^2}\right)^2 k T_{\Phi(\text{late})}(k)^2 \frac{(2\pi)^3}{4\pi} \Delta_\zeta^2(k) . \quad (29.141)$$

For a power-law primordial spectrum (29.132), 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

$$P_m(\eta, k) \propto k^n . \quad (29.142)$$

The proportionality (29.142) explains the origin of the scalar index  $n$ .

**Exercise 29.15 Power spectrum of matter fluctuations: simple approximation.** Use the code you wrote in Exercise 29.10 to compute the transfer function  $T_m(\eta, k)$ , equation (29.138). Deduce the matter power spectrum  $P_m(\eta_0, k)$ , equation (29.137), at the present time,  $\eta = \eta_0$ . Use the normalization and tilt of primordial power measured from Planck, equation (29.135). Compute power spectra for a concordance  $\Lambda$ CDM model,  $\Omega_m = 0.3$ ,  $\Omega_\Lambda = 0.7$ , and a flat matter-dominated Universe,  $\Omega_m = 1$ . Compare your matter power spectrum to data from Anderson (2014), downloadable from [https://www.sdss3.org/science/boss\\_publications.php](https://www.sdss3.org/science/boss_publications.php). The best data set is the “post-reconstruction” DR11 set (Data Release 11). 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

$h \equiv H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$ . Anderson give a covariance matrix for logarithmic powers  $\log_{10} P(k)$ ; for simplicity, take the error in  $\log_{10} P(k)$  to be the square root of the diagonal of that matrix.

As in Exercise 29.10, 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 29.10 was to modify the radiation dipole equation (29.51b) by introducing an artificial damping term, equation (29.56), 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  $\delta_0$  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  $\delta_g$  and matter overdensity  $\delta_m$  (in Fourier space),

$$\delta_g = b\delta_m , \quad (29.143)$$

where  $b$  is the **bias** parameter. Linear bias was introduced by Kaiser (1984), who showed that regions of a Gaussian field (§29.22.3) above a high threshold density are linearly biased.

**Solution.** See Figure 29.15. One of the trickier issues is getting the units right. The BOSS data are given in units where the length scale is such that the comoving Hubble distance at the present time is

$$\frac{c}{a_0 H_0} = \frac{299,792.458 \text{ km s}^{-1}}{100 h \text{ km s}^{-1} \text{ Mpc}} = 2,997.92458 h^{-1} \text{ Mpc}^{-1} . \quad (29.144)$$

My code worked in units where  $c = a_{\text{eq}} = H_{\text{eq}} = 1$ . With  $\Omega_x$  representing values at the present time, the Hubble parameter now  $H_0$  and at matter-radiation equality  $H_{\text{eq}}$  are related by

$$\frac{H_{\text{eq}}}{H_0} = \sqrt{\Omega_r(a_{\text{eq}}/a_0)^{-4} + \Omega_m(a_{\text{eq}}/a_0)^{-3} + \Omega_k(a_{\text{eq}}/a_0)^{-2} + \Omega_\Lambda} . \quad (29.145)$$

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

$$\frac{c}{a_0 H_0} = \frac{1}{3400 \times (6.4 \times 10^{-6})} = 46.0 \text{ program units} . \quad (29.146)$$

The conversion factor between  $h^{-1} \text{ Mpc}$  and program length units was therefore

$$1 \text{ program length unit} = \frac{2,997.92458 h^{-1} \text{ Mpc}}{46.0} = 65.1 h^{-1} \text{ Mpc} . \quad (29.147)$$

For the wavenumber, this meant that the conversion between  $h \text{ Mpc}^{-1}$  and program units was

$$k_{h \text{ Mpc}^{-1}} = \frac{k_{\text{prog}}}{65.1 h^{-1} \text{ Mpc}} . \quad (29.148)$$

The model power spectrum in Figure 29.15 has been multiplied, arbitrarily, by a squared bias factor of

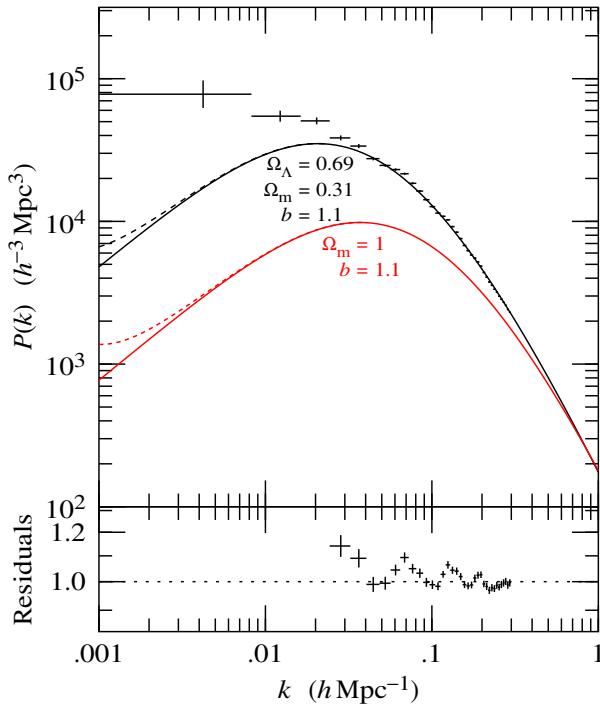


Figure 29.15 Model matter power spectra computed in the simple approximation, compared to observations from the BOSS galaxy survey (Anderson, 2014). Two models are shown, a flat  $\Lambda$ CDM model with concordance parameters  $\Omega_\Lambda = 0.69$  and  $\Omega_m = 0.31$ , and a flat matter-only 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  $\delta_0$  from the overdensity. The  $\Lambda$ CDM model is normalized to the amplitude (29.135) measured by Planck (Ade, 2015), multiplied by a bias squared factor of  $b^2 = 1.1^2$ . The vertical error bars are correlated, being the square root of the diagonal of the full covariance matrix of estimates provided by (Anderson, 2014). The  $\Lambda$ CDM power spectrum calculated here in the simple approximation may be compared to the corresponding power spectra in the hydrodynamic approximation, Figure 31.4, and from a Boltzmann computation, Figure 32.5.

$b^2 = 1.1^2$ , 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 31.4.

## 29.21 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 non-relativistically even in the nonlinear regime, gravity remains well described by the Newtonian limit, equation (29.119c). 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.

## 29.22 Statistics of random fields

### 29.22.1 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,

$$\int P(\rho) d\rho = 1 . \quad (29.149)$$

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 (29.149) 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  $\rho_i$ ,

$$\rho_i \equiv \rho(\mathbf{x}_i) . \quad (29.150)$$

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  $\rho_i$  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,

$$\int P(\rho_1, \dots, \rho_N) d\rho_1 \dots d\rho_N = 1 . \quad (29.151)$$

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.

### 29.22.2 Random fields in Fourier space

Any linear combination  $\tilde{\rho}_i$  of random fields  $\rho_j$  is a random field,

$$\tilde{\rho}_i = \sum_j a_{ij} \rho_j , \quad (29.152)$$

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  $\rho(\mathbf{x})$ ,

$$\rho(\mathbf{k}) \equiv \int \rho(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x , \quad \rho(\mathbf{x}) \equiv \int \rho(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}} \frac{d^3k}{(2\pi)^3} , \quad (29.153)$$

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  $\nabla$ , and the statistical properties of a statistically homogeneous random field commute with the translation operator.

### 29.22.3 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  $\rho$  in §29.22.5.

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  $C_{ij}$ ,

$$C_{ij} \equiv \langle \Delta\rho_i \Delta\rho_j \rangle , \quad (29.154)$$

where  $\Delta\rho_i \equiv \rho_i - \bar{\rho}$  is the deviation of  $\rho_i$  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 (29.154) has dimensions of  $\bar{\rho}^2$ . Commonly, a dimensionless version  $\xi_{ij}$  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 (29.173) derived below,

$$P(\rho_1, \dots, \rho_N) d\rho_1 \dots d\rho_N = \frac{1}{\sqrt{(2\pi)^N |C_{ij}|}} \exp\left(-\frac{1}{2} C_{ij}^{-1} \Delta\rho_i \Delta\rho_j\right) d\rho_1 \dots d\rho_N , \quad (29.155)$$

where  $|C_{ij}|$  is the determinant of the covariance matrix.

Any linear combination  $\sum_j a_{ij} \rho_j$  of Gaussian random fields  $\rho_j$  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.

#### 29.22.4 Moment-generating functions

The proof of the Central Limit Theorem, §29.22.5, goes via moment-generating functions. For simplicity, moment-generating functions are defined in this section for a single random variable  $\rho$ , 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  $\rho$  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

$$M(\mu) \equiv \int e^{\mu\rho} P(\rho) d\rho . \quad (29.156)$$

Expanding the exponential in the integrand as a power series in  $\mu$  implies that the moment-generating function is

$$M(\mu) = 1 + \langle \rho \rangle \mu + \langle \rho^2 \rangle \frac{\mu^2}{2} + \langle \rho^3 \rangle \frac{\mu^3}{3!} + \dots , \quad (29.157)$$

where  $\langle \rho^n \rangle$  is the  $n$ 'th moment of the probability distribution,

$$\langle \rho^n \rangle \equiv \int \rho^n P(\rho) d\rho . \quad (29.158)$$

Equation (29.157) accounts for the name moment-generating function.

Suppose that the measurement of  $\rho$  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  $\rho_i$ )

$$P(\rho_{(1)}, \dots, \rho_{(N)}) = P(\rho_{(1)}) \dots P(\rho_{(N)}) . \quad (29.159)$$

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  $M(\mu)$ ,

$$\begin{aligned} M_N(\mu) &\equiv \int e^{\mu(\rho_{(1)} + \dots + \rho_{(N)})} P(\rho_{(1)}, \dots, \rho_{(N)}) d\rho_{(1)} \dots d\rho_{(N)} \\ &= \int e^{\mu\rho_{(1)}} P(\rho_{(1)}) d\rho_{(1)} \dots \int e^{\mu\rho_{(N)}} P(\rho_{(N)}) d\rho_{(N)} \\ &= M(\mu)^N . \end{aligned} \quad (29.160)$$

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,

$$Z(\mu) \equiv \ln [M(\mu)] . \quad (29.161)$$

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,

$$Z_N(\mu) = N Z(\mu) . \quad (29.162)$$

The coefficients of the series expansion of  $Z(\mu)$  in  $\mu$  define the irreducible moments  $\kappa_n$ ,

$$Z(\mu) = \mu \kappa_1 + \frac{\mu^2}{2} \kappa_2 + \frac{\mu^3}{3!} \kappa_3 + \dots . \quad (29.163)$$

Unlike the moments  $\langle \rho^n \rangle$ , the irreducible moments  $\kappa_n$  have the important property of being additive over sums  $\sum_{i=1}^N \rho_{(i)}$  of independent variables. The defining relation (29.161) between the irreducible  $Z(\mu)$  and standard  $M(\mu)$  moment-generating functions yields the relation between the irreducible moments  $\kappa_n$  and moments  $\langle \rho^n \rangle$ . The relations for the first few moments are, with  $\Delta\rho \equiv \rho - \bar{\rho}$ ,

$$\kappa_1 = \bar{\rho} , \quad (29.164a)$$

$$\kappa_2 = \langle \Delta\rho^2 \rangle , \quad (29.164b)$$

$$\kappa_3 = \langle \Delta\rho^3 \rangle , \quad (29.164c)$$

$$\kappa_4 = \langle \Delta\rho^4 \rangle - 3\langle \Delta\rho^2 \rangle^2 . \quad (29.164d)$$

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  $\kappa_4/\kappa_2^2$ .

More generally, the irreducible-moment-generating function  $Z(\mu_i)$  of a random field  $\rho(\mathbf{x})$  is

$$Z(\mu_i) = \mu_1 \kappa_1 + \frac{\mu_1 \mu_2}{2} \kappa_{12} + \frac{\mu_1 \mu_2 \mu_3}{3!} \kappa_{123} + \dots , \quad (29.165)$$

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

$$\kappa_1 = \bar{\rho} , \quad (29.166a)$$

$$\kappa_{12} = \langle \Delta\rho_1 \Delta\rho_2 \rangle , \quad (29.166b)$$

$$\kappa_{123} = \langle \Delta\rho_1 \Delta\rho_2 \Delta\rho_3 \rangle , \quad (29.166c)$$

$$\kappa_{1234} = \langle \Delta\rho_1 \Delta\rho_2 \Delta\rho_3 \Delta\rho_4 \rangle - \langle \Delta\rho_1 \Delta\rho_2 \rangle \langle \Delta\rho_3 \Delta\rho_4 \rangle - \langle \Delta\rho_1 \Delta\rho_3 \rangle \langle \Delta\rho_2 \Delta\rho_4 \rangle - \langle \Delta\rho_1 \Delta\rho_4 \rangle \langle \Delta\rho_2 \Delta\rho_3 \rangle . \quad (29.166d)$$

### 29.22.5 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  $\rho$ .

As shown in §29.22.4, irreducible moments are additive over sums of independent random variables. Thus the irreducible moment  $\kappa_n$  of a sum  $\sum_{i=1}^N \rho_{(i)}$  of  $N$  independent variables  $\rho_{(i)}$  goes as

$$\kappa_n \propto N . \quad (29.167)$$

The  $n$ 'th irreducible moment  $\kappa_n$  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  $\sigma$ , defined to be the square root of the variance,  $\sigma \equiv \sqrt{\kappa_2} = \sqrt{\langle \Delta\rho^2 \rangle}$ , has the dimension of the  $\bar{\rho}$ . 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  $1/\sqrt{N}$ ,

$$\sigma \equiv \sqrt{\kappa_2} \propto \sqrt{N} , \quad \frac{\sigma}{\bar{\rho}} \equiv \frac{\sqrt{\kappa_2}}{\kappa_1} \propto \frac{1}{\sqrt{N}} . \quad (29.168)$$

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,

$$\frac{\kappa_n^{1/n}}{\kappa_2^{1/2}} \propto N^{1/n-1/2} \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \text{for } n \geq 3 . \quad (29.169)$$

The asymptotic behaviour (29.169) is the CLT: it says that higher order irreducible moments become negligible in the limit of large  $N$ .

## 29.22.6 Gaussian distribution

A Gaussian distribution is defined by the property that its only non-vanishing irreducible moments are the first two, the mean  $\kappa_1$  and variance  $\kappa_2$ . The third and higher irreducible moments of a Gaussian distribution vanish,

$$\kappa_n = 0 \quad (n \geq 3) \quad \text{Gaussian distribution .} \quad (29.170)$$

The irreducible-moment-generating function  $Z(\mu)$  of a Gaussian distribution is, from equation (29.165),

$$Z(\mu) = \mu\bar{\rho} + \frac{\mu^2}{2} \langle \Delta\rho^2 \rangle . \quad (29.171)$$

Accordingly, the moment-generating function  $M(\mu)$  of a Gaussian is

$$M(\mu) = \exp \left( \mu\bar{\rho} + \frac{\mu^2}{2} \langle \Delta\rho^2 \rangle \right) . \quad (29.172)$$

The probability distribution  $P(\rho)$  that, when integrated in accordance with the definition (29.156), yields the Gaussian moment-generating function (29.172), is

$$P(\rho) d\rho = \frac{1}{\sqrt{2\pi\langle\Delta\rho^2\rangle}} \exp\left[-\frac{(\rho - \bar{\rho})^2}{2\langle\Delta\rho^2\rangle}\right] d\rho . \quad (29.173)$$

The 1-point Gaussian probability distribution (29.173) generalizes to the  $N$ -point Gaussian probability distribution (29.155).

## 30

---

### Non-equilibrium processes in the FLRW background

The subject of cosmological perturbations will be resumed in the next Chapter 31. 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, §30.5, which allows to follow the evolution of number densities of species out of thermodynamic equilibrium, and which will be invoked again in Chapter 32 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 ( $\equiv {}^2\text{H}$ ),  ${}^4\text{He}$ ,  ${}^3\text{He}$ , 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 non-relativistic; 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.

### 30.1 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 (10.101), 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

$$T = \frac{T_0}{a} , \quad (30.1)$$

where  $T_0 = 2.725$  K is the CMB temperature today. Equation (30.1) 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  $a_0 = 1$ )

$$n_b = \frac{3\Omega_b H_0^2}{8\pi G m_b a^3} , \quad (30.2)$$

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,

$$X_e \equiv \frac{n_e}{n_+} . \quad (30.3)$$

The definition (30.3) is chosen so that  $X_e = 1$  when the plasma is fully ionized. The nuclear proton density  $n_+$  is

$$n_+ = f_+ n_b , \quad (30.4)$$

where  $f_+ \equiv n_+/n_b$  is the proton fraction. To a good approximation, baryons comprised H and  ${}^4\text{He}$  nuclei, and  $f_+ = 0.875$ , Exercise 30.1.

**Exercise 30.1 Proton and neutron fractions.** Define the proton and neutron fractions  $f_+$  and  $f_n$  by the proton- and neutron-to-baryon ratios

$$f_+ \equiv \frac{n_+}{n_b} = 1 - f_n , \quad f_n \equiv \frac{n_n}{n_b} . \quad (30.5)$$

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

$$n_+ = n_{\text{H}} + 2n_{{}^4\text{He}}, \quad n_n = 2n_{{}^4\text{He}}. \quad (30.6)$$

Show that the primordial  ${}^4\text{He}$  mass fraction defined by  $Y_{\text{He}} \equiv \rho_{{}^4\text{He}} / (\rho_{\text{H}} + \rho_{{}^4\text{He}})$  satisfies

$$Y_{\text{He}} = 2f_n. \quad (30.7)$$

The observed primordial  ${}^4\text{He}$  abundance is  $Y_{\text{He}} = 0.25$  (see Ade, 2015), implying

$$f_n = 0.125. \quad (30.8)$$


---

## 30.2 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 non-relativistic, and their occupation numbers  $f$ , given in thermodynamic equilibrium by equations (10.121), 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 (10.101). 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 (10.137)). 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 10.17. 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  $\sim 10^{-9}$ . As the Universe continued to expand, the occupation number of the now non-relativistic particles, still in thermodynamic equilibrium with photons, fell further as  $f \sim nT^{-3/2} \propto T^{3/2}$ , equation (30.15). Thus the

occupation numbers of non-relativistic electrons and nuclei was

$$f \sim 10^{-9} \left( \frac{T}{m} \right)^{3/2} \ll 1 \quad (30.10)$$

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  $13.6 \text{ eV}$  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 (30.21). 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 (30.21). 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  $\alpha$   $2p-1s$  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.

### 30.3 Energy levels and ionization state in thermodynamic equilibrium

Electrons and nuclei near recombination were non-relativistic, and their occupation numbers were small, and therefore well described by Boltzmann statistics, with occupation number  $f$  given by equation (10.122).

#### 30.3.1 Number density of non-relativistic Boltzmann species in thermodynamic equilibrium

The energy  $E$  of a non-relativistic 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

non-relativistic Boltzmann species is

$$n = \int f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = e^{(\mu-m)/T} \int e^{-p^2/(2mT)} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3}. \quad (30.11)$$

The integral on the right hand side of equation (30.11) is

$$\int e^{-p^2/(2mT)} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = g \left( \frac{mT}{2\pi\hbar^2} \right)^{3/2}, \quad (30.12)$$

so the number density in thermodynamic equilibrium is

$$n = g \left( \frac{mT}{2\pi\hbar^2} \right)^{3/2} e^{(\mu-m)/T}. \quad (30.13)$$

The factor  $(mT/(2\pi\hbar^2))^{3/2}$  defines a length scale  $\lambda_T$  which is a characteristic thermal compton wavelength of the particles,

$$\lambda_T \equiv \left( \frac{mT}{2\pi\hbar^2} \right)^{-1/2}. \quad (30.14)$$

In terms of their number density  $n$ , the occupation number  $f = e^{(\mu-E)/T}$  of a Boltzmann species is

$$f = \frac{n}{g} \left( \frac{mT}{2\pi\hbar^2} \right)^{-3/2} e^{-p^2/(2mT)} = \frac{n\lambda_T^3}{g} e^{-p^2/(2mT)}. \quad (30.15)$$

The condition for the validity of the Boltzmann approximation of small occupation numbers is that there be few particles per compton volume,  $n\lambda_T^3 \ll 1$ .

### 30.3.2 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

$$-E_n = -13.6 \text{ eV}/n^2, \quad (30.16)$$

with  $n = 1, \dots, \infty$  an integer running from the ground state 1 to the continuum  $\infty$ . 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 (30.13),

$$\frac{n_{nl}}{n_{1s}} = (2l+1)e^{(E_n - E_1)/T}. \quad (30.17)$$

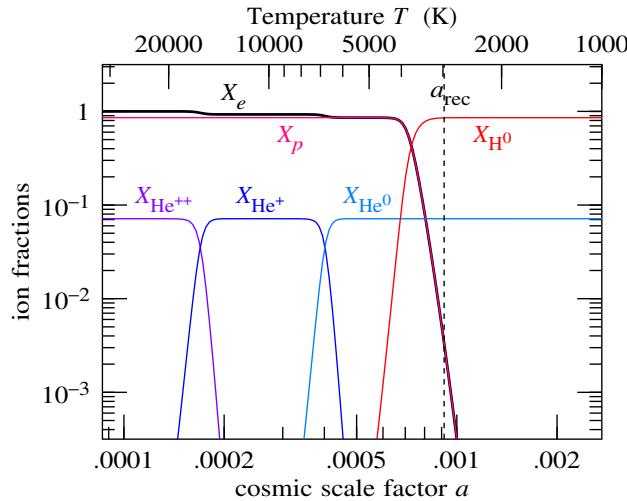


Figure 30.1 Hydrogen and helium ion fractions in thermodynamic equilibrium as a function of cosmic scale factor  $a$  scaled to  $a_0 = 1$ . The total hydrogen and helium fractions are  $X_H = 1 - f_n/f_+ = 0.86$  and  $X_{^4\text{He}} = \frac{1}{2}f_n/f_+ = 0.07$  where  $f_p \equiv 1 - f_n = 0.875$  and  $f_n \equiv n_n/n_b = 0.125$  are the neutron- and proton-to-baryon ratios, Exercise 30.1. The dashed vertical line indicates where recombination actually occurs (where the Thomson scattering optical depth is unity), somewhat later than predicted by equilibrium.

### 30.3.3 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 (10.124). Inserting this equilibrium condition into equation (30.13), valid for non-relativistic Boltzmann species, implies the relation between the number densities  $n_p$ ,  $n_e$ , and  $n_{nl}$  of protons, electrons, and hydrogen atoms in level  $nl$ ,

$$\frac{n_p n_e}{n_{nl}} = \frac{g_p g_e}{g_{nl}} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_n/T}. \quad (30.18)$$

Equation (30.18) is the **Saha equation** for hydrogen. The  $m_e$  on the right hand side of equation (30.18) 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

$$\frac{n_{X^+} n_e}{n_X} = \frac{g_{X^+} g_e}{n_X} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_X/T}. \quad (30.19)$$

Figure 30.1 illustrates the ionization fractions of H and  ${}^4\text{He}$  in thermodynamic equilibrium at the photon temperature  $T$  and baryon density  $n_b$  given by equations (30.1) and (30.2).

**Exercise 30.2 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 (30.17), the ratio of excited  $n = 2$  to ground  $n = 1$  levels in thermodynamic equilibrium is

$$n_{2p} = 3n_{2s} \sim 3e^{\left(\frac{1}{4}-1\right)13.6 \text{ eV}/0.3 \text{ eV}} n_{1s} \sim 10^{-14} n_{1s}, \quad (30.20)$$

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

$$n_3 \sim \frac{9}{4} e^{\left(\frac{1}{9}-\frac{1}{4}\right)13.6 \text{ eV}/0.3 \text{ eV}} n_2 \sim 4 \times 10^{-3} n_2, \quad (30.21)$$

which is also small. Thus the ground state  $n = 1$  dominates the level population, followed by the first excited states  $n = 2$ ,

$$n_1 \gg n_2 \gg n_{n \geq 3}. \quad (30.22)$$

**Exercise 30.3 Ionization state of hydrogen near recombination.** Use the approximation of thermodynamic equilibrium to estimate that 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 (30.18) implies

$$\frac{n_p n_e}{n_{1s}} = \left( \frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_1/T}, \quad (30.23)$$

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 (30.23) is

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_+} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_1/T} = \frac{2^{3/2} G m_b T_0^3}{3\pi^{1/2} \Omega_b H_0^2} \left( \frac{m_e}{T} \right)^{3/2} e^{-E_1/T}, \quad (30.24)$$

where  $X_e$  is the electron fraction, equation (30.3), and  $n_b$  the baryon density, equation (30.2). Recombination occurs at  $X_e \approx \frac{1}{2}$ . Equation (30.24) 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

$$\frac{E_1}{T} = \ln \left[ \frac{2^{3/2} G m_b T_0^3}{3\pi^{1/2} \Omega_b H_0^2} \left( \frac{m_e}{T} \right)^{3/2} \right]. \quad (30.25)$$

Guessing  $T = 10^4 \text{ K}$  yields

$$\frac{E_1}{T} \approx 40, \quad (30.26)$$

which gives the estimated recombination temperature of  $T \approx 4,000 \text{ K}$ . Iterating a second time gives

$$T \approx 3800 \text{ K}. \quad (30.27)$$

**Concept question 30.4 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.

---

## 30.4 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,

$$dN(t, \mathbf{x}, \mathbf{p}) = f(t, \mathbf{x}, \mathbf{p}) \frac{g d^3r d^3p}{(2\pi\hbar)^3} , \quad (30.28)$$

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

$$p^m \equiv e^m \frac{dx^\mu}{d\lambda} = \{E, \mathbf{p}\} = \{E, p^a\} , \quad (30.29)$$

where  $\lambda$  is the affine parameter, related to proper time  $\tau$  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

$$E^2 - p^2 = m^2 . \quad (30.30)$$

## 30.5 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 (30.28),

$$\frac{df}{d\lambda} = 0 . \quad (30.31)$$

Equation (30.31) is the **collisionless Boltzmann equation**. The derivative with respect to affine parameter

$\lambda$  on the left hand side of the Boltzmann equation (30.31) 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,

$$\frac{df}{d\lambda} = C[f] . \quad (30.32)$$

Equation (30.32) 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.

### 30.5.1 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 (30.32) is then

$$\frac{df}{d\lambda} = \frac{dt}{d\lambda} \frac{\partial f}{\partial t} + \frac{dp}{d\lambda} \frac{\partial f}{\partial p} = C[f] . \quad (30.33)$$

To follow lots of particles simultaneously, switch the integration variable from the affine parameter  $\lambda$ , 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  $e_0^{\textcolor{brown}{t}} = 1$ . The relation between conformal time  $t$  and affine parameter  $\lambda$  is

$$\frac{dt}{d\lambda} = p^{\textcolor{brown}{t}} = e_0^{\textcolor{brown}{t}} p^0 = E , \quad (30.34)$$

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  $\eta$  as the integration variable, as will be done later in §32.2, in which case  $e_0^{\textcolor{brown}{\eta}} = 1/a$  and  $d\eta/d\lambda = E/a$ ; for the present purpose however, cosmic time  $t$  is slightly more convenient. As found in Exercise 10.5, 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 (30.33) is

$$\frac{df}{dt} = \frac{\partial f}{\partial t} - \frac{d \ln a}{dt} \frac{\partial f}{\partial \ln p} = \frac{1}{E} C[f] . \quad (30.35)$$

The proper number density  $n$  is an integral (10.117) of the occupation number  $f$  over momenta. Integrating the left hand side of the Boltzmann equation (30.35) gives

$$\begin{aligned} \int \frac{df}{dt} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} &= \int \frac{\partial f}{\partial t} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} - \frac{d \ln a}{dt} \int \frac{\partial f}{\partial \ln p} \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} \\ &= \frac{\partial}{\partial t} \int f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} - \frac{d \ln a}{dt} \left\{ \left[ f \frac{g 4\pi p^3}{(2\pi\hbar)^3} \right] - \int 3f \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} \right\} \\ &= \frac{dn}{dt} + 3 \frac{d \ln a}{dt} n = \frac{1}{a^3} \frac{d n a^3}{dt}. \end{aligned} \quad (30.36)$$

Integrated over momenta, the collisional Boltzmann equation (30.35) is thus

$$\frac{1}{a^3} \frac{d n a^3}{dt} = \int C[f] \frac{g 4\pi p^2 dp}{E(2\pi\hbar)^3}. \quad (30.37)$$

Equation (30.37) holds for both massive and massless particles. In the absence of collisions,  $C[f] = 0$ , the integrated Boltzmann equation (30.37) shows that proper number density  $n$  decreases as  $a^{-3}$ ,

$$n \propto a^{-3}. \quad (30.38)$$

Equation (30.38) says that the number  $n a^3$  of particles in a comoving volume remains constant in the absence of collisions that destroy or create particles.

## 30.6 Collisions

For a 2-body collision of the form

$$1 + 2 \leftrightarrow 3 + 4, \quad (30.39)$$

the rate per unit time and volume at which particles of type 1 leave and enter an interval  $d^3 p_1$  of momentum space is, in units  $c = \hbar = 1$ ,

$$\begin{aligned} C[f_1] \frac{g_1 d^3 p_1}{E_1 (2\pi)^3} &= \int |\mathcal{M}|^2 [-f_1 f_2 (1 \mp f_3)(1 \mp f_4) + f_3 f_4 (1 \mp f_1)(1 \mp f_2)] \\ &\quad (2\pi)^4 \delta_D^4(p_1 + p_2 - p_3 - p_4) \frac{g_1 d^3 p_1}{2E_1 (2\pi)^3} \frac{g_2 d^3 p_2}{2E_2 (2\pi)^3} \frac{d^3 p_3}{2E_3 (2\pi)^3} \frac{d^3 p_4}{2E_4 (2\pi)^3}. \end{aligned} \quad (30.40)$$

All factors in equation (30.40) are Lorentz scalars. On the left hand side, the collision term  $C[f_1]$  and the momentum 3-volume element  $d^3 p_1 / E_1$  are both Lorentz scalars. On the right hand side, the squared amplitude  $|\mathcal{M}|^2$ , the various occupation numbers  $f_a$ , 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  $d^3 p_i / (2E_i)$ , are all Lorentz scalars. The factor of  $1/2$  in each momentum element  $d^3 p_i / (2E_i)$  on the right hand side is associated with the way that the invariant amplitude  $\mathcal{M}$  is conventionally defined in quantum field theory, and arises from integrating  $dE_i d^3 p_i$  over a Dirac delta-function that enforces the “on-shell” relation between energy, momentum and rest mass, equation (10.113), for the incoming and outgoing particles.

The first ingredient in the integrand on the right hand side of the expression (30.40) is the Lorentz-invariant scattering amplitude squared  $|\mathcal{M}|^2$ , calculated using quantum field theory (e.g. Peskin and Schroeder, 1995). For a process involving 4 particles such as (30.39), the invariant amplitude squared  $|\mathcal{M}|^2$  is dimensionless (in units  $c = \hbar = 1$ ). See Concept question 30.5 for a dimensional analysis of equation (30.40).

The invariant amplitude squared  $|\mathcal{M}|^2$  represents a rate averaged over initial spin states and summed over final spin states. To convert the invariant amplitude squared to a rate per unit time and volume, it is necessary to sum over particles in the initial states. This explains why equation (30.40) includes spin factors  $g_1$  and  $g_2$  for the initial states but not for the final states.

The second ingredient in the integrand on the right hand side of expression (30.40) is the combination of rate factors

$$\text{rate}(1 + 2 \rightarrow 3 + 4) \propto f_1 f_2 (1 \mp f_3)(1 \mp f_4) , \quad (30.41a)$$

$$\text{rate}(1 + 2 \leftarrow 3 + 4) \propto f_3 f_4 (1 \mp f_1)(1 \mp f_2) , \quad (30.41b)$$

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:

$$1 - f = \text{Fermi-Dirac blocking factor} , \quad (30.42a)$$

$$1 + f = \text{Bose-Einstein stimulation factor} . \quad (30.42b)$$

The first rate factor (30.41a) 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 (30.41b) 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 (30.41) balance, Exercise 30.6, 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 (30.40) 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:

$$(2\pi)^4 \delta_D^4(p_1 + p_2 - p_3 - p_4) = 2\pi \delta_D(E_1 + E_2 - E_3 - E_4) (2\pi)^3 \delta_D^3(\mathbf{p}_1 + \mathbf{p}_2 - \mathbf{p}_3 - \mathbf{p}_4) . \quad (30.43)$$

**Concept question 30.5 Dimensional analysis of the collision rate.** Carry out a dimensional analysis of equation (30.40). Restore factors of  $c$  and  $\hbar$ . **Answer.** All potential factors of  $c$  are accounted for by interpreting factors of energy  $E$  on both left hand and right hand sides of equation (30.40) as being in units of momentum (that is, replace  $E$  by  $p^0 = E/c$ ). After that replacement is done, then the dimensions of both left and right hand sides are  $1/\text{length}^4$  in units  $\hbar = 1$ . On the left hand side, the units of  $C[f] = df/d\lambda$  are  $1/\lambda$ , which is mass/time, or equivalently momentum/length, or equivalently  $\hbar/\text{length}^2$ . The phase space factor on the left hand side has units momentum<sup>2</sup> (with  $E$  replaced by  $p^0$ ) or equivalently  $\hbar^2/\text{length}^2$ . Thus

the units of the left hand side are  $\hbar^3/\text{length}^4$ . On the right hand side, the units of the delta function are  $1/\text{momentum}^4$ , while the units of each phase space factor are momentum<sup>2</sup> (again with each  $E$  interpreted as  $p^0$ ). The units of the combined delta function and phase space factors are momentum<sup>4</sup>, or equivalently  $\hbar^4/\text{length}^4$ . The length dimensions of the left and right hand sides match (both  $1/\text{length}^4$ ) provided that  $|\mathcal{M}|^2$  is dimensionless in units  $c = \hbar = 1$ . If  $|\mathcal{M}|^2$  is treated as dimensionless even in dimensionful units, then the left times  $\hbar^3$  and the right times  $\hbar^4$  have the same dimensionful unit  $1/\text{length}^4$ . To make equation (30.40) yield a rate in units of particles per unit time per unit volume, both sides should be multiplied further by  $c$ .

### Exercise 30.6 Detailed balance.

1. Show that the rates balance in thermodynamic equilibrium,

$$f_1 f_2 (1 \mp f_3) (1 \mp f_4) = f_3 f_4 (1 \mp f_1) (1 \mp f_2) . \quad (30.44)$$

2. Conclude that, if each particle type  $i$  has a thermodynamic distribution with its own temperature  $T_i$  and chemical potential  $\mu_i$ , then

$$\begin{aligned} & - f_1 f_2 (1 \mp f_3) (1 \mp f_4) + f_3 f_4 (1 \mp f_1) (1 \mp f_2) \\ &= f_1 f_2 (1 \mp f_3) (1 \mp f_4) \left[ -1 + \exp \left( \frac{E_1 - \mu_1}{T_1} + \frac{E_2 - \mu_2}{T_2} + \frac{-E_3 + \mu_3}{T_3} + \frac{-E_4 + \mu_4}{T_4} \right) \right] . \end{aligned} \quad (30.45)$$

#### Solution.

1. Equation (30.44) is true if and only if

$$\frac{f_1}{1 \mp f_1} \frac{f_2}{1 \mp f_2} = \frac{f_3}{1 \mp f_3} \frac{f_4}{1 \mp f_4} . \quad (30.46)$$

But

$$\frac{f}{1 \mp f} = e^{(-E+\mu)/T} , \quad (30.47)$$

so (30.46) is true if and only if

$$\frac{-E_1 + \mu_1}{T} + \frac{-E_2 + \mu_2}{T} = \frac{-E_3 + \mu_3}{T} + \frac{-E_4 + \mu_4}{T} , \quad (30.48)$$

which is true in thermodynamic equilibrium because

$$E_1 + E_2 = E_3 + E_4 , \quad \mu_1 + \mu_2 = \mu_3 + \mu_4 . \quad (30.49)$$


---

## 30.7 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 (30.40) with stimulated processes from protons, electrons, and hydrogen atoms neglected because of their small occupation numbers,

$$C_i[f_e] \frac{g_p d^3 p_p}{m_p (2\pi)^3} = \int |\mathcal{M}|_i^2 [-f_p f_e (1 + f_\gamma) + f_i f_\gamma] (2\pi)^4 \delta_D^4(p_p + p_e - p_i - p_\gamma) \frac{g_p d^3 p_p}{2m_p (2\pi)^3} \frac{g_e d^3 p_e}{2m_e (2\pi)^3} \frac{d^3 p_i}{2m_p (2\pi)^3} \frac{d^3 p_\gamma}{2p_\gamma (2\pi)^3}. \quad (30.50)$$

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 invariant squared amplitude  $|\mathcal{M}|_i^2$  is essentially independent of the proton and hydrogen momenta. Integrating the collision integral (30.50) over the proton and hydrogen momenta yields

$$C_i[f_e] = \frac{1}{8m_p^2} \int \mathcal{M}|_i^2 2\pi \delta_D(E_p + E_e - E_i - E_\gamma) [-n_p f_e (1 + f_\gamma) + n_i (g_p/g_i) f_\gamma] \frac{d^3 p_\gamma}{2p_\gamma (2\pi)^3}, \quad (30.51)$$

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  $4\pi$ . Integrating over the photon energy  $p_\gamma$  swallows the energy-conserving delta-function, yielding

$$C_i[f_e] = \frac{p_\gamma}{16\pi m_p^2} |\mathcal{M}|_i^2 [-n_p f_e (1 + f_\gamma) + n_i (g_p/g_i) f_\gamma]. \quad (30.52)$$

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,

$$\frac{p_e^2}{2m_e} + E_n = p_\gamma. \quad (30.53)$$

In the situation of cosmological recombination under consideration, the photons, whose numbers overwhelm those of electrons, have a thermal (Planckian) momentum distribution at temperature  $T_\gamma$ . 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 (30.15) at temperature  $T_e$  for the electron occupation number  $f_e$ , and the Planckian distribution (10.126) at temperature  $T_\gamma$  for the photon occupation number  $f_\gamma$ , brings the electron collision integral (30.52) to

$$C_i[f_e] = \frac{p_\gamma}{16\pi m_p^2} |\mathcal{M}|_i^2 \left[ \frac{-n_p n_e (1/g_e) (m_e T_e / 2\pi)^{-3/2} e^{-p_e^2 / (2m_e T_e)} + n_i (g_p / g_i) e^{-p_\gamma / T_\gamma}}{1 - e^{-p_\gamma / T_\gamma}} \right]. \quad (30.54)$$

Finally, integrating the collision integral (30.54) over electron momenta gives

$$\int C_i[f_e] \frac{g_e d^3 p_e}{m_p (2\pi)^3} = -n_p n_e [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + n_i \beta_i(T_\gamma), \quad (30.55)$$

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 direction 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  $T_\gamma$ . 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

$$\beta_i(T) = \alpha_i(T) \left( \frac{n_p n_e}{n_i} \right)_{\text{TE}} = \alpha_i(T) \frac{g_p g_e}{g_i} \left( \frac{m_e T}{2\pi \hbar^2} \right)^{3/2} e^{-E_n/T}. \quad (30.56)$$

The Boltzmann equation for electrons, equation (30.37), is a sum over recombinations to and photoionizations out of bound states  $i$ ,

$$\frac{1}{a^3} \frac{dn_e a^3}{dt} = -n_p n_e \sum_i [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + \sum_i n_i \beta_i(T_\gamma). \quad (30.57)$$

Let  $X_i$  denote the ratio of the number density of species  $i$  to the nuclear proton density  $n_+$ ,

$$X_i \equiv \frac{n_i}{n_+}. \quad (30.58)$$

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 (30.57) can be written as an equation for the evolution of the electron fraction,

$$\frac{dX_e}{dt} = -X_p X_e n_+ \sum_i [\alpha_i(T_e) + \alpha_i^{\text{stim}}(T_e, T_\gamma)] + \sum_i X_i \beta_i(T_\gamma). \quad (30.59)$$

Equation (30.59) 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.

### 30.8 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 (30.22), 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 (30.22), 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 photonionizations out of the  $n = 2$  level,

$$\frac{dX_p}{dt} = -X_p X_e n_{+} \alpha_2 + X_2 \beta_2 . \quad (30.60)$$

Equation (30.60) 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  $\alpha$  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  $\alpha$  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, §30.10. Thus the rate of change of the fraction  $X_1 \equiv n_1/n_+$  of hydrogen atoms in the ground  $n = 1$  level is

$$\frac{dX_1}{dt} = X_2 A_{21} - X_1 B_{12} , \quad (30.61)$$

where the effective spontaneous decay rate  $A_{21}$  from the  $n = 2$  levels to the ground  $n = 1$  level is

$$A_{21} = \frac{g_{2s}}{g_2} A_{2s-1s} + \frac{g_{2p}}{g_2} A_{2p-1s} P_S , \quad (30.62)$$

with  $P_S$  the Sobolev escape probability given by equation (30.100). Equation (30.62) 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 (30.101),

$$\frac{g_{2p}}{g_2} A_{2p-1s} P_S = \frac{1}{X_1 n_+} \frac{g_1}{g_2} \frac{8\pi H}{\lambda_{2p-1s}^3} , \quad (30.63)$$

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 (30.60), must vanish in thermodynamic equilibrium (TE), so the ratio of photonionization to recombination rate coefficients must be

$$\frac{\beta_2}{\alpha_2} = \left( \frac{n_p n_e}{n_2} \right)_{\text{TE}} = \frac{g_p g_e}{g_2} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{3/2} e^{-E_2/T}, \quad (30.64)$$

the statistical weight factor being  $g_p g_e / g_2 = \frac{1}{4}$ . Hence equation (30.60) may be written

$$\frac{dX_p}{dt} = X_p X_e n_+ \alpha_2 \left( -1 + \frac{1}{b_{p2}} \right), \quad (30.65)$$

where the departure coefficient  $b_{p2}$  is the value of  $n_p n_e / n_2$  relative to its value in thermodynamic equilibrium,

$$b_{p2} \equiv \frac{n_p n_e}{n_2} / \left( \frac{n_p n_e}{n_2} \right)_{\text{TE}} = \frac{X_p X_e n_+}{X_2} \frac{g_2}{g_p g_e} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{-3/2} e^{E_2/T}. \quad (30.66)$$

Similarly, detailed balance requires that  $dX_1/dt$ , equation (30.61), must vanish in thermodynamic equilibrium, so the ratio of radiative excitation to decay rate coefficients must be

$$\frac{B_{12}}{A_{21}} = \left( \frac{X_2}{X_1} \right)_{\text{TE}} = \frac{g_2}{g_1} e^{-E_{12}/T}, \quad (30.67)$$

with  $E_{12} \equiv E_1 - E_2$  and  $g_2/g_1 = 4$ . Thus equation (30.61) may be written

$$\frac{dX_1}{dt} = X_1 B_{12} (b_{21} - 1), \quad (30.68)$$

where the departure coefficient  $b_{21}$  is the value of  $n_2/n_1$  relative to its value in thermodynamic equilibrium,

$$b_{21} \equiv \frac{n_2}{n_1} / \left( \frac{n_2}{n_1} \right)_{\text{TE}} = \frac{X_2}{X_1} \frac{g_1}{g_2} e^{E_{12}/T}. \quad (30.69)$$

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  $X_1$ ,

$$\frac{dX_2}{dt} = -\frac{dX_p}{dt} - \frac{dX_1}{dt} = X_p X_e n_+ \alpha_2 - X_2 \beta_2 - X_2 A_{21} + X_1 B_{12} \approx 0. \quad (30.70)$$

The approximation (30.70) of vanishing  $dX_2/dt$  allows  $X_2$  to be eliminated in favour of  $X_1$ ,

$$\frac{X_2}{X_1} = \frac{(X_p X_e / X_1) n_+ \alpha_2 + B_{12}}{\beta_2 + A_{21}}. \quad (30.71)$$

Given the detailed balance relations (30.64) between  $\beta_2$  and  $\alpha_2$ , and (30.67) between  $B_{12}$  and  $A_{21}$ , the relation (30.71) may also be written as an expression for the departure coefficient  $b_{21}$ ,

$$b_{21} = \frac{b_{p2} \beta_2 + A_{21}}{\beta_2 + A_{21}}, \quad (30.72)$$

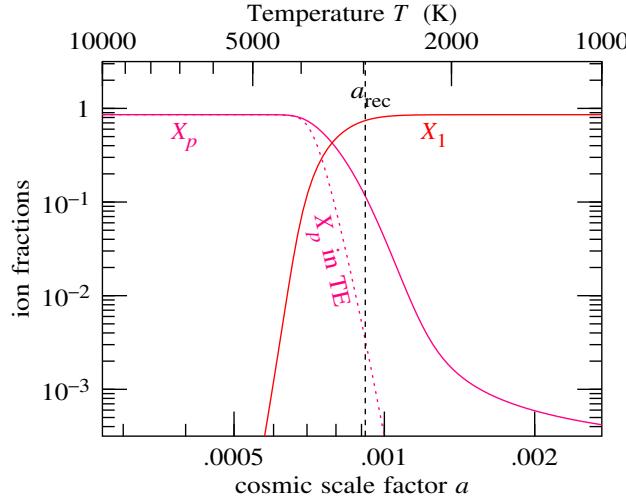


Figure 30.2 Non-equilibrium hydrogen ion fractions as a function of cosmic scale factor  $a$  scaled to  $a_0 = 1$ . The total hydrogen fraction,  $X_H = f_H = 0.86$ , is the fraction of nuclear protons that are hydrogen nuclei, equation (30.82).

in terms of the departure coefficient  $b_{p1}$ , the value of  $n_p n_e / n_1$  relative to its value in thermodynamic equilibrium,

$$b_{p1} \equiv \frac{n_p n_e}{n_1} \left/ \left( \frac{n_p n_e}{n_1} \right)_{\text{TE}} \right. = \frac{X_p X_e n_+}{X_1} \frac{g_1}{g_p g_e} \left( \frac{m_e T}{2\pi\hbar^2} \right)^{-3/2} e^{E_1/T}. \quad (30.73)$$

The statistical weight factor is  $g_1/(g_p g_e) = 1$ . Equation (30.72) allows the departure coefficients  $b_{21}$  and  $b_{p2} \equiv b_{p1}/b_{21}$  in the recombination equations (30.65) and (30.68) to be eliminated in favour of  $b_{p1}$ , yielding

$$\frac{d \ln X_p}{dt} = X_e n_+ \alpha_2 \frac{A_{21}}{\beta_2 + A_{21}} \left( \frac{1}{b_{p1}} - 1 \right), \quad (30.74a)$$

$$\frac{d \ln X_1}{dt} = B_{12} \frac{\beta_2}{\beta_2 + A_{21}} (b_{p1} - 1). \quad (30.74b)$$

Equations (30.74a) and (30.74b) combine to give  $d(X_p + X_1)/dt = 0$  in accordance with the condition (30.70), but each of equations (30.74) is written in a form that remains finite as respectively  $X_p \rightarrow 0$  and  $X_1 \rightarrow 0$ .

Equations (30.74) combine to give the rate of change of the logarithmic departure coefficient  $\ln b_{p1}$ ,

$$\frac{d \ln b_{p1}}{dt} = \frac{d \ln X_p}{dt} + \frac{d \ln X_e}{dt} - \frac{d \ln X_1}{dt} + \frac{d}{dt} \ln \left( n_+ T^{-3/2} e^{E_1/T} \right). \quad (30.75)$$

Given that helium is largely neutral by the time of recombination, charge conservation in the pure hydrogen gas implies that

$$X_e = X_p, \quad (30.76)$$

so the time derivative of  $\ln X_e$  in equation (30.75) is the same as that for  $\ln X_p$ . The time derivatives of the

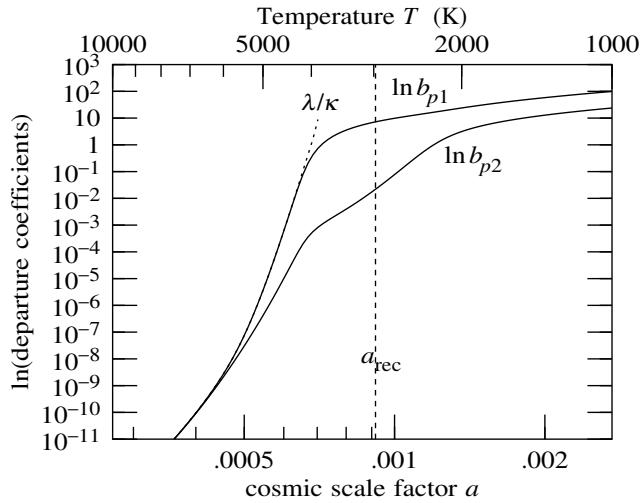


Figure 30.3 Logarithmic ionized-to-bound level departure coefficients  $\ln b_{p1}$  and  $\ln b_{p2}$ , equations (30.73) and (30.66). Logarithmic departure coefficients are zero in thermodynamic equilibrium (the departure coefficients themselves are unity). The increasingly positive values of  $\ln b_{p1}$  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  $\lambda/\kappa$ , equation (30.80), indicated by the dotted line.

temperature and density follow from  $T \propto a^{-1}$  and  $n_+ \propto a^{-3}$ , equations (30.1) and (30.4). The differential equation (30.75) is stiff. Near thermodynamic equilibrium, the logarithmic departure coefficient  $\ln b_{p1}$  is near zero, and then the factors involving  $b_{p1}$  on the right hand sides of equations (30.74) become small,

$$\frac{1}{b_{p1}} - 1 = e^{-\ln b_{p1}} - 1 \approx -\ln b_{p1} , \quad b_{p1} - 1 = e^{\ln b_{p1}} - 1 \approx \ln b_{p1} . \quad (30.77)$$

The  $d \ln X_i / dt$  derivatives in equation (30.75) are then proportional to  $\ln b_{p1}$  with a negative coefficient  $-\kappa$ ,

$$\frac{d \ln X_p}{dt} + \frac{d \ln X_e}{dt} - \frac{d \ln X_1}{dt} \approx -\kappa \ln b_{p1} . \quad (30.78)$$

Thus the differential equation (30.75) takes the form

$$\frac{d \ln b_{p1}}{dt} \approx -\kappa \ln b_{p1} + \lambda , \quad (30.79)$$

where the forcing term  $\lambda$  is the remaining, last, term on the right hand side of the differential equation (30.75). The  $\kappa$  term tends to drive  $\ln b_{p1}$  exponentially to zero, that is, into thermodynamic equilibrium, while the forcing term  $\lambda$  drives  $\ln b_{p1}$  away from zero. The differential equation (30.79) is stiff when  $\kappa$  is much larger than the absolute value of  $\lambda$ .

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 (30.79), which is

$$\ln b_{p1} \approx \frac{\lambda}{\kappa}. \quad (30.80)$$

A differential equation solver that can cope with stiff equations will in effect impose the solution (30.80). 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

$$X_1 = \frac{4f_{\text{H}}^2 e^q}{(1 + \sqrt{1 + 4f_{\text{H}}e^q})^2}, \quad X_p = \frac{2f_{\text{H}}}{1 + \sqrt{1 + 4f_{\text{H}}e^q}}, \quad (30.81)$$

where  $f_{\text{H}}$ ,

$$f_{\text{H}} \equiv n_{\text{H}}/n_+ = 1 - f_n/f_+, \quad (30.82)$$

is the fraction of nuclear protons that are hydrogen nuclei, and  $q$  is

$$q \equiv \ln \left( \frac{X_1}{X_p X_e} \right) = \ln \left[ n_+ \frac{g_1}{g_p g_e} \left( \frac{m_e T}{2\pi \hbar^2} \right)^{-3/2} \right] + \frac{\chi_{\text{H}}}{T} - \ln b_{p1}. \quad (30.83)$$

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 (30.74a),

$$\frac{dX_p}{dt} = \left( X_1 \frac{g_2}{g_1} \beta_2 e^{(E_2 - E_1)/T} - X_e X_p n_+ \alpha_2 \right) \frac{A_{21}}{\beta_2 + A_{21}}. \quad (30.84)$$

The statistical weight factor is  $g_2/g_1 = 4$ . Figure 30.2 shows the resulting non-equilibrium H ion fractions, and Figure 30.3 shows the logarithmic departure coefficients  $\ln b_{p1}$  and  $\ln b_{p2}$ . Exercise 30.7 asks you to write code to solve equation (30.84).

**Exercise 30.7 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  $\kappa$  and  $\lambda$  that go into equation (30.79) in thermodynamic equilibrium. If  $\kappa/|\lambda|$  exceeds some threshold, then set the logarithmic departure coefficient to  $\ln b_{p1} = \lambda/\kappa$ , equation (30.80). Thence deduce the ionization fractions  $X_1$  and  $X_p$ , equation (30.81).
2. Once  $\kappa/|\lambda|$  falls below the threshold, solve the evolution equation (30.84) for the proton fraction  $X_p$  numerically.

**Solution.** See Figures 30.2 and 30.3.

### 30.9 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  $\text{He}^+$ , plus one each of  $e$ ,  $p$ ,  $\text{H}^-$ , and  $\text{He}^{++}$ , plus the ground levels of molecular hydrogen species  $\text{H}_2$  and  $\text{H}_2^+$ .

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 (30.84) an equivalent equation for helium, protons  $p$  being replaced by singly-ionized helium  $\text{He}^+$  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 (30.62),

$$A_{\text{He}21} = \frac{g_{\text{He}2s}}{g_{\text{He}2}} A_{\text{He}2s-1s} + \frac{g_{\text{He}2p}}{g_{\text{He}2}} A_{\text{He}2p-1s} P_S e^{(E_{\text{He}2s} - E_{\text{He}2p})/T}. \quad (30.85)$$

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 (30.63), the value of  $A_{\text{He}2p-1s}$  cancels against the Sobolev probability  $P_S$ , equation (30.101),

$$\frac{g_{\text{He}2p}}{g_{\text{He}2}} A_{\text{He}2p-1s} P_S = \frac{1}{X_{\text{He}1} n_+} \frac{g_{\text{He}1}}{g_{\text{He}2}} \frac{8\pi H}{\lambda_{\text{He}2p-1s}^3}, \quad (30.86)$$

the statistical weight factor being  $g_{\text{He}1}/g_{\text{He}2} = \frac{1}{4}$ .

In thermodynamic equilibrium,  $\text{He}^{++}$  combines to  $\text{He}^+$  at a redshift of  $z \sim 6,000$ , a factor of 6 higher than recombination, Figure 30.1. By the time recombination approaches, little  $\text{He}^{++}$  remains.  $\text{He}^{++}$  is well-approximated throughout as being in thermodynamic equilibrium with  $\text{He}^+$ .

Charge conservation implies that the electron fraction density  $X_e$  is

$$X_e = X_p + X_{\text{He}^+} + 2X_{\text{He}^{++}}. \quad (30.87)$$

The relevant atomic physics is as follows. The wavelengths of the  $2 \rightarrow 1$  transitions of hydrogen and helium are

$$\lambda_{\text{H}2p-1s} = 121.5682 \text{ nm}, \quad \lambda_{\text{He}2p-1s} = 58.4334 \text{ nm}, \quad \lambda_{\text{He}2s-1s} = 60.1404 \text{ nm}. \quad (30.88)$$

Ionization energies of hydrogen and helium, commonly quoted in units of  $\text{cm}^{-1}$ , are

$$\chi_{\text{H}} = 10,967,877.17 \text{ cm}^{-1}, \quad \chi_{\text{He}} = 19,831,066.9 \text{ cm}^{-1}, \quad \chi_{\text{He}^+} = 43,890,887.89 \text{ cm}^{-1}. \quad (30.89)$$

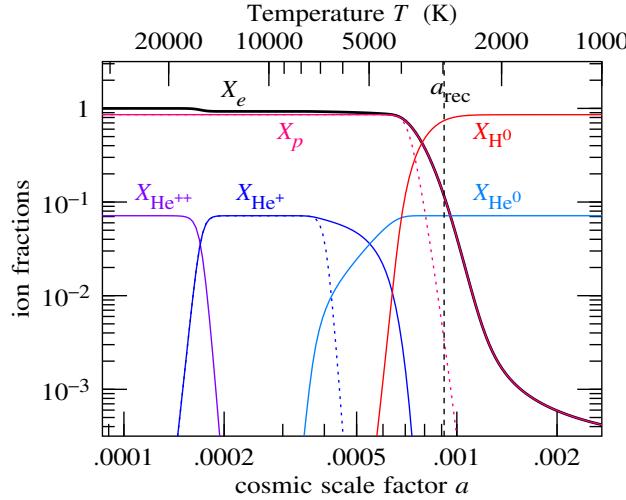


Figure 30.4 Non-equilibrium hydrogen and helium ion fractions as a function of cosmic scale factor  $a$  scaled to  $a_0 = 1$ . Dotted lines show  $X_p$  and  $X_{\text{He}^+}$  in thermodynamic equilibrium. The total hydrogen and helium fractions are  $X_{\text{H}} = 1 - f_n/f_+ = 0.86$  and  $X_{\text{He}^0} = \frac{1}{2}f_n/f_+ = 0.07$ .

The spontaneous 2-photon  $2s \rightarrow 1s$  transition rates of hydrogen and neutral helium are

$$A_{\text{H}2s-1s} = 8.22458 \text{ s}^{-1}, \quad A_{\text{He}2s-1s} = 51.3 \text{ s}^{-1}. \quad (30.90)$$

The effective recombination rates to  $n = 2$  levels of hydrogen and neutral helium are

$$\alpha_{\text{H}2}(T) = 1.14 \times 10^{-19} \frac{4.309 (T/10^4 \text{ K})^{-0.6166}}{1 + 0.6703 (T/10^4 \text{ K})^{0.5300}} \text{ m}^3 \text{ s}^{-1}, \quad (30.91a)$$

$$\alpha_{\text{He}2}(T) = 10^{-16.744} \left[ \sqrt{T/3 \text{ K}} \left( 1 + \sqrt{T/3 \text{ K}} \right)^{1-p} \left( 1 + \sqrt{T/10^{5.114} \text{ K}} \right)^{1+p} \right]^{-1} \text{ m}^3 \text{ s}^{-1}, \quad (30.91b)$$

with  $p = 0.711$ . The factor of 1.14 in the hydrogenic recombination rate (30.91a) 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 (30.91b) 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 (30.64), is  $g_{\text{He}^+}g_e/g_{\text{He}2} = 2 \times 2/4 = 1$ .

Figure 30.4 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.

### 30.10 Sobolev escape probability

The Sobolev escape probability formalism applies to a uniformly expanding medium such as a 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  $\phi_\lambda$ , which can be taken to be normalized to

$$\int_0^\infty \phi_\lambda d \ln \lambda = 1 . \quad (30.92)$$

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  $\sigma_\lambda$  for absorption at wavelength  $\lambda$  is

$$\sigma_\lambda = \sigma \phi_\lambda , \quad (30.93)$$

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

$$\sigma = \frac{1}{8\pi c} \frac{g_2}{g_1} \lambda_{21}^3 A_{21} . \quad (30.94)$$

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

$$d\tau_\lambda = n_1 \sigma_\lambda dl = n_1 c \sigma \phi_\lambda dt . \quad (30.95)$$

The medium is expanding with Hubble parameter  $H$ , and the photon wavelength  $\lambda$  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

$$d\tau_\lambda = \tau_S \phi_\lambda d \ln \lambda , \quad (30.96)$$

where  $\tau_S$  is the Sobolev optical depth

$$\tau_S \equiv \frac{n_1 c \sigma}{H} = n_1 \frac{g_2}{g_1} \frac{\lambda_{21}^3 A_{21}}{8\pi H} . \quad (30.97)$$

The optical depth  $\tau_\lambda$  for the photon to redshift from an emitted wavelength  $\lambda$  to infinite wavelength is

$$\tau_\lambda \equiv \tau_S \int_\lambda^\infty \phi_{\lambda'} d \ln \lambda' . \quad (30.98)$$

The probability for a photon emitted at wavelength  $\lambda$  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  $P_S$ ,

$$\begin{aligned} P_S &\equiv \int_0^\infty e^{-\tau_\lambda} \phi_\lambda d \ln \lambda = \int_0^\infty \exp \left( -\tau_S \int_\lambda^\infty \phi_{\lambda'} d \ln \lambda' \right) \phi_\lambda d \ln \lambda \\ &= \frac{1 - e^{-\tau_S}}{\tau_S}, \end{aligned} \quad (30.99)$$

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  $\tau_S$ , the Sobolev escape probability approximates the reciprocal of the Sobolev optical depth (30.97),

$$P_S = \frac{1}{\tau_S} \quad (\tau_S \gg 1). \quad (30.100)$$

The rate per unit time and volume at which photons are emitted and escape is then

$$n_2 A_{21} P_S = \frac{n_2}{n_1} \frac{g_1}{g_2} \frac{8\pi H}{\lambda_{21}^3}. \quad (30.101)$$

# 31

---

## Cosmological perturbations: the hydrodynamic approximation

The simple model in Chapter 29 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 29 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  $\sqrt{1/3}$ , §31.4. 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, §31.10. 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 33.7. 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 31.4, analogous to the acoustic oscillations in the CMB power spectrum.

A second important effect missing from the simple model of Chapter 29 is dissipation that results from the finite mean free path of electron-photon scattering, which causes photons and baryons not to be perfectly coupled, §31.7. 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, §31.11. 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

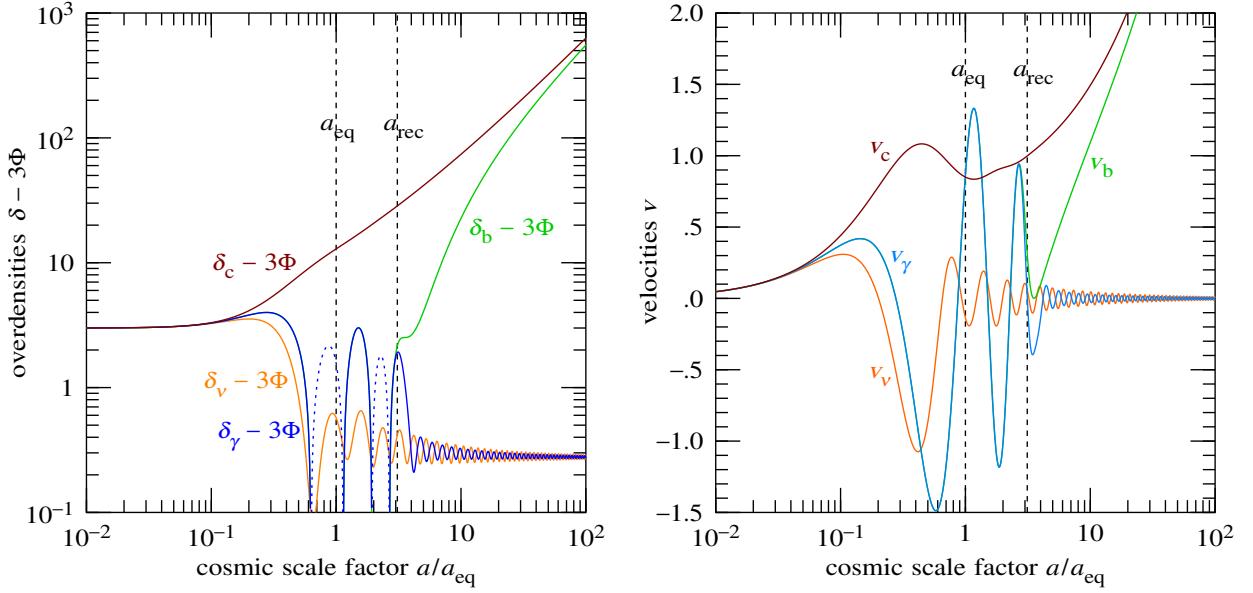


Figure 31.1 (Left) Overdensities  $\delta - 3\Phi$ , and (right) bulk velocities  $v$  in the hydrodynamic approximation as a function of cosmic scale factor  $a/a_{\text{eq}}$ , at wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ , for non-baryonic dark matter (c), baryons (b), photons ( $\gamma$ ), and neutrinos ( $\nu$ ). The cosmological model is the standard model adopted in this book, a flat  $\Lambda$ CDM model with concordance parameters  $\Omega_\Lambda = 0.69$  and  $\Omega_m = 0.31$ , and adiabatic initial conditions, §31.3. The overdensities and velocities of relativistic species are related to their monopole and dipole moments by  $\delta_\gamma - 3\Phi = 3(\Theta_0 - \Phi)$ ,  $\delta_\nu - 3\Phi = 3(N_0 - \Phi)$ ,  $v_\gamma = 3\Theta_1$ ,  $v_\nu = 3N_1$ . The results may be compared to those in the simple approximation, Figure 29.2, and from a Boltzmann computation, Figure 32.1.

treatment, Chapter 32. 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 31.1 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  $\Lambda$ CDM cosmological model at an illustrative wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ . Figure 31.2 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 32.

This chapter starts, §31.2, with a summary of the equations in the hydrodynamic approximation. The

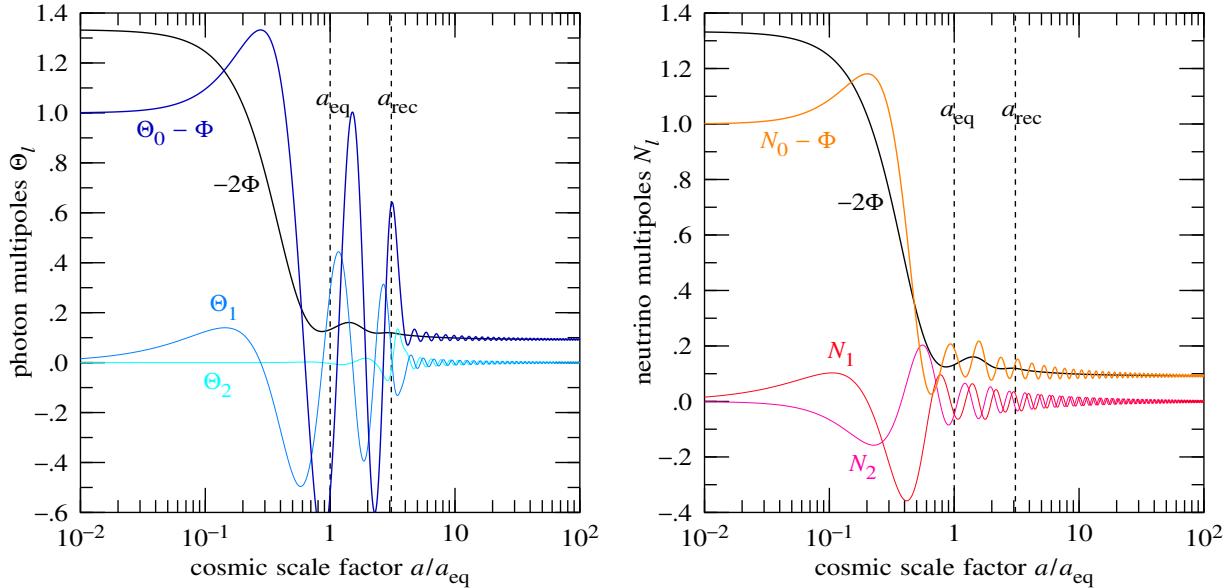


Figure 31.2 (Left) Photon and (right) neutrino multipoles in the hydrodynamic approximation as a function of cosmic scale factor  $a/a_{\text{eq}}$ , at wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ . The cosmological model is the same as in Figures 31.1–32.1, §31.3. The multipoles may be compared to those from a Boltzmann computation, Figure 32.2.

remainder of the chapter is concerned with finding approximations to the hydrodynamic system of equations (31.6)–(31.13), so as to gain a physical understanding of their solutions.

Section 31.4 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 29, but the mass density contributed by baryons reduces the sound speed of the fluid.

Sections 31.6–31.10 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 31.11 considers neutrinos, which stream freely. After recombination, photons also stream freely, as do baryons.

### 31.1 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

$$l_{\text{T}}^{-1} \equiv \bar{n}_e \sigma_{\text{T}} a , \quad (31.1)$$

where  $\sigma_{\text{T}}$  is the Thomson cross-section. The Thomson cross-section is proportional to the square of the classical electron radius  $r_e$ ,

$$\sigma_{\text{T}} = \frac{8\pi}{3} r_e^2 , \quad r_e = \frac{e^2}{m_e c^2} . \quad (31.2)$$

The inverse comoving mean free path  $l_{\text{T}}^{-1}$  is evaluated in Exercise 31.1. In calculating fluctuations in the CMB, Chapter 33, it is convenient to introduce the (dimensionless) Thomson scattering optical depth  $\tau$ , which starts from zero,  $\tau_0 = 0$ , at the present time, and increases going backwards in time  $\eta$  to higher redshift,

$$\tau \equiv \int_{\eta}^{\eta_0} \bar{n}_e \sigma_{\text{T}} a d\eta . \quad (31.3)$$

The conformal time derivative of the Thomson optical depth  $\tau$  equals minus the inverse comoving mean free path,

$$\dot{\tau} \equiv \frac{d\tau}{d\eta} \equiv -\bar{n}_e \sigma_{\text{T}} a . \quad (31.4)$$

**Exercise 31.1 Thomson scattering rate.** Let  $f_+$  be the proton fraction (30.5), and  $X_e$  be the ionization fraction (30.3). Show that the (dimensionless) ratio of the inverse comoving electron-photon (Thomson) mean free path  $l_{\text{T}}^{-1} = \bar{n}_e \sigma_{\text{T}} a$  to the inverse comoving Hubble distance  $a_{\text{eq}} H_{\text{eq}}/c$  at matter-radiation equality is

$$\begin{aligned} \frac{c \bar{n}_e \sigma_{\text{T}} a}{a_{\text{eq}} H_{\text{eq}}} &= \frac{3c \sigma_{\text{T}} f_+ X_e H_{\text{eq}} \Omega_b}{16\pi G m_b \Omega_m} \left( \frac{a}{a_{\text{eq}}} \right)^{-2} \\ &= 0.033 h f_+ X_e \frac{H_{\text{eq}}}{H_0} \frac{\Omega_b}{\Omega_m} \left( \frac{a}{a_{\text{eq}}} \right)^{-2} = 500 X_e \left( \frac{a}{a_{\text{eq}}} \right)^{-2} , \end{aligned} \quad (31.5)$$

the Hubble parameter  $H_{\text{eq}}$  at matter-radiation equality being related to the present-day Hubble parameter  $H_0$  by equation (29.39).

## 31.2 Summary of equations in the hydrodynamic approximation

The hydrodynamic approximation is derived by suitably truncating the full set of Boltzmann equations, §32.1, 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, §29.7, and in a full Boltzmann treatment, §32.1). The equations for non-baryonic cold dark matter (c) follow from conservation

of energy-momentum, and are the same as those (29.50) in the simple approximation (recall that overdot signifies the derivative  $d/d\eta$  with respect to conformal time),

$$\dot{\delta}_c - k v_c - 3 \dot{\Phi} = 0 , \quad (31.6a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k \Psi = 0 . \quad (31.6b)$$

Equations for baryons (b) are similar to those (31.6) 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,

$$\dot{\delta}_b - k v_b - 3 \dot{\Phi} = 0 , \quad (31.7a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k \Psi = - \frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) . \quad (31.7b)$$

The equations of conservation of energy and momentum of photons ( $\gamma$ ) are

$$\dot{\Theta}_0 - k \Theta_1 - \dot{\Phi} = 0 , \quad (31.8a)$$

$$\dot{\Theta}_1 + \frac{k}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3} \Psi = \frac{1}{3} |\dot{\tau}| (v_b - 3\Theta_1) . \quad (31.8b)$$

The photon quadrupole moment  $\Theta_2$  can be approximated by an expression that interpolates between the large scattering limit  $|\dot{\tau}| \gg k_s$  and the free-streaming limit  $|\dot{\tau}| \ll k_s$ , 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  $\Lambda$ CDM cosmological parameters),

$$\Theta_2 = - \frac{1}{(1 + |\dot{\tau}|/k_s)^2} \left[ \frac{|\dot{\tau}|}{k_s} \frac{8k}{15k_s} \Theta_1 + (\Theta_0 + \Psi) + \frac{3}{k\eta} \Theta_1 \right] \rightarrow \begin{cases} - \frac{8k}{15|\dot{\tau}|} & |\dot{\tau}| \gg k_s , \\ - (\Theta_0 + \Psi) - \frac{3}{k\eta} \Theta_1 & |\dot{\tau}| \ll k_s . \end{cases} \quad (31.9)$$

As commented after equation (31.67), the factor  $\frac{8}{15}$  in equation (31.9) includes the effect of polarization; without polarization, the factor is  $\frac{4}{9}$ . Energy-momentum conservation of neutrinos ( $\nu$ ) implies

$$\dot{\mathcal{N}}_0 - k \mathcal{N}_1 - \dot{\Phi} = 0 , \quad (31.10a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3} (\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3} \Psi = 0 . \quad (31.10b)$$

The neutrino quadrupole  $\mathcal{N}_2$  may be approximated by, equation (33.48),

$$\mathcal{N}_2 = - (\mathcal{N}_0 + \Psi) - \frac{3}{k\eta} \mathcal{N}_1 . \quad (31.11)$$

The Einstein energy equation is

$$- k^2 \Phi - 3 \frac{\dot{a}}{a} F = 4\pi G a^2 (\bar{\rho}_c \delta_c + \bar{\rho}_b \delta_b + 4\bar{\rho}_\gamma \Theta_0 + 4\bar{\rho}_\nu \mathcal{N}_0) , \quad (31.12)$$

where  $F$  is defined by equation (29.53). The non-vanishing photon and neutrino quadrupoles  $\Theta_2$  and  $\mathcal{N}_2$  are a source for the difference  $\Psi - \Phi$  in scalar gravitational potentials, equation (28.40d). However, the hydrodynamic approximations (31.9) and (31.11) 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 (29.55),

$$\Psi = \Phi . \quad (31.13)$$

**Exercise 31.2 Program the equations in the hydrodynamic approximation.** Upgrade the code you wrote in Exercise 29.10 to implement the hydrodynamic approximation, equations (31.6)–(31.13). Explore the evolution of the gravitational potential  $\Phi$ , and of the 4 species of mass-energy, non-baryonic dark matter, baryons, photons, and neutrinos.

**Solution.** See Figures 31.1, 31.2, and 31.3.

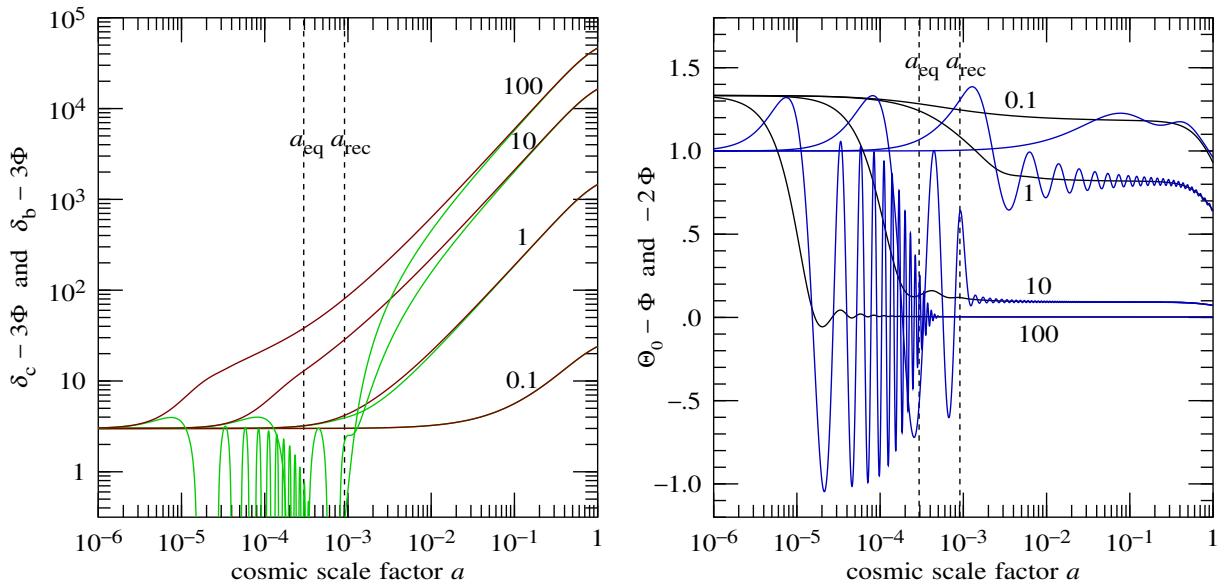


Figure 31.3 (Left) Overdensities  $\delta_c - 3\Phi$  and  $\delta_b - 3\Phi$  of non-baryonic dark matter (brown) and baryonic matter (green), and (right) radiation monopole  $\Theta_0 - 3\Phi$  (blue), and minus twice the scalar potential,  $-2\Psi$  (black), as a function of cosmic scale factor  $a$  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 29.1, and using a Boltzmann computation, Figure 32.3.

**Exercise 31.3 Power spectrum of matter fluctuations: hydrodynamic approximation.** Upgrade the code you wrote in Exercise 29.15 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.

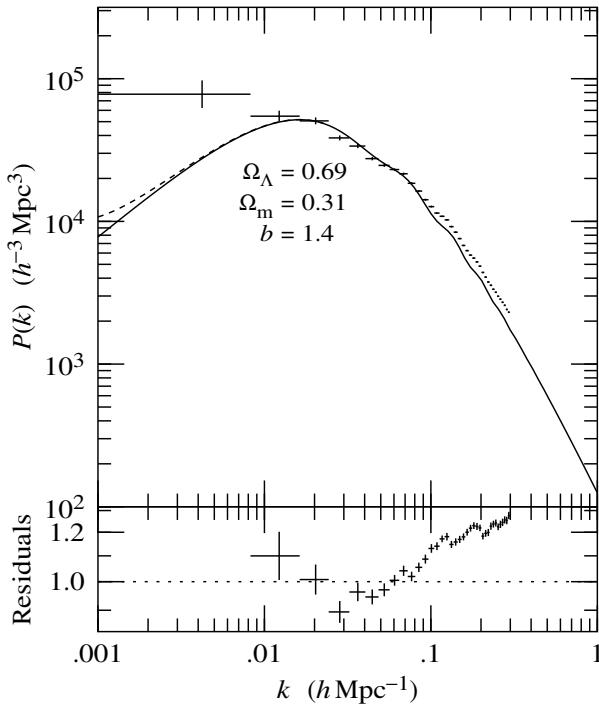


Figure 31.4 Model matter power spectrum computed in the hydrodynamic approximation, compared to observations from the BOSS galaxy survey (Anderson, 2014). The predicted power spectrum has been multiplied, arbitrarily, by a bias factor of  $b^2 = 1.4^2$ . The model power spectrum may be compared to those computed in the simple approximation, Figure 29.15, and from a Boltzmann computation, Figure 32.5.

**Solution.** See Figure 31.4. The cosmological model is the standard flat  $\Lambda$ CDM model described in §31.3. The model power spectrum differs from that in the simple approximation, Figure 29.15, 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. Anderson (2014) were interested primarily in the scale of the BAO. In Figure 10.3 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 31.4 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 non-relativistic.

### 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  $c = 1$ )

$$\rho(m, T) = \int_0^\infty \sqrt{p^2 + m^2} \frac{1}{e^{p/T} + 1} \frac{4\pi p^2 dp}{(2\pi\hbar)^3} = \frac{7\pi^2 T^4}{240 \hbar^3} R(m/T), \quad (31.14)$$

where  $R(\mu)$  is the integral

$$R(\mu) \equiv \frac{120}{7\pi^4} \int_0^\infty \frac{\sqrt{x^2 + \mu^2}}{e^x + 1} x^2 dx \rightarrow \begin{cases} 1 & \mu \rightarrow 0, \\ \alpha\mu & \mu \rightarrow \infty, \end{cases} \quad (31.15)$$

with

$$\alpha = \frac{180\zeta(3)}{7\pi^4} = 0.3173. \quad (31.16)$$

The neutrino pressure  $p(m, T)$  (not to be confused with neutrino particle momentum  $p$ ) is

$$p(m, T) = \frac{1}{3} \int_0^\infty \frac{p^2}{\sqrt{p^2 + m^2}} \frac{1}{e^{p/T} + 1} \frac{4\pi p^2 dp}{(2\pi\hbar)^3}, \quad (31.17)$$

which can be expressed in terms of the neutrino density  $\rho(m, T)$  as

$$p(m, T) = \frac{1}{3} \left[ \rho(m, T) - \frac{\partial \rho(m, T)}{\partial \ln m} \right]. \quad (31.18)$$

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 (31.18), is

$$R(\mu) \approx \sqrt{\frac{1 + \beta\mu^2 + \gamma\alpha^2\mu^4}{1 + \gamma\mu^2}}, \quad (31.19)$$

where the constants  $\alpha$  (equation (31.16)),  $\beta, \gamma$  are chosen such that both density  $\rho$  and pressure  $p$  have the correct asymptotic behaviour at both  $\mu \rightarrow 0$  and  $\mu \rightarrow \infty$ ,

$$\beta = \frac{10}{7\pi^2} + \gamma = 0.2902, \quad \gamma = \frac{10[-7\pi^2 + 3240\zeta(3)^2]}{49\pi^8 - 486000\zeta(3)\zeta(5)} = 0.1454. \quad (31.20)$$

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 non-relativistic at  $T = \alpha m$ ,

$$\rho(m, T) \approx \frac{7\pi^2 T^3}{240 \hbar^3} \begin{cases} T & T \geq \alpha m \\ \alpha m & T \leq \alpha m \end{cases}, \quad (31.21)$$

2. The approximation (31.21) for the neutrino density suggests adopting an abrupt transition of the neutrino evolution equations from relativistic, equations (31.10) and (31.11), to non-relativistic at  $T = \alpha m$ , with  $\alpha$  from equation (31.16). The non-relativistic equations are as for non-baryonic cold dark matter, equations (31.6). Conservation of energy and momentum at the transition requires that the neutrino overdensity  $\delta_\nu$  and bulk velocity  $v_\nu$  are

$$\delta_\nu = 3\mathcal{N}_0, \quad (31.22a)$$

$$v_\nu = 3\mathcal{N}_1. \quad (31.22b)$$

### 31.3 Standard cosmological parameters

Unless otherwise stated, all computations of cosmological perturbations carried out in this book are for a standard flat  $\Lambda$ CDM cosmological model with parameters consistent with those reported by the Planck collaboration (Ade, 2015). This section gives the standard parameters adopted in this book.

The CMB power spectrum constrains the physical density  $\Omega h^2$  of dark matter and baryonic components more precisely than the density  $\Omega$  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,

$$\Omega_c h^2 = 0.12, \quad \Omega_b h^2 = 0.022. \quad (31.23)$$

The conversion factor between  $\Omega h^2$  and mass density  $\rho$  today is

$$\rho = \frac{3\Omega H^2}{8\pi G c^2} = 6.44932 \times 10^{-26} \Omega h^2 \text{ kg m}^{-3}. \quad (31.24)$$

The matter density  $\Omega_m$  in non-baryonic cold dark matter and baryonic components today is taken to be, in the standard model,

$$\Omega_m = 0.31. \quad (31.25)$$

The individual non-baryonic cold dark matter and baryonic densities are then

$$\Omega_c = 0.262, \quad \Omega_b = 0.048. \quad (31.26)$$

Together, equations (31.23) and (31.25) yield a Hubble parameter  $H_0$  today of

$$H_0 \equiv 100 h \text{ km s}^{-1} \text{ Mpc}^{-1} = 67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}. \quad (31.27)$$

The CMB temperature  $T_0$  today is (Fixsen, 2009)

$$T_0 = 2.7255 \text{ K}, \quad (31.28)$$

implying a physical photon density of, Exercise 10.2,

$$\Omega_\gamma h^2 = 2.4728 \times 10^{-5} . \quad (31.29)$$

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 10.20. The energy-weighted effective number of relativistic particle species at recombination is then, equation (10.149b),

$$g_\rho = 2 \left[ 1 + \left( \frac{4}{11} \right)^{4/3} \frac{7}{8} N_{\text{eff}} \right] = 3.36 . \quad (31.30)$$

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  $g_\rho$  of relativistic species at recombination is correspondingly slightly larger. It is conventional to quote the increase in  $g_\rho$  as if it were an increase in the effective number of neutrino types in equation (31.30),  $N_{\text{eff}} = 3.04$  (Mangano et al., 2002). In the approximation of  $N_{\text{eff}} = 3$  massless neutrinos adopted here, the physical density of neutrinos today is

$$\Omega_\nu h^2 = 1.68 \times 10^{-5} . \quad (31.31)$$

The ratio of the physical matter density from equations (31.23) to the physical radiation density implied by equations (31.29) and (31.31) implies a redshift of matter-radiation equality of

$$1 + z_{\text{eq}} = 3415 . \quad (31.32)$$

If neutrinos have masses as indicated by neutrino oscillation data, §10.25, then at least 2 of the 3 neutrino species are non-relativistic today, even though they were relativistic at recombination. If the third species is taken to be massless, then the neutrino masses are

$$m_{\nu 1} = 0 \text{ eV} , \quad m_{\nu 2} = 0.01 \text{ eV} , \quad m_{\nu 3} = 0.05 \text{ eV} . \quad (31.33)$$

The corresponding physical neutrino mass density today is, in place of equation (31.31),

$$\Omega_\nu h^2 = 1.3 \times 10^{-3} . \quad (31.34)$$

The assumption of spatial flatness implies vanishing spatial curvature,

$$\Omega_k = 0 . \quad (31.35)$$

In the standard  $\Lambda$ CDM model, the remaining density is taken to be vacuum energy, equivalent to a cosmological constant, with density

$$\Omega_\Lambda = 1 - \Omega_k - \Omega_m - \Omega_r = 0.69 . \quad (31.36)$$

Recombination is affected by the helium mass fraction  $Y_{^4\text{He}} \equiv \rho_{^4\text{He}} / (\rho_{\text{H}} + \rho_{^4\text{He}})$ , taken to be (Ade, 2015)

$$Y_{^4\text{He}} = 0.25 . \quad (31.37)$$

Given the helium fraction (31.37) and the Peebles approximation to recombination, §30.8, along with the standard parameters adopted here, the redshift of recombination, where the Thomson optical depth is unity, is

$$1 + z_{\text{rec}} = 1092 . \quad (31.38)$$

In integrating the simple or hydrodynamic or Boltzmann equations, §29.7 or §31.2 or §32.1, 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

$$\frac{a_0}{a_{\text{eq}}} = 3415 , \quad \frac{H_0}{H_{\text{eq}}} = 6.363 \times 10^{-6} . \quad (31.39)$$

The scale factor and Hubble parameter at recombination, where the Thomson optical depth  $\tau$  is one, are related to those at matter-radiation equality by

$$\frac{a_{\text{rec}}}{a_{\text{eq}}} = 3.13 , \quad \frac{H_{\text{rec}}}{H_{\text{eq}}} = 0.147 . \quad (31.40)$$

The Hubble distance today relative to those at matter-radiation equality and at recombination are

$$\frac{c}{a_0 H_0} = 46.0 \frac{c}{a_{\text{eq}} H_{\text{eq}}} = 21.1 \frac{c}{a_{\text{rec}} H_{\text{rec}}} . \quad (31.41)$$

In cosmology, distances are commonly reported in units of  $h^{-1}$  Mpc, or, if the Hubble parameter  $h$  today is considered to be known, in Mpc. The Hubble distance today is

$$\frac{c}{a_0 H_0} = 2997.92458 h^{-1} \text{ Mpc} = 4.43 \text{ Gpc} . \quad (31.42)$$

The horizon distance today is

$$\eta_0 = 147 \frac{c}{a_{\text{eq}} H_{\text{eq}}} = 3.20 \frac{c}{a_0 H_0} = 9600 h^{-1} \text{ Mpc} = 14.2 \text{ Gpc} . \quad (31.43)$$

The age of the Universe today is, equation (10.15),

$$t_0 = 0.955 H_0^{-1} = 13.8 \text{ Gyr} . \quad (31.44)$$

## 31.4 The photon-baryon fluid in the tight-coupling approximation

Prior to recombination, non-relativistic 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 approaches. The density loading decreased the

sound speed of the photon-baryon fluid, equation (31.50), 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 non-relativistic, have negligible pressure, so their energy is just that of their rest mass. Consequently the energy conservation equation (29.13a) holds separately for each of the photon and baryon fluids. However, the exchange of momentum means that the momentum equation (29.13b) 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,

$$v_b = v_\gamma , \quad (31.45)$$

the photon bulk velocity being related to the photon dipole by  $v_\gamma = 3\Theta_1$ , equation (29.15). The approximation (31.45) is called the **tight-coupling** approximation. The right panel of Figure 31.1 illustrates that the equality (31.45) 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,

$$R \equiv \frac{3\bar{\rho}_b}{4\bar{\rho}_\gamma} = R_a \frac{a}{a_{eq}} , \quad R_a = \frac{3g_\rho\Omega_b}{8\Omega_m} \approx 0.2 , \quad (31.46)$$

with  $g_\rho = 3.36$  being the energy-weighted effective number of relativistic particle species at around the time of recombination, equation (10.149b). The energy flux of the combined photon-baryon fluid is, from equation (29.9b),

$$f_\gamma + f_b = (\bar{\rho}_\gamma + \bar{\rho}_b)v_\gamma + \bar{\rho}_b v_b = \frac{4}{3}\bar{\rho}_\gamma v(1+R) , \quad (31.47)$$

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 (29.13b) with  $w = 1/3$ , and  $R$  times the baryon-velocity equation (29.13b) with  $w = 0$ . The resulting momentum conservation equation is

$$(1+R)\dot{v} + R\frac{\dot{a}}{a}v + \frac{k}{3}\delta_\gamma = -k(1+R)\Psi . \quad (31.48)$$

Combining the photon energy conservation equation (29.13a) with the momentum conservation equation (31.48), and substituting  $\delta_\gamma = 3\Theta_0$ , equation (29.15), yields

$$\left[ \frac{d^2}{d\eta^2} + \frac{R}{1+R} \frac{\dot{a}}{a} \frac{d}{d\eta} + \frac{k^2}{3(1+R)} \right] (\Theta_0 - \Phi) = -\frac{k^2}{3(1+R)} [(1+R)\Psi + \Phi] , \quad (31.49)$$

which coincides with equation (29.14) for  $w = 1/[3(1+R)]$ , and which goes over to the earlier radiation equation (29.45) 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 (31.49) is an adiabatic damping term. In the absence of this term, and in the absence of a driving potential, equation (31.49) would reduce to a wave equation with sound speed

$$c_s = \sqrt{\frac{1}{3(1+R)}} . \quad (31.50)$$

The coefficient of the adiabatic damping term in equation (31.49) is, given that  $R \propto a$ , equation (31.46),

$$\frac{R}{1+R} \frac{\dot{a}}{a} = -2 \frac{\dot{c}_s}{c_s} . \quad (31.51)$$

The **sound horizon distance**  $\eta_s$  is defined to be the distance travelled by a sound wave since the initial time  $\eta = 0$ ,

$$\eta_s \equiv \int_0 c_s d\eta . \quad (31.52)$$

Recast in terms of the sound horizon distance  $\eta_s$ , the differential equation (31.49) is

$$\left( \frac{d^2}{d\eta_s^2} - \frac{c'_s}{c_s} \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = -k^2 [(1+R)\Psi + \Phi] , \quad (31.53)$$

where prime ' denotes derivatives with respect to the sound horizon distance,  $c'_s = dc_s/d\eta_s$ .

### 31.5 WKB approximation

Equation (31.53) 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  $2\kappa$ . The homogeneous version of equation (31.53) is then

$$\left( \frac{d^2}{d\eta_s^2} + 2\kappa \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = 0 . \quad (31.54)$$

In case being considered, the damping rate  $\kappa$  is the adiabatic rate

$$\kappa = -\frac{1}{2} \frac{c'_s}{c_s} , \quad (31.55)$$

but the WKB method works for more general  $\kappa$ , provided that  $\kappa$  is small compared to the wavenumber of the sound wave,  $\kappa \ll k$ . If the damping rate  $\kappa$  is treated as approximately constant, then the homogeneous wave equation (31.54) can be solved by introducing a frequency  $\omega$  defined by

$$\Theta_0 - \Phi \propto e^{\int \omega d\eta_s} . \quad (31.56)$$

The homogeneous wave equation (31.54) is then equivalent to

$$\omega' + \omega^2 + 2\kappa\omega + k^2 = 0 . \quad (31.57)$$

To the extent that the damping parameter is much smaller than the frequency,  $\kappa \ll k \sim \omega$ , the frequency  $\omega$

is approximately constant, so that  $\omega'$  can be neglected in equation (31.57). With  $\omega'$  neglected, the solution of equation (31.57) is

$$\omega = -\kappa \pm i\sqrt{k^2 - \kappa^2} \approx -\kappa \pm ik , \quad (31.58)$$

where the last approximation holds because  $\kappa \ll k$ . Equation (31.58) is called the **WKB approximation**. Thus the homogeneous solutions of the wave equation (31.54) are approximately

$$\Theta_0 - \Phi \propto e^{-\int \kappa d\eta_s \pm ik\eta_s} . \quad (31.59)$$

### 31.5.1 Radiation in the tight-coupling approximation

In the tight-coupling approximation, the damping rate  $\kappa$  in the differential equation (31.54) is the adiabatic damping rate (31.55). The integral of the adiabatic damping term is  $\int \kappa_a d\eta_s = -\frac{1}{2} \ln c_s$ , whose exponential is

$$e^{-\int \kappa_a d\eta_s} = \sqrt{c_s} . \quad (31.60)$$

In the WKB approximation, the homogeneous solutions to the wave equation (31.54) are

$$\Theta_0 - \Phi \propto \sqrt{c_s} e^{\pm ik\eta_s} . \quad (31.61)$$

This shows that, as the sound speed decreases thanks to the increasing baryon-to-photon density in the expanding Universe, the amplitude of a sound wave decreases as the square root of the sound speed.

## 31.6 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, §31.7.

The scalar part of the momentum conservation equation (28.56b) 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,

$$T_{\text{quad}}^{ab} = (\bar{\rho} + \bar{p})q \left( \frac{3}{2}\hat{k}_a\hat{k}_b - \frac{1}{2}\delta_{ab} \right) , \quad (\bar{\rho} + \bar{p})q \equiv \left( \hat{k}_a\hat{k}_b - \frac{1}{3}\delta_{ab} \right) T^{ab} . \quad (31.62)$$

For relativistic species such as photons, the dimensionless quadrupole  $q$  is related to the quadrupole moment  $\Theta_2$  by, equation (??),

$$q = -2\Theta_2 . \quad (31.63)$$

In the presence of a quadrupole pressure, the momentum conservation equation (28.56b) includes a term

$$D_m T_{\text{quad}}^{ma} = \frac{1}{a}(\bar{\rho} + \bar{p})\nabla_a q . \quad (31.64)$$

The net effect is to modify all momentum conservation equations by replacing  $\Psi \rightarrow \Psi + q$ . The scalar bulk velocity equation (29.13b) is thus modified to

$$\dot{v} + (1 - 3w) \frac{\dot{a}}{a} v + wk\delta = -k(\Psi + q) . \quad (31.65)$$

### 31.7 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 (31.84). 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 (32.81) 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 §31.4, 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  $\ell$ . 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 (32.81) is the lowest order multipole, which acts as a driver. Solution of the Boltzmann equations (32.81) then requires that

$$\Theta_{\ell+1} \sim \frac{k}{|\dot{\tau}|} \Theta_\ell \quad \text{for } \ell \geq 2 . \quad (31.66)$$

The relation (31.66) 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  $\Theta_2$  in the Boltzmann hierarchy (32.81) of photon multipole equations, but still neglecting the higher multipoles,  $\Theta_\ell = 0$  for  $\ell \geq 3$ . According to the estimate (31.66), 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 (32.81c) reduces to

$$\Theta_2 = -\frac{4k}{9|\dot{\tau}|} \Theta_1 . \quad (31.67)$$

Substituted into the photon momentum equation (31.8b), the photon quadrupole  $\Theta_2$  (31.67) acts as a source of friction on the photon dipole  $\Theta_1$ . 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 (32.81c), the factor  $\frac{4}{9}$  in equation (31.67) is increased by a factor of  $\frac{6}{5}$  to  $\frac{8}{15}$ , equation (34.53), as already adopted in equation (31.9).

The diffusive damping resulting from a small photon quadrupole  $\Theta_2$  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 §31.6.

### 31.8 Viscous baryon drag damping

A second source of damping of sound waves, distinct from the photon diffusion of §31.7, 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 (31.8b) or (31.7b) for photons or baryons,

$$v_b - 3\Theta_1 = \frac{3}{|\dot{\tau}|} \left( \dot{\Theta}_1 + \frac{k}{3}\Theta_0 + \frac{k}{3}\Psi \right) = -\frac{R}{|\dot{\tau}|} \left( \dot{v}_b + \frac{\dot{a}}{a}v_b + k\Psi \right) \approx -\frac{3R}{|\dot{\tau}|} \left( \dot{\Theta}_1 + \frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}\Psi \right). \quad (31.68)$$

The bulk velocity difference  $v_b - 3\Theta_1$  is small because the scattering factor  $|\dot{\tau}|$  is large. The final approximation of equations (31.68) 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 (31.68) so as to eliminate  $\dot{\Theta}_1$  gives

$$v_b - 3\Theta_1 = \frac{3R}{(1+R)|\dot{\tau}|} \left( \frac{k}{3}\Theta_0 - \frac{\dot{a}}{a}\Theta_1 \right). \quad (31.69)$$

On the right hand side of equation (31.69), 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

$$v_b - 3\Theta_1 \approx \frac{Rk}{(1+R)|\dot{\tau}|} \Theta_0. \quad (31.70)$$

It is tempting to insert the approximation (31.70) directly into the right hand sides of the photon and baryon momentum conservation equations (31.8b) and (31.7b), 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 (31.70). 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 (31.8b) and (31.7b) for photons and baryons,

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi + \frac{R}{3} \left( \dot{v}_b + \frac{\dot{a}}{a}v_b + k\Psi \right) = 0. \quad (31.71)$$

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 (31.71) to

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi + R\left(\dot{\Theta}_1 + \frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}\Psi\right) + \frac{R}{3}\left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)(v_b - 3\Theta_1) = 0. \quad (31.72)$$

The term in equation (31.72) involving the velocity difference  $v_b - 3\Theta_1$  is proportional to

$$\left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)(v_b - 3\Theta_1) = \left(\frac{d}{d\eta} + \frac{\dot{a}}{a}\right)\frac{Rk}{(1+R)|\dot{\tau}|}\Theta_0 \approx \frac{Rk}{(1+R)|\dot{\tau}|}\dot{\Theta}_0 \approx \frac{Rk^2}{(1+R)|\dot{\tau}|}\Theta_1, \quad (31.73)$$

the second step of which invokes the approximation (31.70), 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 (31.73), and the diffusive approximation (31.67) for the radiation quadrupole  $\Theta_2$ , brings the photon-baryon momentum conservation equation (31.72) to

$$(1+R)\dot{\Theta}_1 + R\frac{\dot{a}}{a}\Theta_1 + \frac{k}{3}[\Theta_0 + (1+R)\Psi] + \frac{k^2}{3|\dot{\tau}|}\left(\frac{8}{9} + \frac{R^2}{1+R}\right)\Theta_1 = 0. \quad (31.74)$$

The final terms proportional to the comoving Thomson mean free path  $1/(|\dot{\tau}|)$  on the left hand side of equation (31.74) are the dissipative terms. The  $8/9$  term is from photon diffusion, while the  $R^2/(1+R)$  term is from baryon drag.

### 31.9 Photon-baryon wave equation with dissipation

Eliminating the dipole  $\Theta_1$  in equation (31.74) in favour of the monopole  $\Theta_0$  using the photon monopole equation (31.8) yields a second order differential equation for  $\Theta_0 - \Phi$ ,

$$\left\{ \frac{d^2}{d\eta^2} + \left[ \frac{R}{(1+R)}\frac{\dot{a}}{a} + \frac{k^2}{3(1+R)|\dot{\tau}|} \left( \frac{8}{9} + \frac{R^2}{1+R} \right) \right] \frac{d}{d\eta} + \frac{k^2}{3(1+R)} \right\} (\Theta_0 - \Phi) = -\frac{k^2}{3(1+R)} [(1+R)\Psi + \Phi]. \quad (31.75)$$

Recast in terms of the sound horizon distance  $\eta_s$  defined by equation (31.52), the wave equation (31.75) becomes

$$\boxed{\left\{ \frac{d^2}{d\eta_s^2} + \left[ -\frac{c'_s}{c_s} + \frac{k^2 c_s}{|\dot{\tau}|} \left( \frac{8}{9} + \frac{R^2}{1+R} \right) \right] \frac{d}{d\eta_s} + k^2 \right\} (\Theta_0 - \Phi) = -k^2 [(1+R)\Psi + \Phi]}, \quad (31.76)$$

where prime ' denotes derivative with respect to sound horizon distance,  $c'_s = dc_s/d\eta_s$ . Equations (31.75) and (31.76) differ from the earlier dissipation-free equations (31.49) and (31.53) 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 (31.76) was discussed in §31.5.

In Exercise 34.3 it is found that polarization increases the photon diffusion contribution in equation (31.76) by a factor of  $\frac{6}{5}$  from  $\frac{8}{9}$  to  $\frac{16}{15}$ ,

$$\frac{8}{9} \rightarrow \frac{16}{15}. \quad (31.77)$$

The terms proportional to the linear derivative  $d/d\eta_s$  in equation (31.76) are damping terms, which may be collected into an overall damping coefficient  $\kappa$ ,

$$\left( \frac{d^2}{d\eta_s^2} + 2\kappa \frac{d}{d\eta_s} + k^2 \right) (\Theta_0 - \Phi) = -k^2 [(1+R)\Psi + \Phi] . \quad (31.78)$$

The damping coefficient  $\kappa$  is a sum of adiabatic  $\kappa_a$  and dissipative  $\kappa_d$  parts,

$$\kappa \equiv \kappa_a + \kappa_d , \quad \kappa_a = -\frac{1}{2} \frac{d \ln c_s}{d\eta_s} , \quad \kappa_d = \frac{k^2 c_s}{2|\dot{\tau}|} \left( \frac{16}{15} + \frac{R^2}{1+R} \right) . \quad (31.79)$$

In the dissipative damping coefficient  $\kappa_d$ , 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, §31.5, the homogeneous solutions of equation (31.78) are

$$\Theta_0 - \Phi \propto \sqrt{c_s} e^{-\int \kappa_d d\eta_s} e^{\pm ik\eta_s} . \quad (31.80)$$

The dissipative factor  $e^{-\int \kappa_d d\eta_s}$  involves an integral over the dissipative damping coefficient, which may be written

$$\int \kappa_d d\eta_s = \frac{k^2}{k_d^2} , \quad (31.81)$$

where  $k_d^{-1}$  is the damping scale defined by

$$\frac{1}{k_d^2} \equiv \int \frac{c_s}{2|\dot{\tau}|} \left( \frac{16}{15} + \frac{R^2}{1+R} \right) d\eta_s = \int \frac{1}{6|\dot{\tau}|(1+R)} \left( \frac{16}{15} + \frac{R^2}{1+R} \right) d\eta . \quad (31.82)$$

The damping scale  $k_d^{-1}$  is roughly the geometric mean of the scattering mean free path  $l_T$  and the horizon distance  $\eta$ , as might be expected for a random walk by increments  $l_T$  over a time  $\eta$ ,

$$k_d^{-1} \sim \sqrt{l_T \eta} . \quad (31.83)$$

The resulting dissipative damping factor is

$$e^{-\int \kappa_d d\eta_s} = e^{-k^2/k_d^2} . \quad (31.84)$$

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 31.6.

### 31.10 Baryon loading

The driving potential on the right hand side of the wave equation (31.78) causes  $\Theta_0 - \Phi$  to oscillate not around zero, but rather around the offset  $-(1+R)\Psi + \Phi$ . 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 (31.78) well inside the horizon is

$$\Theta_0 + (1+R)\Psi \propto \sqrt{c_s} e^{-k^2/k_d^2} e^{\pm ik\eta_s}. \quad (31.85)$$

As will be seen in Chapter 33, equation (33.15), the monopole contribution to CMB fluctuations is not the photon monopole  $\Theta_0$  by itself, but rather  $\Theta_0 + \Psi$ , which is the monopole redshifted by the potential  $\Psi$ . This redshifted monopole is

$$\Theta_0 + \Psi = -R\Psi + A\sqrt{c_s} e^{-k^2/k_d^2} e^{\pm ik\eta_s}, \quad (31.86)$$

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 31.5 Behaviour of radiation in the presence of damping.** Confirm that, for  $\kappa \ll k$ , the homogeneous solutions of equation (31.78) are approximately  $\Theta_0 - \Phi \propto e^{-\int \kappa d\eta_s} \pm ik\eta_s$ . Hence find the retarded Green's function, and write down the general solution to equation (31.78). Convince yourself that  $\Theta_0 - \Phi$  oscillates around  $-(1+R)\Psi + \Phi$ .

**Solution.** The general solution of equation (31.78) is, with  $y \equiv k\eta_s$  and  $\beta \equiv \int \kappa d\eta_s$ ,

$$\Theta_0(y) - \Phi(y) = e^{-\beta} (A_0 \cos y + A_1 \sin y) - \int_0^y \{[1+R(y')] \Psi(y') + \Phi(y')\} e^{-(\beta-\beta')} \sin(y-y') dy', \quad (31.87)$$

where  $A_0$  and  $A_1$  are constants.

**Exercise 31.6 Diffusion scale.** Show that the dimensionless ratio of the damping scale  $k_d$  defined by (31.82) to the comoving Hubble distance  $c/(a_{\text{eq}}H_{\text{eq}})$  at matter-radiation equality is given by

$$\frac{a_{\text{eq}}^2 H_{\text{eq}}^2}{c^2 k_d^2} = \frac{8\sqrt{2}\pi G m_b}{9c\sigma_T f_+ H_{\text{eq}}} \frac{\Omega_m}{\Omega_b} \int_0^{a/a_{\text{eq}}} \frac{(a/a_{\text{eq}})^2}{X_e \sqrt{1+(a/a_{\text{eq}})(1+R)}} \left( \frac{16}{15} + \frac{R^2}{1+R} \right) d(a/a_{\text{eq}}). \quad (31.88)$$

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 (31.88) can be done analytically. With  $a$  normalized to  $a_{\text{eq}} = 1$ ,

$$f(a) \equiv \int_0^a \frac{a^2}{\sqrt{1+a}} \frac{\frac{16}{15}(1+R) + R^2}{(1+R)^2} da \approx \int_0^a \frac{a^2 da}{\sqrt{1+a}}. \quad (31.89)$$

The last approximation is correct to order unity for any  $a$ . Conclude that, neglecting the effect of recombination on the electron fraction  $X_e$ ,

$$\frac{a_{\text{eq}}^2 H_{\text{eq}}^2}{c^2 k_d^2} = \frac{6.83 h^{-1}}{f_+ f_H} \frac{H_0}{H_{\text{eq}}} \frac{\Omega_m}{\Omega_b} f(a/a_{\text{eq}}) = 6 \times 10^{-4} f(a/a_{\text{eq}}) \approx 0.0035, \quad (31.90)$$

the final value being the approximate value at recombination.

### 31.11 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 31.7 you will show that, in an approximation developed in §33.6.2, 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 31.7 Generic behaviour of neutrinos.** Insert the approximate value (33.48) of the neutrino quadrupole  $\mathcal{N}_2$  into the neutrino energy and momentum conservation equations to obtain the differential equation

$$\left( \frac{d^2}{d\eta^2} + \frac{2}{\eta} \frac{d}{d\eta} + k^2 \right) (\mathcal{N}_0 - \Phi) = -k^2(\Psi + \Phi) . \quad (31.91)$$

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  $\mathcal{N}_0 - \Phi$  oscillates around  $-(\Psi + \Phi)$ .

**Solution.** The Green's function solution of equation (31.91) is with  $y \equiv k\eta$ ,

$$\mathcal{N}_0 - \Phi = \zeta_\nu \frac{\sin y}{y} - \int_0^y [\Psi(y') + \Phi(y')] \sin(y - y') \frac{y'}{y} dy' . \quad (31.92)$$

## 32

### Cosmological perturbations: Boltzmann treatment

Chapters 29 and 31 treated cosmological perturbations in the approximations that matter and radiation behaved as respectively perfect and imperfect fluids. The fluid approximation truncates the momentum dis-

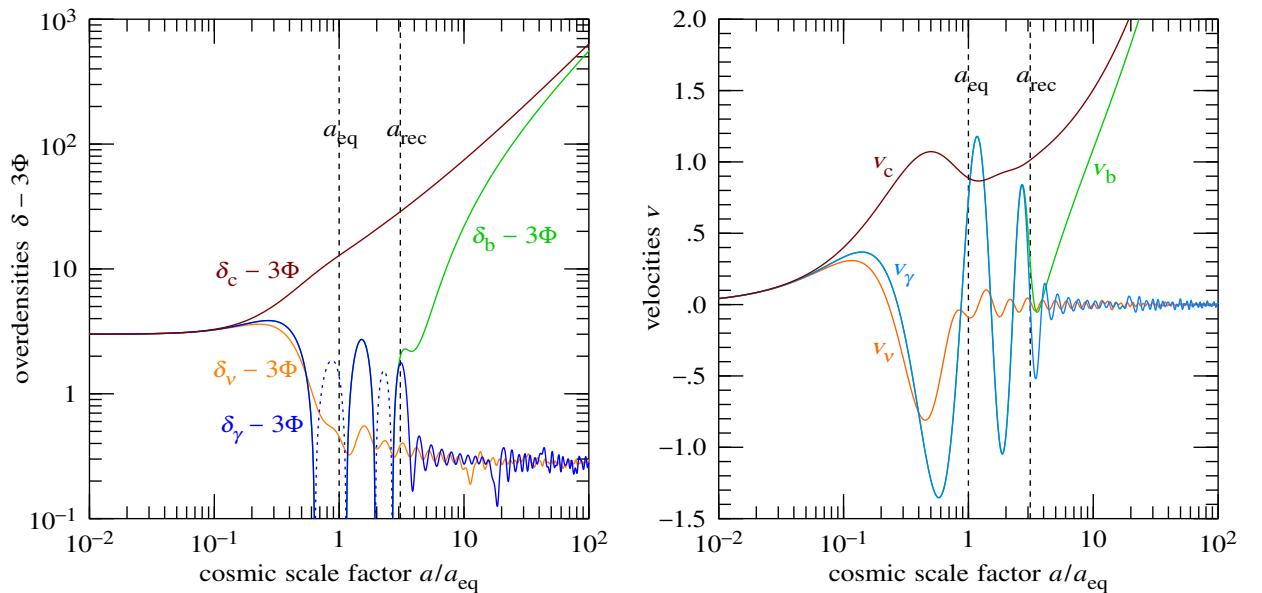


Figure 32.1 (Left) Overdensities  $\delta - 3\Phi$ , and (right) bulk velocities  $v$  in a Boltzmann treatment as a function of cosmic scale factor  $a/a_{\text{eq}}$ , at wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ , for non-baryonic dark matter (c), baryons (b), photons ( $\gamma$ ), and neutrinos ( $\nu$ ). The cosmological model is the standard model adopted in this book, a flat  $\Lambda$ CDM model with concordance parameters  $\Omega_\Lambda = 0.69$  and  $\Omega_m = 0.31$  and adiabatic initial conditions, §31.3. 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  $\ell_{\text{max}} = 31$ . Compare these results to the simple and hydrodynamic computations, Figures 29.2 and 31.1.

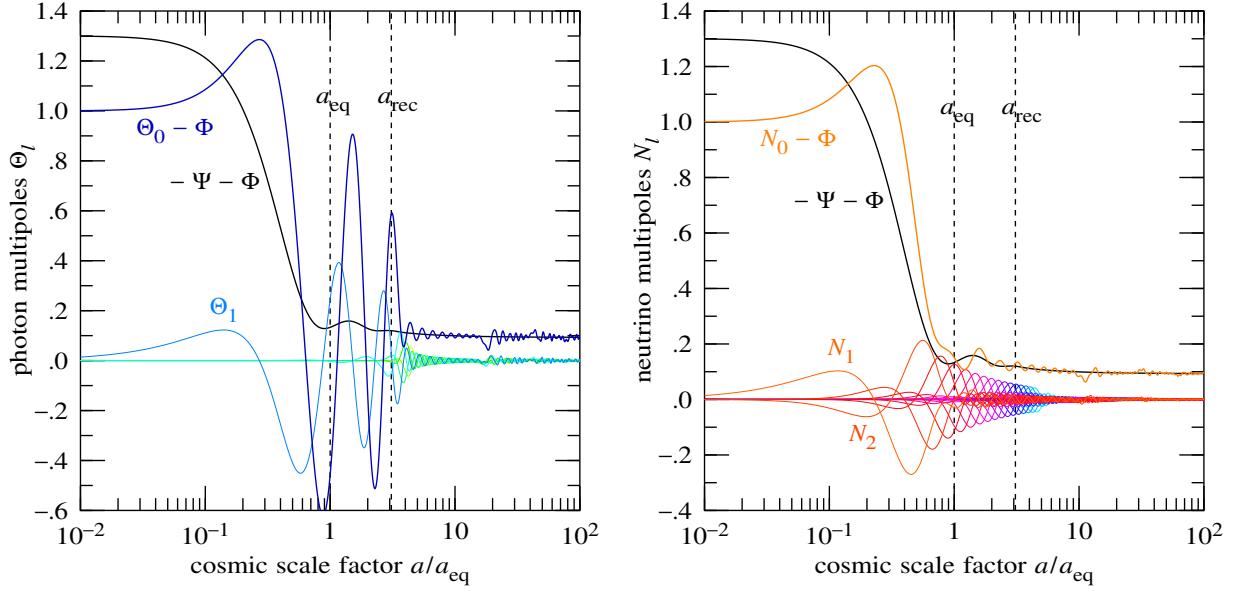


Figure 32.2 (Left) Photon multipoles up to  $\ell = 6$ , and (right) neutrino multipoles up to  $\ell = 31$ , as a function of cosmic scale factor  $a/a_{\text{eq}}$ , at wavenumber  $k/(a_{\text{eq}}H_{\text{eq}}) = 10$ . The cosmological model is the same as in Figure 32.1, §31.3. The thick (black) 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  $\ell_{\text{max}} = 31$ . 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 31.2.

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 31. The purpose of the present chapter is to give an account of the Boltzmann equation as it applies to cosmological perturbation theory.

Figure 32.1 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 32.2 shows photon multipoles up to  $\ell = 6$  and neutrino multipoles up to  $\ell = 31$  in a Boltzmann computation that include multipoles up to  $\ell_{\text{max}} = 31$  for both photons and neutrinos.

### 32.1 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  $\ell_{\max}$ . The hierarchy is truncated by invoking a suitable approximation for the  $(\ell_{\max}+1)$ 'th harmonic. The Boltzmann treatment yields the hydrodynamic approximation, §31.2, when  $\ell_{\max} = 1$ .

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 (31.6) and (31.7),

$$\dot{\delta}_c - k v_c - 3 \dot{\Phi} = 0 , \quad (32.1a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k \Psi = 0 , \quad (32.1b)$$

and

$$\dot{\delta}_b - k v_b - 3 \dot{\Phi} = 0 , \quad (32.2a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k \Psi = -\frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) . \quad (32.2b)$$

The equations for photons ( $\gamma$ ) are given by the Boltzmann hierarchy (32.81),

$$\dot{\Theta}_0 - k \Theta_1 - \dot{\Phi} = 0 , \quad (32.3a)$$

$$\dot{\Theta}_1 + \frac{k}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3} \Psi = \frac{1}{3} |\dot{\tau}| (v_b - 3\Theta_1) , \quad (32.3b)$$

$$\dot{\Theta}_2 + \frac{k}{5} (2\Theta_1 - 3\Theta_3) = -\frac{3}{4} |\dot{\tau}| \Theta_2 , \quad (32.3c)$$

$$\dot{\Theta}_\ell + \frac{k}{2\ell+1} [\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] = -|\dot{\tau}| \Theta_\ell \quad (\ell \geq 3) . \quad (32.3d)$$

As commented after equations (32.81), the factor  $\frac{3}{4}$  in equation (32.3c) includes the effect of polarization; without polarization, the factor is  $\frac{9}{10}$ . The  $(\ell_{\max}+1)$ 'th harmonic  $\Theta_{\ell_{\max}+1}$  may be approximated by an expression that interpolates between the large scattering limit  $|\dot{\tau}| \gg k_s$  and the free-streaming limit  $|\dot{\tau}| \ll k_s$ ,

$$\begin{aligned} \Theta_{\ell_{\max}+1} &= -\frac{1}{(1+|\dot{\tau}|/k_s)^2} \left[ \left(1 + \frac{1}{3}\delta_{\ell_{\max}1}\right) \frac{|\dot{\tau}|}{k_s} \frac{(\ell_{\max}+1)k}{(2\ell_{\max}+3)k_s} \Theta_{\ell_{\max}} + (\Theta_{\ell_{\max}-1} + \delta_{\ell_{\max}1}\Psi) + \frac{2\ell_{\max}+1}{k\eta} \Theta_{\ell_{\max}} \right] \\ &\rightarrow \begin{cases} -\frac{(1 + \frac{1}{3}\delta_{\ell_{\max}1})(\ell_{\max}+1)k}{(2\ell_{\max}+3)|\dot{\tau}|} & |\dot{\tau}| \gg k_s , \\ -(\Theta_0 + \Psi) - \frac{2\ell_{\max}+1}{k\eta} \Theta_1 & |\dot{\tau}| \ll k_s . \end{cases} \end{aligned} \quad (32.4)$$

Equation (32.4) reduces to the hydrodynamic approximation (31.9) when  $\ell_{\max} = 1$ . 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  $\Lambda$ CDM cosmological parameters). The equations for neutrinos ( $\nu$ ) are given by a Boltzmann hierarchy (32.89) which looks like that for photons, but without the scattering

terms,

$$\dot{\mathcal{N}}_0 - k\mathcal{N}_1 - \dot{\Phi} = 0 , \quad (32.5a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3}(\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3}\Psi = 0 , \quad (32.5b)$$

$$\dot{\mathcal{N}}_\ell + \frac{k}{2\ell+1}[\ell\mathcal{N}_{\ell-1} - (\ell+1)\mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) . \quad (32.5c)$$

The  $(\ell_{\max}+1)$ 'th harmonic  $\mathcal{N}_{\ell_{\max}+1}$  may be approximated by, equation (32.90),

$$\mathcal{N}_{\ell_{\max}+1} = -(\mathcal{N}_{\ell_{\max}-1} + \delta_{\ell_{\max}1}\Psi) - \frac{2\ell_{\max}+1}{k\eta}\mathcal{N}_{\ell_{\max}} . \quad (32.6)$$

The Einstein energy and quadrupole pressure equations are

$$-k^2\Phi - 3\frac{\dot{a}}{a}F = 4\pi Ga^2(\bar{\rho}_c\delta_c + \bar{\rho}_b\delta_b + 4\bar{\rho}_\gamma\Theta_0 + 4\bar{\rho}_\nu\mathcal{N}_0) , \quad (32.7a)$$

$$k^2(\Psi - \Phi) = -32\pi Ga^2(\bar{\rho}_\gamma\Theta_2 + \bar{\rho}_\nu\mathcal{N}_2) , \quad (32.7b)$$

where  $F$  is defined by equation (29.53).

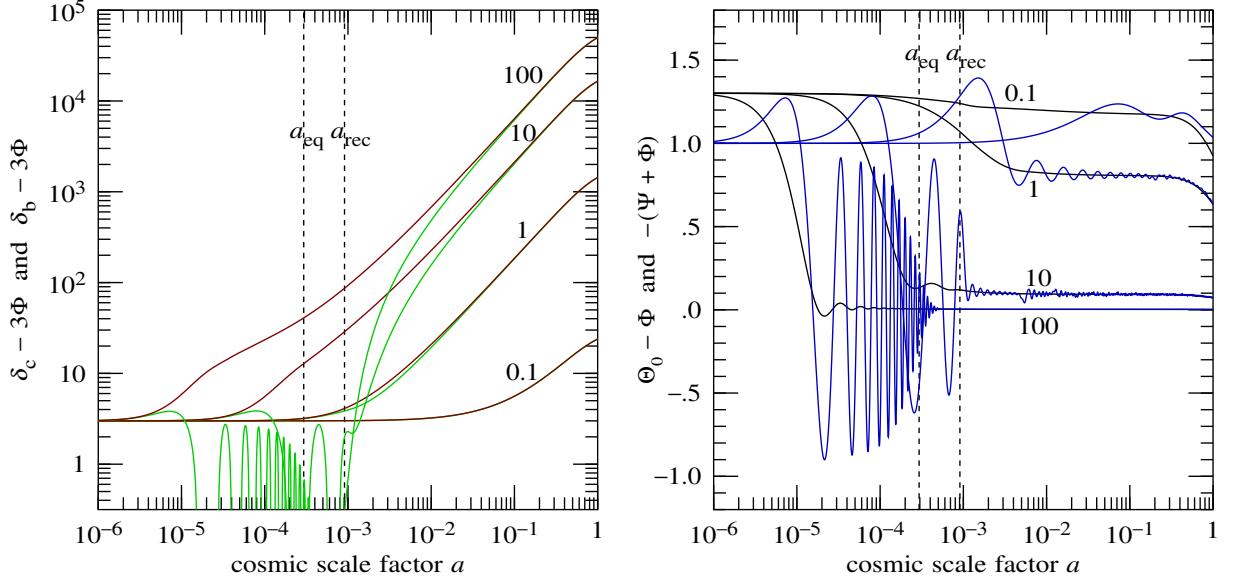


Figure 32.3 (Left) Overdensities  $\delta_c - 3\Phi$  and  $\delta_b - 3\Phi$  of non-baryonic dark matter (brown) and baryonic matter (green), and (right) radiation monopole  $\Theta_0 - 3\Phi$  (blue), and minus the sum of the scalar potentials,  $-(\Psi + \Phi)$  (black), as a function of cosmic scale 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 §31.3. Compare these results to the simple and hydrodynamic computations, Figures 29.1 and 31.3.

**Exercise 32.1 Program the Boltzmann equations.** Upgrade the code you wrote in Exercise 31.2 to implement the system of Boltzmann equations (31.6)–(31.13). Initial conditions for neutrinos, and for the two scalar potentials  $\Psi$  and  $\Phi$ , are derived in Exercise 32.5. 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 32.1, 32.2, 32.3 and 32.4. The computations here included photon and neutrino multipoles up to  $\ell_{\max} = 31$ .

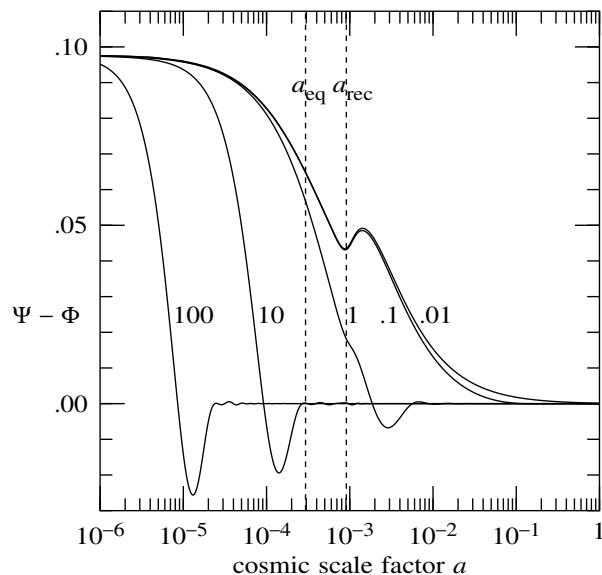


Figure 32.4 Difference  $\Psi - \Phi$  in scalar potentials as a function of cosmic scale factor  $a$ . The cosmological model is the same as in Figure 32.1, §31.3. Curves are labelled with the wavenumber  $k/(a_{\text{eq}}H_{\text{eq}})$  in units of the Hubble distance at matter-radiation equality. 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  $\ell_{\max} = 31$ .

**Exercise 32.2 Power spectrum of matter fluctuations: Boltzmann treatment.** Upgrade the code you wrote in Exercise 31.3 to compute the power spectrum of matter fluctuations in a Boltzmann computation.

**Solution.** See Figure 32.5.

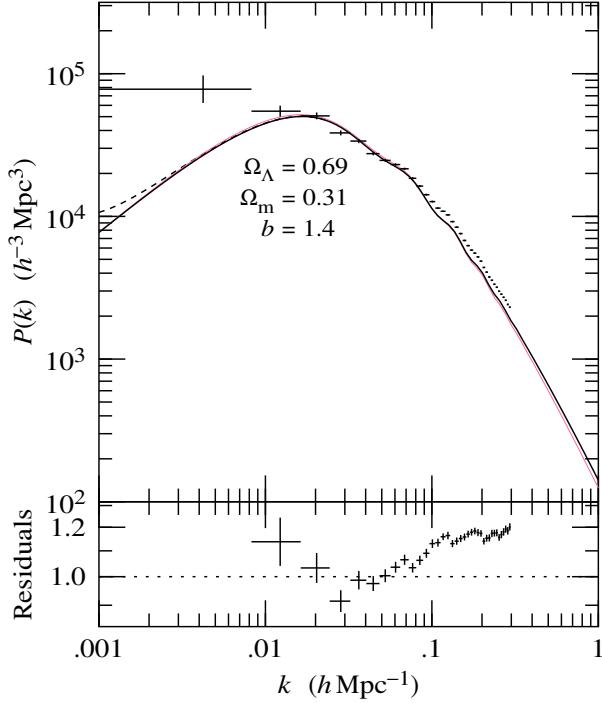


Figure 32.5 Model matter power spectrum computed from a Boltzmann computation, compared to observations from the BOSS galaxy survey (Anderson, 2014). The cosmological model is the same as in Figure 32.1, §31.3. The model power spectrum may be compared to those computed in the simple and hydrodynamic approximations, Figure 29.15 and Figure 31.4. The thin (pink) line is the model power spectrum in the hydrodynamic approximation.

### 32.2 Boltzmann equation in a perturbed FLRW geometry

The Boltzmann equation was introduced in §30.5. The left hand side of the Boltzmann equation (30.32) is, for either massless or massive particles,

$$\frac{df}{d\lambda} = p^m \partial_m f + \frac{dp^a}{d\lambda} \frac{\partial f}{\partial p^a} = E \partial_0 f + p^a \partial_a f + \frac{d\hat{\mathbf{p}}}{d\lambda} \cdot \frac{\partial f}{\partial \hat{\mathbf{p}}} + \frac{dp}{d\lambda} \frac{\partial f}{\partial p}. \quad (32.8)$$

Here  $\lambda$  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

$$\frac{df}{d\lambda} = E \partial_0 f + p^a \partial_a f + \frac{dp}{d\lambda} \frac{\partial f}{\partial p}. \quad (32.9)$$

The expression (32.9) 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

$$\frac{dp^k}{d\lambda} + \Gamma_{mn}^k p^m p^n = 0 . \quad (32.10)$$

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

$$p \frac{dp}{d\lambda} = E \frac{dE}{d\lambda} . \quad (32.11)$$

The equation of motion for the tetrad-frame energy  $E \equiv p^0$  is

$$\frac{dE}{d\lambda} = -\Gamma_{mn}^0 p^m p^n = \Gamma_{0ao} p^a E + \Gamma_{0ab} p^a p^b . \quad (32.12)$$

From this it follows that

$$\frac{d \ln p}{d\lambda} = \frac{E}{p^2} \frac{dE}{d\lambda} = E \left( \frac{E \hat{p}^a}{p} \Gamma_{0ao} + \hat{p}^a \hat{p}^b \Gamma_{0ab} \right) = E \left( -\frac{\dot{a}}{a^2} + \frac{E \hat{p}^a}{p} \Gamma_{0ao} + \hat{p}^a \hat{p}^b \Gamma_{0ab}^1 \right) , \quad (32.13)$$

where in the last expression the tetrad connection  $\Gamma_{0ab}$ , equation (28.24b), has been separated into its unperturbed and perturbed parts  $-(\dot{a}/a^2)\delta_{ab}$  and  $\Gamma_{0ab}^1$ .

In practice, the integration variable used to evolve equations is the conformal time  $\eta$ , not the affine parameter  $\lambda$ . The relation between conformal time  $\eta$  and affine parameter  $\lambda$  is

$$\frac{d\eta}{d\lambda} = p^{\eta} = e_m^{\eta} p^m = (\delta_m^n + \varphi_m{}^n) e_n^{\eta} p^m = \frac{1}{a} [E(1 - \varphi_{00}) - p^a \varphi_{a0}] , \quad (32.14)$$

whose reciprocal is to linear order

$$\frac{d\lambda}{d\eta} = \frac{a}{E} \left( 1 + \varphi_{00} + \frac{p^a}{E} \varphi_{a0} \right) . \quad (32.15)$$

With conformal time  $\eta$  as the integration variable, the equation of motion (32.13) for the magnitude  $p$  of the tetrad-frame momentum becomes, to linear order,

$$\frac{d \ln p}{d\eta} = -\frac{\dot{a}}{a} \left( 1 + \varphi_{00} + \frac{p^a}{E} \varphi_{a0} \right) + \frac{E \hat{p}^a}{p} a \Gamma_{0ao} + \hat{p}^a \hat{p}^b a \Gamma_{0ab}^1 . \quad (32.16)$$

With the collision term restored, the Boltzmann equation (30.32) expressed with respect to conformal time  $\eta$  is

$$\frac{df}{d\eta} = \frac{\partial f}{\partial \eta} + v^a \nabla_a f + \frac{d \ln p}{d\eta} \frac{\partial f}{\partial \ln p} = \frac{d\lambda}{d\eta} C[f] , \quad (32.17)$$

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 (32.15)

and (32.16). 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 32.3. In conformal Newtonian gauge, the factor  $d\lambda/d\eta$ , equation (32.15), is

$$\frac{d\lambda}{d\eta} = \frac{a}{E}(1 + \Psi) . \quad (32.18)$$

In conformal Newtonian gauge, and including only scalar fluctuations, the factor  $d\ln p/d\eta$ , equation (32.16), is

$$\frac{d\ln p}{d\eta} = -\frac{\dot{a}}{a} + \dot{\Phi} - \frac{E\hat{p}^a}{p}\nabla_a\Psi . \quad (32.19)$$

To unperturbed order, the Boltzmann equation (32.17) is

$$\frac{d\overset{0}{f}}{d\eta} = \frac{\partial\overset{0}{f}}{\partial\eta} - \frac{\dot{a}}{a}\frac{\partial\overset{0}{f}}{\partial\ln p} = \frac{a}{E}C[\overset{0}{f}] , \quad (32.20)$$

where  $C[\overset{0}{f}]$  is the unperturbed collision term, the factor  $a/E$  coming from  $d\lambda/d\eta = a/E$  to unperturbed order, equation (32.15). The second term in the middle expression of equation (32.20) 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 (10.66).

Subtracting off the unperturbed part (32.20) of the Boltzmann equation (32.17) gives the perturbation of the Boltzmann equation

$$\boxed{\frac{d\overset{1}{f}}{d\eta} = \frac{\partial\overset{1}{f}}{\partial\eta} + v^a\nabla_a\overset{1}{f} - \frac{\dot{a}}{a}\frac{\partial\overset{1}{f}}{\partial\ln p} + G\frac{\partial\overset{0}{f}}{\partial\ln p} = \frac{a}{E}C[\overset{1}{f}] + \frac{d\lambda}{d\eta}C[\overset{0}{f}]} , \quad (32.21)$$

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 (32.16), and  $G$  defined by

$$G \equiv \frac{d\ln(ap)}{d\eta} \quad (32.22)$$

expresses the peculiar gravitational redshifting of particles. In conformal Newtonian gauge, and including only scalar fluctuations, the gravitational redshift term  $G$  is

$$G = \dot{\Phi} - \frac{E\hat{p}^a}{p}\nabla_a\Psi . \quad (32.23)$$

In conformal Newtonian gauge, the perturbed part of  $d\lambda/d\eta$  which appears on the right hand side of the Boltzmann equation (32.21) is

$$\frac{d\lambda}{d\eta} = \frac{a}{E}\Psi . \quad (32.24)$$

**Exercise 32.3 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 (32.15) is

$$\frac{d\lambda}{d\eta} = \frac{a}{E} \left[ 1 + \psi + \frac{p^a}{E} (\nabla_a \tilde{w} + \tilde{w}_a) \right] , \quad (32.25)$$

while equation (32.16) is

$$\begin{aligned} \frac{d \ln ap}{d\eta} &= \dot{\phi} + \frac{E \hat{p}^a}{p} \left[ -\nabla_a \psi + \left( \frac{\partial}{\partial \eta} + \frac{\dot{a}}{a} \frac{m^2}{E^2} \right) (\nabla_a \tilde{w} + \tilde{w}_a) \right] \\ &\quad + \hat{p}^a \hat{p}^b \left[ -\nabla_a \nabla_b (w - \dot{h}) - \frac{1}{2} (\nabla_a W_b + \nabla_b W_a) + \nabla_b \tilde{w}_a + \dot{h}_{ab} \right] . \end{aligned} \quad (32.26)$$

### 32.3 Non-baryonic cold dark matter

Non-baryonic cold dark matter is by assumption non-relativistic and collisionless. The unperturbed mean density is  $\bar{\rho}_c$ , which evolves with cosmic scale factor  $a$  as

$$\bar{\rho}_c \propto a^{-3} . \quad (32.27)$$

Since dark matter particles are non-relativistic, the energy of a dark matter particle is its rest-mass energy,  $E_c = m_c$ , and its momentum is the non-relativistic 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 (10.118). The energy and momentum moments of the distribution define the dark matter overdensity  $\delta_c$  and bulk velocity  $\mathbf{v}_c$ , 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  $\mathbf{v}$ ),

$$T_c^{00} \equiv \int f_c m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \equiv \bar{\rho}_c (1 + \delta_c) , \quad (32.28a)$$

$$T_c^{0a} \equiv \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \equiv \bar{\rho}_c v_c^a , \quad (32.28b)$$

$$T_c^{ab} \equiv \int f_c m_c v_c^a v_c^b \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = 0 . \quad (32.28c)$$

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

$$\frac{df_c}{d\eta} = 0 . \quad (32.29)$$

The energy and momentum moments of the Boltzmann equation (32.17) yield equations for the overdensity

$\delta_c$  and bulk velocity  $v_c$ , which in the conformal Newtonian gauge are

$$\begin{aligned} 0 &= \int \frac{df_c}{d\eta} m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = \frac{\partial}{\partial \eta} \int f_c m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} + \nabla_a \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} - \int \left( \frac{\dot{a}}{a} - \dot{\Phi} \right) \frac{\partial f}{\partial \ln p} m_c \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &= \frac{\partial \bar{\rho}_c (1 + \delta_c)}{\partial \eta} + \nabla_a (\bar{\rho}_c v_c^a) + 3 \left( \frac{\dot{a}}{a} - \dot{\Phi} \right) \bar{\rho}_c , \end{aligned} \quad (32.30a)$$

$$\begin{aligned} 0 &= \int \frac{df_c}{d\eta} m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} = \frac{\partial}{\partial \eta} \int f_c m_c v_c^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} + \nabla_b \int f_c m_c v_c^a v_c^b \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &\quad - \int \left( \frac{\dot{a}}{a} - \dot{\Phi} + \frac{E \dot{p}^b}{p} \nabla_b \Psi \right) \frac{\partial f}{\partial \ln p} m_c v^a \frac{g_c d^3 p_c}{(2\pi\hbar)^3} \\ &= \frac{\partial \bar{\rho}_c v_c^a}{\partial \eta} + 4 \left( \frac{\dot{a}}{a} - \dot{\Phi} \right) \bar{\rho}_c v_c^a + \bar{\rho}_c \nabla_a \Psi . \end{aligned} \quad (32.30b)$$

The  $\dot{\Phi} \bar{\rho}_c v_c^a$  term on the last line of equation (32.30b) can be dropped, since the potential  $\Phi$  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 (32.30a) and (32.30b) gives equations for the dark matter overdensity  $\delta_c$  and bulk velocity  $v_c$ ,

$$\dot{\delta}_c + \nabla \cdot \mathbf{v}_c - 3\dot{\Phi} = 0 , \quad (32.31a)$$

$$\dot{\mathbf{v}}_c + \frac{\dot{a}}{a} \mathbf{v}_c + \nabla \Psi = 0 . \quad (32.31b)$$

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

$$\mathbf{v}_c = -i\hat{\mathbf{k}}v_c + \mathbf{v}_{c,\perp} . \quad (32.32)$$

For the scalar modes under consideration, only the scalar part of the dark matter equations (32.31) is relevant:

$$\dot{\delta}_c - kv_c - 3\dot{\Phi} = 0 , \quad (32.33a)$$

$$\dot{v}_c + \frac{\dot{a}}{a} v_c + k\Psi = 0 . \quad (32.33b)$$

Equations (32.33) reproduce the equations (29.50) derived previously from conservation of energy and momentum.

**Exercise 32.4 Moments of the non-baryonic cold dark matter Boltzmann equation.** Confirm equations (32.30).

### 32.4 Boltzmann equation for the temperature fluctuation

The full hierarchy of Boltzmann equations is needed only to describe relativistic species (photons and neutrinos); non-relativistic 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  $\Theta$ ,

$$\Theta(\eta, \mathbf{x}, \mathbf{p}) \equiv \frac{\delta T(\eta, \mathbf{x}, \mathbf{p})}{T(\eta)} . \quad (32.34)$$

In most of this book  $\Theta$  refers to photons, but in this section the temperature fluctuation  $\Theta$  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  $\Theta$ , and not the perturbation  $\dot{f}$  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  $\Theta$  to become anisotropic.

For a relativistic species, the unperturbed occupation number in thermodynamic equilibrium is

$$\overset{0}{f} = \frac{1}{e^{p/T} \mp 1} , \quad (32.35)$$

where the sign is  $-$  for bosons (photons) and  $+$  for fermions (neutrinos). The unperturbed Boltzmann equation (32.20) can be recast as an equation for the background temperature  $T(\eta)$ ,

$$\frac{d \ln(aT)}{d\eta} = \frac{a}{E} C[\overset{0}{f}] / \frac{\partial \overset{0}{f}}{\partial \ln T} , \quad (32.36)$$

where it follows from equation (32.35) that (the partial derivative with respect to temperature  $\partial/\partial \ln T$  is at constant momentum  $p$ )

$$\frac{\partial \overset{0}{f}}{\partial \ln T} = \overset{0}{f}(1 \pm \overset{0}{f}) \frac{p}{T} , \quad (32.37)$$

with  $+$  and  $-$  for bosons and fermions respectively. Equation (32.36) shows that if the collision term  $C[\overset{0}{f}]$  vanishes, then the background temperature redshifts as  $T \propto a^{-1}$ . In practice, the collision term  $C[\overset{0}{f}]$  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 (10.101). 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 (32.34) of the temperature fluctuation  $\Theta$  is to be interpreted as meaning that the perturbation to the occupation number is

$$\overset{1}{f} = \frac{\partial \overset{0}{f}}{\partial \ln T} \delta \ln T = \frac{\partial \overset{0}{f}}{\partial \ln T} \Theta . \quad (32.38)$$

Two of the terms on the left hand side of the perturbed Boltzmann equation (32.21) rearrange to

$$\begin{aligned} \frac{\partial \overset{1}{f}}{\partial \eta} + \frac{d \ln p}{d \eta} \frac{\partial \overset{1}{f}}{\partial \ln p} &= \frac{\partial \overset{0}{f}}{\partial \ln T} \dot{\Theta} + \left[ \left( \frac{d \ln T}{d \eta} \frac{\partial}{\partial \ln T} - \frac{\dot{a}}{a} \frac{\partial}{\partial \ln p} \right) \frac{\partial \overset{0}{f}}{\partial \ln T} \right] \Theta \\ &= \frac{\partial \overset{0}{f}}{\partial \ln T} \left[ \dot{\Theta} + \frac{d \ln(aT)}{d \eta} \frac{\partial \ln(\partial \overset{0}{f}/\partial \ln T)}{\partial \ln T} \Theta \right]. \end{aligned} \quad (32.39)$$

The collision terms on the right hand side of the perturbed Boltzmann equation (32.21) are

$$\frac{a}{E} C[\overset{1}{f}] + \frac{d\lambda}{d\eta} C[\overset{0}{f}] = \frac{\partial \overset{0}{f}}{\partial \ln T} \left[ C[\Theta] + \frac{d \ln(aT)}{d \eta} \frac{E}{a} \frac{d\lambda}{d\eta} \right], \quad (32.40)$$

where  $C[\overset{0}{f}]$  has been eliminated in favour of  $d \ln(aT)/d\eta$  using equation (32.36), and  $C[\Theta]$  is the scaled collision term defined by

$$C[\Theta] \equiv \frac{a}{E} C[\overset{1}{f}] / \frac{\partial \overset{0}{f}}{\partial \ln T}. \quad (32.41)$$

The perturbed Boltzmann equation (32.21) thus becomes

$$\frac{d\Theta}{d\eta} = \frac{\partial \Theta}{\partial \eta} + v^a \nabla_a \Theta - G + \frac{d \ln(aT)}{d \eta} \frac{\partial \ln(\partial \overset{0}{f}/\partial \ln T)}{\partial \ln T} \Theta = C[\Theta] + \frac{d \ln(aT)}{d \eta} \frac{E}{a} \frac{d\lambda}{d\eta}, \quad (32.42)$$

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 term  $d \ln(aT)/d\eta$  is negligible for both photons and neutrinos during the time before and through recombination when anisotropies in the CMB are developing. The Boltzmann equation (32.42) then reduces to

$$\frac{d\Theta}{d\eta} = \frac{\partial \Theta}{\partial \eta} + v^a \nabla_a \Theta - G = C[\Theta]. \quad (32.43)$$

As long as the particles are relativistic, the particle velocity is one,  $v = 1$ ; but equation (32.43) allows for a general non-unit velocity  $v$  to accommodate neutrinos, which have small masses.

Fourier transforming the Boltzmann equation (32.43) over spatial position  $\mathbf{x}$  yields the Boltzmann equation for the Fourier components  $\Theta(\eta, \mathbf{k}, \mathbf{p})$  of the temperature fluctuation,

$$\frac{d\Theta}{d\eta} = \frac{\partial \Theta}{\partial \eta} - ivk\mu\Theta - G = C[\Theta], \quad (32.44)$$

where  $\mu \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}$ . In Fourier space, the gravitational redshift term  $G$ , in conformal Newtonian gauge and including only scalar fluctuations, is, equation (32.23),

$$G = \dot{\Phi} + \frac{ik\mu}{v} \Psi. \quad (32.45)$$

### 32.5 Spherical harmonics of the temperature fluctuation

It is natural to expand the (photon or neutrino) temperature fluctuation  $\Theta$  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  $\Theta$  can be written

$$\begin{aligned}\Theta(\eta, \mathbf{k}, \mathbf{p}) &= \sum_{\ell=0}^{\infty} (-i)^{\ell} \sqrt{4\pi(2\ell+1)} \Theta_{\ell}(\eta, \mathbf{k}, p) Y_{\ell 0}(\hat{\mathbf{p}}) \\ &= \sum_{\ell=0}^{\infty} (-i)^{\ell} (2\ell+1) \Theta_{\ell}(\eta, \mathbf{k}, p) P_{\ell}(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}}),\end{aligned}\quad (32.46)$$

where  $P_{\ell}$  are Legendre polynomials, §32.15. The choice of normalization of the scalar harmonics  $\Theta_{\ell}$  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  $\Theta_{\ell}$  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  $\Theta_{\ell}$  are angular integrals of the temperature fluctuation  $\Theta$  over momentum directions  $\hat{\mathbf{p}}$ ,

$$\Theta_{\ell}(\eta, \mathbf{k}, p) = i^{\ell} \int \Theta(\eta, \mathbf{k}, \mathbf{p}) P_{\ell}(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}}) \frac{d\sigma_{\mathbf{p}}}{4\pi}.\quad (32.47)$$

Expanded into the scalar harmonics  $\Theta_{\ell}(\eta, \mathbf{k}, p)$ , equation (32.46), the left hand side of the Boltzmann equation (32.44) in conformal Newtonian gauge is

$$\frac{d\Theta_0}{d\eta} = \dot{\Theta}_0 - vk\Theta_1 - \dot{\Phi},\quad (32.48a)$$

$$\frac{d\Theta_1}{d\eta} = \dot{\Theta}_1 + \frac{vk}{3} (\Theta_0 - 2\Theta_2) + \frac{k}{3v} \Psi,\quad (32.48b)$$

$$\frac{d\Theta_{\ell}}{d\eta} = \dot{\Theta}_{\ell} + \frac{vk}{2\ell+1} [\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] \quad (\ell \geq 2).\quad (32.48c)$$

### 32.6 The Boltzmann equation for massless particles

The Boltzmann equation (32.44) and its harmonic expansion (32.48) are valid for massive as well as massless particles, to allow for neutrino masses. The case of massive neutrinos will be resumed in §32.14; 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  $\Theta$  is equation (32.43) with  $v = 1$ . The left hand side of the Boltzmann equation expands in scalar harmonics  $\Theta_\ell$  as equations (32.48) again with  $v = 1$ .

As remarked at the beginning of §32.4, at early times photons and neutrinos have distributions in thermodynamic equilibrium, as a result of which their initial temperature fluctuations  $\Theta$  are independent of particle momentum  $\mathbf{p}$ . For massless particles ( $v = 1$ ) the left hand side of the Boltzmann equation (32.43) is independent of the magnitude  $p$  of the particle momentum. As will be seen in §32.8, equation (32.67), Thomson scattering leaves the magnitude  $p_\gamma$  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,

$$\Theta(\eta, \mathbf{k}, \mathbf{p}) = \Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}) \quad \text{for photons and relativistic neutrinos .} \quad (32.49)$$

### 32.7 Energy-momentum tensor for massless particles

Perturbations  $\overset{1}{T}{}^{kl}$  to the energy-momentum tensor of particles involve integrals (10.118) 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  $\hat{\mathbf{p}}$ ,

$$\int \overset{1}{f} p^2 F(\hat{\mathbf{p}}) \frac{g d^3 p}{p(2\pi\hbar)^3} = \int \frac{\partial \overset{0}{f}}{\partial \ln T} p^2 \frac{g 4\pi p^2 dp}{p(2\pi\hbar)^3} \int \Theta F(\hat{\mathbf{p}}) \frac{d\omega_{\mathbf{p}}}{4\pi} = 4\bar{\rho} \int \Theta F(\hat{\mathbf{p}}) \frac{d\omega_{\mathbf{p}}}{4\pi} , \quad (32.50)$$

in which the last expression is true because

$$\int \frac{\partial \overset{0}{f}}{\partial \ln T} p^2 \frac{g 4\pi p^2 dp}{p(2\pi\hbar)^3} = 4 \int \overset{0}{f} p \frac{g 4\pi p^2 dp}{(2\pi\hbar)^3} = 4\bar{\rho} , \quad (32.51)$$

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

$$\overset{1}{T}{}^{00} = 4\bar{\rho}\Theta_0 , \quad (32.52a)$$

$$\hat{k}_a \overset{1}{T}{}^{0a} = -i 4\bar{\rho}\Theta_1 , \quad (32.52b)$$

$$\frac{1}{3} \delta_{ab} \overset{1}{T}{}^{ab} = \frac{4}{3} \bar{\rho}\Theta_0 , \quad (32.52c)$$

$$\left( \frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab} \right) \overset{1}{T}{}^{ab} = -4\bar{\rho}\Theta_2 . \quad (32.52d)$$

### 32.8 Nonrelativistic electron-photon (Thomson) scattering

The dominant process that couples photons and baryons is electron-photon scattering

$$e + \gamma \leftrightarrow e' + \gamma' . \quad (32.53)$$

The Lorentz-invariant amplitude squared for unpolarized non-relativistic 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  $c = \hbar = 1$  (Peskin and Schroeder, 1995, p. 162)

$$|\mathcal{M}|^2 = (8\pi\alpha)^2 \frac{1 + (\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})^2}{2}, \quad (32.54)$$

where  $\alpha \equiv e^2/(\hbar c)$  is the fine-structure constant. The unpolarized invariant amplitude squared (32.54) is the polarized amplitude squared (34.36) averaged over polarization states of the incoming photon, and summed over polarization states of the scattered photon. The differential cross-section  $d\sigma_T/do'$  for unpolarized Thomson scattering into an interval  $do'$  of solid angle about scattered photon direction  $\hat{\mathbf{p}}'$  is related to the squared invariant amplitude  $|\mathcal{M}|^2$  by

$$\frac{d\sigma_T}{do'} = \frac{|\mathcal{M}|^2}{(8\pi m_e)^2} = \frac{\alpha^2}{m_e^2} \frac{1 + (\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})^2}{2}. \quad (32.55)$$

The coefficient of the differential cross-section is, with units  $c$  and  $\hbar$  restored,

$$\left(\frac{\alpha\hbar}{m_e c}\right)^2 = r_e^2 = \left(\frac{e^2}{m_e c^2}\right)^2, \quad (32.56)$$

where  $r_e \equiv e^2/m_e c^2$  is the classical electron radius. The total Thomson cross-section  $\sigma_T$  is

$$\sigma_T \equiv \int \frac{d\sigma_T}{do'} do' = \frac{8\pi}{3} r_e^2. \quad (32.57)$$

### 32.9 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 30.6,

$$C[\overset{0}{f}_\gamma] = 0. \quad (32.58)$$

Thanks to detailed balance, the combination of rates in the collision integral (30.40) 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 (30.40). To unperturbed order, the energies of the electrons, which are non-relativistic, 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 (30.40),

$$C[\overset{1}{f}_\gamma] = \frac{1}{16} \int |\mathcal{M}|^2 [-f_e f_\gamma (1 + f_{\gamma'}) + f_{e'} f_{\gamma'} (1 + f_\gamma)] (2\pi)^4 \delta_D^4(p_e + p_\gamma - p_{e'} - p_{\gamma'}) \frac{2 d^3 p_e}{m_e (2\pi)^3} \frac{d^3 p_{e'}}{m_e (2\pi)^3} \frac{d^3 p_{\gamma'}}{p_{\gamma'} (2\pi)^3}. \quad (32.59)$$

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  $\gamma'$ , which is scattered into an interval  $do_{\gamma'}$  of solid angle:

$$\int 2\pi \delta_D(E_e + E_\gamma - E_{e'} - E_{\gamma'}) \frac{d^3 p_{\gamma'}}{E_{\gamma'} (2\pi)^3} = p_{\gamma'} \frac{do_{\gamma'}}{(2\pi)^2} \approx p_\gamma \frac{do_{\gamma'}}{(2\pi)^2}. \quad (32.60)$$

The approximation in the last step of equation (32.60), replacing the energy  $p_{\gamma'}$  of the scattered photon by the energy  $p_\gamma$  of the incoming photon, is valid because, thanks to the smallness of the combination of rates in the collision integral (32.59), it suffices to treat the photon energy to unperturbed order. As seen below, equation (32.65), 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 ( $e + \gamma \rightarrow e' + \gamma'$ ),

$$\int (2\pi)^3 \delta_D^3(\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{e'} - \mathbf{p}_{\gamma'}) \frac{d^3 p_{e'}}{m_e (2\pi)^3} = \frac{1}{m_e}, \quad (32.61)$$

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 (32.59) becomes

$$-f_e f_\gamma (1 + f_{\gamma'}) + f_e f_{\gamma'} (1 + f_\gamma) = f_e (-f_\gamma + f_{\gamma'}). \quad (32.62)$$

Notice that the stimulated terms cancel. The energy- and momentum-conserving integrations (32.60) and (32.61) bring the photon collision term (32.59) to

$$C[\overset{1}{f}_\gamma] = \frac{p_\gamma}{16\pi m_e^2} \int |\mathcal{M}|^2 f_e (-f_\gamma + f_{\gamma'}) \frac{2 d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi}. \quad (32.63)$$

The collision integral (32.63) involves the difference  $-f_\gamma + f_{\gamma'}$  between the occupancy of the initial and final photon states. To linear order, the difference is

$$-f_\gamma + f_{\gamma'} = -\overset{0}{f}(\mathbf{p}_\gamma) + \overset{0}{f}(\mathbf{p}_{\gamma'}) - \overset{1}{f}(\mathbf{p}_\gamma) + \overset{1}{f}(\mathbf{p}_{\gamma'}) = \frac{\partial \overset{0}{f}_\gamma}{\partial \ln T} \left[ \frac{p_\gamma - p_{\gamma'}}{p_\gamma} - \Theta(\mathbf{p}_\gamma) + \Theta(\mathbf{p}_{\gamma'}) \right]. \quad (32.64)$$

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:

$$\begin{aligned}
p_\gamma - p_{\gamma'} &= E_{e'} - E_e \\
&= \left( m_e + \frac{p_e'^2}{2m_e} \right) - \left( m_e + \frac{p_e^2}{2m_e} \right) \\
&= \frac{(\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{\gamma'})^2 - p_e^2}{2m_e} \\
&= \frac{(\mathbf{p}_\gamma - \mathbf{p}_{\gamma'}) \cdot (2\mathbf{p}_e + \mathbf{p}_\gamma - \mathbf{p}_{\gamma'})}{2m_e} \\
&\approx (\mathbf{p}_\gamma - \mathbf{p}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e}, \tag{32.65}
\end{aligned}$$

the last line of which follows because the photon momentum is small compared to the electron momentum,

$$p_\gamma \sim T \sim \frac{p_e^2}{m_e} \ll p_e. \tag{32.66}$$

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  $\Theta$  as being a function only of the direction  $\hat{\mathbf{p}}_\gamma$  of the photon momentum, not of its energy:

$$\Theta(\mathbf{p}_\gamma) \approx \Theta(\hat{\mathbf{p}}_\gamma). \tag{32.67}$$

The linear approximations (32.65) and (32.67) bring the difference (32.64) between the initial and final photon occupancies to

$$-f_\gamma + f_{\gamma'} = \frac{\partial \overset{\circ}{f}_\gamma}{\partial \ln T} \left[ (\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right]. \tag{32.68}$$

Inserting this difference in occupancies into the collision integral (32.63) yields

$$C[\overset{\circ}{f}_\gamma] = \frac{p_\gamma}{16\pi m_e^2} \frac{\partial \overset{\circ}{f}_\gamma}{\partial \ln T} \int |\mathcal{M}|^2 f_e \left[ (\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right] \frac{2d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi}, \tag{32.69}$$

or, switching to  $C[\Theta]$  defined by equation (32.41),

$$C[\Theta] = \frac{a}{16\pi m_e^2} \int |\mathcal{M}|^2 f_e \left[ (\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \frac{\mathbf{p}_e}{m_e} - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'}) \right] \frac{2d^3 p_e}{(2\pi)^3} \frac{do_{\gamma'}}{4\pi}. \tag{32.70}$$

The invariant amplitude squared  $|\mathcal{M}|^2$ , equation (32.54), is independent of electron momenta, so the integration over electron momentum in the collision integral (32.70) is straightforward. The unperturbed electron density  $\bar{n}_e$  and the electron bulk velocity  $\mathbf{v}_e$  are defined by

$$\bar{n}_e \equiv \int \overset{\circ}{f}_e \frac{2d^3 p_e}{(2\pi)^3}, \quad \bar{n}_e \mathbf{v}_e \equiv \int \overset{\circ}{f}_e \frac{\mathbf{p}_e}{m_e} \frac{2d^3 p_e}{(2\pi)^3}. \tag{32.71}$$

Coulomb scattering keeps electrons and ions tightly coupled, so the electron bulk velocity  $\mathbf{v}_e$  equals the baryon bulk velocity  $\mathbf{v}_b$ ,

$$\mathbf{v}_e = \mathbf{v}_b . \quad (32.72)$$

Integration over the electron momentum brings the collision integral (32.70) to

$$C[\Theta] = \frac{\bar{n}_e a}{16\pi m_e^2} \int |\mathcal{M}|^2 [(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'})] \frac{d\sigma_{\gamma'}}{4\pi} . \quad (32.73)$$

Finally, the collision integral (32.73) must be integrated over the direction  $\hat{\mathbf{p}}_{\gamma'}$  of the scattered photon. The integration is facilitated if the angular dependence of the invariant amplitude squared  $|\mathcal{M}|^2$  given by equation (32.54) is expanded in Legendre polynomials,  $\frac{1}{2}(1+\mu^2) = \frac{2}{3}[1 + \frac{1}{2}P_2(\mu)]$ . Inserting the invariant amplitude squared  $|\mathcal{M}|^2$  into the collision integral (32.73) brings it to

$$C[\Theta] = |\dot{\tau}| \int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] [(\hat{\mathbf{p}}_\gamma - \hat{\mathbf{p}}_{\gamma'}) \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma) + \Theta(\hat{\mathbf{p}}_{\gamma'})] \frac{d\sigma_{\gamma'}}{4\pi} , \quad (32.74)$$

where  $\dot{\tau} \equiv -\bar{n}_e \sigma_T a$  is the scattering rate (31.4). Equation (32.74) (unlike equation (32.73)) 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 (32.74) is odd, and vanishes on angular integration:

$$\int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] \hat{\mathbf{p}}_{\gamma'} \frac{d\sigma_{\gamma'}}{4\pi} = 0 . \quad (32.75)$$

Similarly, the angular integral over the quadrupole of quantities independent of  $\hat{\mathbf{p}}_{\gamma'}$  vanishes:

$$\int P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'}) [\hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b - \Theta(\hat{\mathbf{p}}_\gamma)] \frac{d\sigma_{\gamma'}}{4\pi} = 0 . \quad (32.76)$$

The collision integral (32.74) thus reduces to

$$C[\Theta(\mathbf{x}, \hat{\mathbf{p}}_\gamma)] = |\dot{\tau}| \left\{ \hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b(\mathbf{x}) - \Theta(\mathbf{x}, \hat{\mathbf{p}}_\gamma) + \int [1 + \frac{1}{2}P_2(\hat{\mathbf{p}}_\gamma \cdot \hat{\mathbf{p}}_{\gamma'})] \Theta(\mathbf{x}, \hat{\mathbf{p}}_{\gamma'}) \frac{d\sigma_{\gamma'}}{4\pi} \right\} , \quad (32.77)$$

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 (32.103), expand the temperature fluctuation  $\Theta(\mathbf{k}, \hat{\mathbf{p}}_{\gamma'})$  in scalar multipole moments according to equation (32.46), and invoke orthogonality of the spherical harmonics. With  $\mu_\gamma$  defined by

$$\mu_\gamma \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma , \quad (32.78)$$

these manipulations bring the photon collision integral (32.77) at last to

$$C[\Theta(\mathbf{k}, \hat{\mathbf{p}}_\gamma)] = |\dot{\tau}| [-i\mu_\gamma v_b(\mathbf{k}) - \Theta(\mathbf{k}, \mu_\gamma) + \Theta_0(\mathbf{k}) - \frac{1}{2}\Theta_2(\mathbf{k})P_2(\mu_\gamma)] . \quad (32.79)$$

### 32.10 Boltzmann equation for photons

Inserting the collision term (32.79) into equation (32.44) 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,

$$\frac{d\Theta}{d\eta} = \frac{\partial\Theta}{\partial\eta} - ik\mu_\gamma\Theta - \dot{\Phi} - ik\mu_\gamma\Psi = |\dot{\tau}| \left[ -i\mu_\gamma v_b - \Theta + \Theta_0 - \frac{1}{2}\Theta_2 P_2(\mu_\gamma) \right]. \quad (32.80)$$

Expanded into the scalar harmonics  $\Theta_\ell(\eta, \mathbf{k})$ , the photon Boltzmann equation (32.80) yields the hierarchy of photon multipole equations

$$\dot{\Theta}_0 - k\Theta_1 - \dot{\Phi} = 0, \quad (32.81a)$$

$$\dot{\Theta}_1 + \frac{k}{3}(\Theta_0 - 2\Theta_2) + \frac{k}{3}\Psi = \frac{1}{3}|\dot{\tau}|(v_b - 3\Theta_1), \quad (32.81b)$$

$$\dot{\Theta}_2 + \frac{k}{5}(2\Theta_1 - 3\Theta_3) = -\frac{9}{10}|\dot{\tau}|\Theta_2, \quad (32.81c)$$

$$\dot{\Theta}_\ell + \frac{k}{2\ell+1}[\ell\Theta_{\ell-1} - (\ell+1)\Theta_{\ell+1}] = -|\dot{\tau}|\Theta_\ell \quad (\ell \geq 3). \quad (32.81d)$$

When polarization is included, the factor  $\frac{9}{10}$  on the right hand side of equation (32.81c) is decreased by a factor  $\frac{5}{6}$  to  $\frac{3}{4}$ , Exercise 34.3,

$$\frac{9}{10} \rightarrow \frac{3}{4}. \quad (32.82)$$

The Boltzmann hierarchy (32.81) shows that all the photon multipoles except the photon monopole  $\Theta_0$  are affected by electron-photon scattering, but only the photon dipole  $\Theta_1$  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.

### 32.11 Baryons

The equations governing baryonic matter are similar to those governing non-baryonic cold dark matter, §32.3, 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 (32.58),

$$C[\overset{0}{f}_b] = 0. \quad (32.83)$$

For the perturbed baryon distribution, only the first and second moments of the phase-space distribution are important, since these govern the baryon overdensity  $\delta_b$  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,

$$\int C[\dot{f}_e] \frac{2d^3p_e}{m_e(2\pi)^3} = 0 . \quad (32.84)$$

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:

$$\int C[\dot{f}_e] m_e \mathbf{v}_e \frac{2d^3p_e}{m_e(2\pi)^3} + \int C[\dot{f}_\gamma] \mathbf{p}_\gamma \frac{2d^3p_\gamma}{p_\gamma(2\pi)^3} = 0 . \quad (32.85)$$

Substituting the expression (32.77) for the photon collision integral into equation (32.85), separating out factors depending on the magnitude  $p_\gamma$  and direction  $\hat{\mathbf{p}}_\gamma$  of the photon momentum, and taking into consideration that the integral terms in equation (32.77), 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

$$\int C[\dot{f}_e] m_e \mathbf{v}_e \frac{2d^3p_e}{m_e(2\pi)^3} = \bar{n}_e \sigma_{\text{TA}} a \int \frac{\partial \dot{f}_\gamma}{\partial \ln T} p_\gamma^2 \frac{24\pi p_\gamma^2 dp_\gamma}{p_\gamma(2\pi)^3} \int [-\hat{\mathbf{p}}_\gamma \cdot \mathbf{v}_b + \Theta(\hat{\mathbf{p}}_\gamma)] \hat{\mathbf{p}}_\gamma \frac{do_\gamma}{4\pi} . \quad (32.86)$$

The integral over the magnitude  $p_\gamma$  of the photon momentum in equation (32.86) yields  $4\bar{\rho}$ , in accordance with equation (??). Transformed into Fourier space, and with only scalar terms retained, the collision integral (32.86) becomes, with  $\mu_\gamma \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma$ ,

$$\begin{aligned} \hat{\mathbf{k}} \cdot \int C[\dot{f}_e] m_e \mathbf{v}_e \frac{2d^3p_e}{m_e(2\pi)^3} &= 4\bar{\rho}_\gamma \bar{n}_e \sigma_{\text{TA}} a \int [i\mu_\gamma v_b + \Theta] \mu_\gamma \frac{do_\gamma}{4\pi} \\ &= \frac{4}{3} i \bar{\rho}_\gamma \bar{n}_e \sigma_{\text{TA}} a (v_b - 3\Theta_1) . \end{aligned} \quad (32.87)$$

The result is that the equations governing the baryon overdensity  $\delta_b$  and scalar bulk velocity  $v_b$  look like those (32.33) governing non-baryonic cold dark matter, except that the velocity equation has an additional source (32.87) arising from momentum transfer with photons through electron-photon scattering:

$$\dot{\delta}_b - kv_b - 3\dot{\Phi} = 0 , \quad (32.88a)$$

$$\dot{v}_b + \frac{\dot{a}}{a} v_b + k\Psi = -\frac{|\dot{\tau}|}{R} (v_b - 3\Theta_1) , \quad (32.88b)$$

where  $R \equiv \frac{3}{4}\bar{\rho}_b/\bar{\rho}_\gamma$  is  $\frac{3}{4}$  the baryon-to-photon density ratio, equation (31.46).

### 32.12 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 neutrino types were relativistic until some time after

recombination, equation (10.108). As long as neutrinos are relativistic, the hierarchy of Boltzmann equations is the same as that for photons, equations (32.81), but without the scattering terms,

$$\dot{\mathcal{N}}_0 - k\mathcal{N}_1 - \dot{\Phi} = 0 , \quad (32.89a)$$

$$\dot{\mathcal{N}}_1 + \frac{k}{3}(\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3}\Psi = 0 , \quad (32.89b)$$

$$\dot{\mathcal{N}}_\ell + \frac{k}{2\ell+1}[\ell\mathcal{N}_{\ell-1} - (\ell+1)\mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) . \quad (32.89c)$$

The radiative transfer equation for neutrinos can be solved explicitly, equation (33.44). 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 (32.89) 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 (33.44), is the approximation (33.47),

$$\mathcal{N}_{\ell_{\max}+1} \approx -(\mathcal{N}_{\ell_{\max}-1} + \delta_{\ell_{\max}1}\Psi) - \frac{2\ell_{\max}+1}{k\eta}\mathcal{N}_{\ell_{\max}} , \quad (32.90)$$

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 32.4. In Exercise 32.5 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 32.5 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 32.4. 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.

- Initially, only the neutrino monopole  $\mathcal{N}_0$  is finite. In the Boltzmann hierarchy (32.89) 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 (32.89) reduces to, with  $y \equiv k\eta$ ,

$$\frac{d(\mathcal{N}_0 - \Phi)}{dy} = 0 , \quad (32.91a)$$

$$\frac{d\mathcal{N}_1}{dy} = -\frac{1}{3}(\mathcal{N}_0 + \Psi) , \quad (32.91b)$$

$$\frac{d\mathcal{N}_\ell}{dy} = -\frac{\ell}{2\ell+1}\mathcal{N}_{\ell-1} \quad (\ell \geq 2) . \quad (32.91c)$$

Show that the initial behaviour of the neutrino multipoles is

$$\mathcal{N}_\ell = \frac{(-y)^\ell}{(2\ell+1)!!} (\mathcal{N}_0 + \Psi) \quad (\ell \geq 1). \quad (32.92)$$

2. Let  $f_\gamma$  and  $f_\nu$  be the photon and neutrino fraction of the total radiation density,

$$f_\gamma \equiv \frac{\bar{\rho}_\gamma}{\bar{\rho}_\gamma + \bar{\rho}_\nu} = 1 - f_\nu, \quad f_\nu \equiv \frac{\bar{\rho}_\nu}{\bar{\rho}_\gamma + \bar{\rho}_\nu} = \frac{6 \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}}{2 + 6 \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}} \approx 0.405. \quad (32.93)$$

Show that the Einstein energy equation implies, initially,

$$-\Psi = 2(\Phi + \zeta_r), \quad (32.94)$$

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 implies, initially,

$$\Psi - \Phi = -\frac{4}{5} f_\nu (\Psi + \Phi + \zeta_\nu). \quad (32.95)$$

3. Conclude that, for adiabatic initial conditions  $\zeta_\nu = \zeta_\gamma$ ,

$$\Phi = (1 + \frac{2}{5} f_\nu) \Psi. \quad (32.96)$$


---

### 32.13 Truncating the photon Boltzmann hierarchy

Photons are tightly bound 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 §31.7, prior to recombination, when  $|\dot{\tau}|$  is large, scattering keeps successive multipoles smaller by factors of  $k/|\dot{\tau}|$ , equation (31.66). Keeping only the dominant  $\Theta_{\ell=1}$  term on the left hand side of the Boltzmann hierarchy (32.80) for  $\ell \geq 2$  implies

$$\Theta_{\ell+1} \approx -(1 + \frac{1}{3} \delta_{\ell 1}) \frac{(\ell+1)k}{(2\ell+3)|\dot{\tau}|} \Theta_\ell \quad \text{for } \ell \geq 2. \quad (32.97)$$

The factor  $\frac{4}{3}$  in equation (32.97) for  $\ell = 1$  includes the effects of polarization; without polarization the factor is  $\frac{10}{9}$ .

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 (32.90),

$$\Theta_{\ell_{\max}+1} \approx -(\Theta_{\ell_{\max}-1} + \delta_{\ell_{\max} 1} \Psi) - \frac{2\ell_{\max} + 1}{k\eta} \Theta_{\ell_{\max}}. \quad (32.98)$$

The truncation of the photon Boltzmann hierarchy adopted in equation (32.4) is an interpolation between the scattering and free-streaming regimes (32.97) and (32.98).

## 32.14 Massive neutrinos

Once neutrinos become non-relativistic, 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 non-relativistic, equation (10.108). 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.

### 32.14.1 Simplified treatment of massive neutrinos

A full treatment of massive neutrinos, §32.14.2, requires integrating a Boltzmann hierarchy of multipole equations for each of a spectrum of neutrino momenta  $p_\nu$ . 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 non-relativistic neutrinos at some time. This was the strategy suggested in Exercise 31.4.

Another possible simplified strategy is to follow the Boltzmann hierarchy (32.100) for just a single representative neutrino momentum  $p_\nu$  near the peak of the distribution,  $p_\nu/T_\nu = 1$ .

### 32.14.2 Full treatment of massive neutrinos

The Boltzmann equation for collisionless neutrinos with mass is equation (32.43) with zero collision term. For massive neutrinos, the neutrino velocity  $v_\nu$  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  $p_\nu$ . Since the temperature fluctuation  $\mathcal{N}$  is already of first order, it suffices to treat the particle velocity  $v_\nu$  in the Boltzmann equation (32.43) 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  $p_\nu/T_\nu$ ,

$$\mathcal{N}(\eta, \mathbf{k}, \mathbf{p}_\nu) = \mathcal{N}(\eta, \mathbf{k}, p_\nu/T_\nu, \hat{\mathbf{p}}_\nu) . \quad (32.99)$$

The harmonics  $\mathcal{N}_\ell(\eta, \mathbf{k}, p_\nu/T_\nu)$  of the temperature fluctuation are functions of the ratio  $p_\nu/T_\nu$ .

At the risk of being repetitious, the Boltzmann hierarchy of equations for a species of massive neutrino is,

equations (32.48),

$$\dot{\mathcal{N}}_0 - v_\nu k \mathcal{N}_1 - \dot{\Phi} = 0 , \quad (32.100a)$$

$$\dot{\mathcal{N}}_1 + \frac{v_\nu k}{3} (\mathcal{N}_0 - 2\mathcal{N}_2) + \frac{k}{3v_\nu} \Psi = 0 , \quad (32.100b)$$

$$\dot{\mathcal{N}}_\ell + \frac{v_\nu k}{2\ell+1} [\ell\mathcal{N}_{\ell-1} - (\ell+1)\mathcal{N}_{\ell+1}] = 0 \quad (\ell \geq 2) , \quad (32.100c)$$

which differs from the massless neutrino hierarchy (32.89) 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 non-relativistic, a full treatment requires following neutrino with different momenta  $p_\nu$  separately. In due course the neutrinos become non-relativistic, and the equations re-simplify to the non-relativistic limit.

Because the massive neutrino multipoles  $\mathcal{N}_\ell(p_\nu/T_\nu)$  depend on neutrino momentum  $p_\nu$ , the perturbed neutrino energy-momentum tensor  $\dot{T}_\nu^{kl}$  is more complicated than the massless case, equations (32.52). The perturbed energy density, energy flux, monopole pressure, and quadrupole pressure of massive neutrinos are

$$\dot{T}_\nu^{00} = \int \frac{\partial \dot{f}_\nu^0}{\partial \ln T_\nu} \mathcal{N}_0(p_\nu/T_\nu) E_\nu \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (32.101a)$$

$$\hat{k}_a T_\nu^{0a} = -i \int \frac{\partial \dot{f}_\nu^0}{\partial \ln T_\nu} \mathcal{N}_1(p_\nu/T_\nu) p_\nu^a \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (32.101b)$$

$$\frac{1}{3} \delta_{ab} \dot{T}_\nu^{ab} = \frac{1}{3} \int \frac{\partial \dot{f}_\nu^0}{\partial \ln T_\nu} \mathcal{N}_0(p_\nu/T_\nu) \frac{p_\nu^2}{E_\nu} \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} , \quad (32.101c)$$

$$\left(\frac{3}{2} \hat{k}_a \hat{k}_b - \frac{1}{2} \delta_{ab}\right) T_\nu^{ab} = - \int \frac{\partial \dot{f}_\nu^0}{\partial \ln T_\nu} \mathcal{N}_2(p_\nu/T_\nu) \frac{p_\nu^2}{E_\nu} \frac{4\pi p_\nu^2 dp_\nu}{(2\pi\hbar)^3} . \quad (32.101d)$$

### 32.15 Appendix: Legendre polynomials

The Legendre polynomials  $P_\ell(\mu)$  satisfy the orthogonality relations

$$\int_{-1}^1 P_\ell(\mu) P_{\ell'}(\mu) d\mu = \frac{2}{2\ell+1} \delta_{\ell\ell'} , \quad (32.102)$$

the addition theorem

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}^*(\hat{\mathbf{a}}) Y_{\ell m}(\hat{\mathbf{b}}) = \frac{2\ell+1}{4\pi} P_\ell(\hat{\mathbf{a}} \cdot \hat{\mathbf{b}}) , \quad (32.103)$$

the recurrence relation

$$\mu P_\ell(\mu) = \frac{1}{2\ell+1} [\ell P_{\ell-1}(\mu) + (\ell+1) P_{\ell+1}(\mu)] , \quad (32.104)$$

and the derivative relation

$$\frac{dP_\ell(\mu)}{d\mu} = \frac{\ell+1}{1-\mu^2} [\mu P_{\ell-1}(\mu) - P_{\ell+1}(\mu)] . \quad (32.105)$$

The first few Legendre polynomials are

$$P_0(\mu) = 1 , \quad P_1(\mu) = \mu , \quad P_2(\mu) = -\frac{1}{2} + \frac{3}{2}\mu^2 . \quad (32.106)$$

# 33

---

## 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 (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).

### 33.1 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 (32.80) for the temperature fluctuation  $\Theta(\eta, \mathbf{k}, \mu_\gamma)$ , where  $\mu_\gamma \equiv \hat{\mathbf{k}} \cdot \hat{\mathbf{p}}_\gamma$  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

$$\left( \frac{\partial}{\partial \eta} - ik\mu_\gamma + |\dot{\tau}| \right) (\Theta + \Psi) = I + |\dot{\tau}| S , \quad (33.1)$$

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 (33.1) is the Integrated Sachs-Wolfe (ISW) term,

$$I(\eta, \mathbf{k}) \equiv \dot{\Psi}(\eta, \mathbf{k}) + \dot{\Phi}(\eta, \mathbf{k}) , \quad (33.2)$$

so-called because, as will be seen in equation (33.15), 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 (33.1) embodies

source terms arising from Thomson scattering, a sum of monopole, dipole, and quadrupole harmonics,

$$\begin{aligned} S(\eta, \mathbf{k}, \mu_\gamma) &\equiv \Theta_0(\eta, \mathbf{k}) + \Psi(\eta, \mathbf{k}) - i\mu_\gamma v_b(\eta, \mathbf{k}) - \frac{1}{2}\Theta_2(\eta, \mathbf{k})P_2(\mu_\gamma) \\ &= \sum_{n=0}^2 (-i)^n S_n(\eta, \mathbf{k}) P_n(\mu_\gamma) , \end{aligned} \quad (33.3)$$

with harmonic coefficients  $S_n(\eta, \mathbf{k})$ ,

$$S_0 \equiv \Theta_0 + \Psi , \quad (33.4a)$$

$$S_1 \equiv v_b , \quad (33.4b)$$

$$S_2 \equiv \frac{1}{2}\Theta_2 . \quad (33.4c)$$

The electron-photon (Thomson) scattering optical depth  $\tau$  is defined by equation (31.3). The optical depth is zero,  $\tau_0 = 0$ , at zero redshift, and increases going backwards in time  $\eta$  to higher redshift. The radiative transfer equation (33.1) can be written (note that  $\dot{\tau}$  is negative)

$$e^{ik\mu_\gamma\eta+\tau} \frac{d}{d\eta} [e^{-ik\mu_\gamma\eta-\tau} (\Theta + \Psi)] = I - \dot{\tau}S . \quad (33.5)$$

The solution for the photon distribution  $\Theta(\eta_0, \mathbf{k}, \mu_\gamma)$  today is obtained by integrating the radiative transfer equation (33.5) over the line of sight from the Big Bang to the present time,

$$\Theta(\eta_0, \mathbf{k}, \mu_\gamma) + \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [I(\eta, \mathbf{k}) - \dot{\tau}S(\eta, \mathbf{k}, \mu_\gamma)] e^{-ik\mu_\gamma(\eta-\eta_0)-\tau} d\eta . \quad (33.6)$$

Notice that the left hand side of the solution (33.6) of the radiative transfer equation is not the temperature fluctuation  $\Theta(\eta_0, \mathbf{k}, \mu_\gamma)$  by itself, but rather the temperature fluctuation redshifted by the potential,  $\Theta(\eta_0, \mathbf{k}, \mu_\gamma) + \Psi(\eta_0, \mathbf{k})$ . 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.

### 33.1.1 Visibility function

Introduce the **visibility function**  $g(\eta)$  defined by

$$g(\eta) \equiv -\dot{\tau}e^{-\tau} . \quad (33.7)$$

The visibility function  $g(\eta)$ , illustrated in Figure 33.1, 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,

$$\int_0^{\eta_0} g(\eta) d\eta = \int_0^0 -e^{-\tau} d\tau = [e^{-\tau}]_0^0 = 1 . \quad (33.8)$$

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.

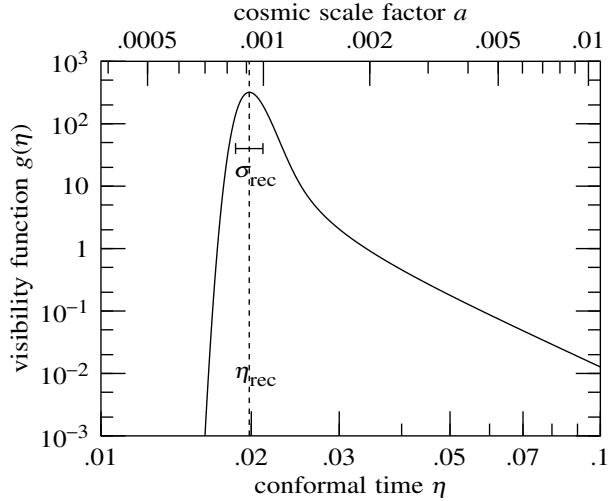


Figure 33.1 Visibility function  $g(\eta)$  as a function of conformal time  $\eta$  in units of the conformal time today,  $\eta_0 = 1$ . The visibility function here is calculated from the Peebles approximation to recombination, Exercise 30.7. The dashed vertical line marks the time  $\eta_{\text{rec}}$  of recombination, where the optical depth is one. The width  $\sigma_{\text{rec}}$  of recombination is the standard deviation of a Gaussian fit to the core of the visibility function.

The solution (33.6) of the radiative transfer equation can be written in terms of the visibility function  $g(\eta)$  as

$$\Theta(\eta_0, \mathbf{k}, \mu_\gamma) + \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [e^{-\tau} I(\eta, \mathbf{k}) + g(\eta) S(\eta, \mathbf{k}, \mu_\gamma)] e^{-ik\mu_\gamma(\eta-\eta_0)} d\eta . \quad (33.9)$$

### 33.2 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  $\Theta(\eta_0, \mathbf{k}, \mu_\gamma) + \Psi(\eta_0, \mathbf{k})$  in harmonics, equation (32.46).

The  $e^{-ik\mu_\gamma(\eta-\eta_0)}$  factor in the integrand on the right hand side of equation (33.9) can be expanded in harmonics through the general formula

$$e^{-iy\mu_\gamma} = \sum_{\ell=0}^{\infty} (-i)^\ell (2\ell+1) P_\ell(\mu_\gamma) j_\ell(y) , \quad (33.10)$$

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  $e^{-ik\mu_\gamma(\eta-\eta_0)}$  in the integrand of equation (33.9) is a sum of harmonics, equation (33.3). It is useful to introduce modified spherical Bessel functions  $j_{\ell n}(y)$  defined by an expansion analogous to (33.10),

$$(-i)^n P_n(\mu_\gamma) e^{-iy\mu_\gamma} = \sum_{\ell=0}^{\infty} (-i)^\ell (2\ell+1) P_\ell(\mu_\gamma) j_{\ell n}(y) . \quad (33.11)$$

The Legendre functions  $P_n(\mu_\gamma)$  are polynomials in  $\mu_\gamma$ . Acting on  $e^{-iy\mu_\gamma}$ , these polynomials can be replaced by derivatives with respect to  $y$  through

$$\mu_\gamma^n e^{-iy\mu_\gamma} = \left( i \frac{\partial}{\partial y} \right)^n e^{-iy\mu_\gamma} . \quad (33.12)$$

The resulting modified spherical Bessel functions  $j_{\ell n}(y)$  with  $n = 0, 1, 2$  relevant here are

$$j_{\ell 0} = j_\ell , \quad j_{\ell 1} = j'_\ell , \quad j_{\ell 2} = \frac{1}{2} j_\ell + \frac{3}{2} j''_\ell , \quad (33.13)$$

where ' denotes the total derivative,  $j'_\ell \equiv dj_\ell(y)/dy$ . The modified spherical Bessel functions are even or odd as  $j_{l n}(-y) = (-)^{l+n} j_{l n}(y)$ . The harmonic expansion of equation (33.9) is thus

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = \int_0^{\eta_0} \left\{ e^{-\tau} I(\eta, \mathbf{k}) j_\ell [k(\eta_0 - \eta)] + g(\eta) \sum_{n=0}^2 S_n(\eta, \mathbf{k}) j_{\ell n} [k(\eta - \eta_0)] \right\} d\eta , \quad (33.14)$$

where  $g(\eta)$  is the visibility function defined by equation (33.7). With the ISW and scattering source terms  $I$  and  $S_n$  written out explicitly, equation (33.14) is

$$\begin{aligned} \Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} [\dot{\Psi}(\eta, \mathbf{k}) + \dot{\Phi}(\eta, \mathbf{k})] j_\ell [k(\eta - \eta_0)] && \text{ISW} \\ &\quad + g(\eta) \left\{ [\Theta_0(\eta, \mathbf{k}) + \Psi(\eta, \mathbf{k})] j_\ell [k(\eta - \eta_0)] \right. && \text{monopole} \\ &\quad + v_b(\eta, \mathbf{k}) j_{\ell 1} [k(\eta - \eta_0)] && \text{dipole} \\ &\quad \left. + \frac{1}{2} \Theta_2(\eta, \mathbf{k}) j_{\ell 2} [k(\eta - \eta_0)] \right\} d\eta && \text{quadrupole} . \end{aligned} \quad (33.15)$$

The term in the first line on the right hand side of equation (33.15) 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 (33.4). Note that the monopole term (on both sides of equation (33.15)) is not the temperature fluctuation  $\Theta_0$  by itself, but rather  $\Theta_0 + \Psi$ , which is the temperature fluctuation redshifted by the potential  $\Psi$ . In the tight-coupling approximation, the baryon velocity on the third line approximates the photon velocity,  $v_b \approx 3\Theta_1$ .

Figure 33.2 shows an illustrative example of the factors that go into the monopole and dipole contributions to the integrand of the solution (33.15) of the radiative transfer equation.

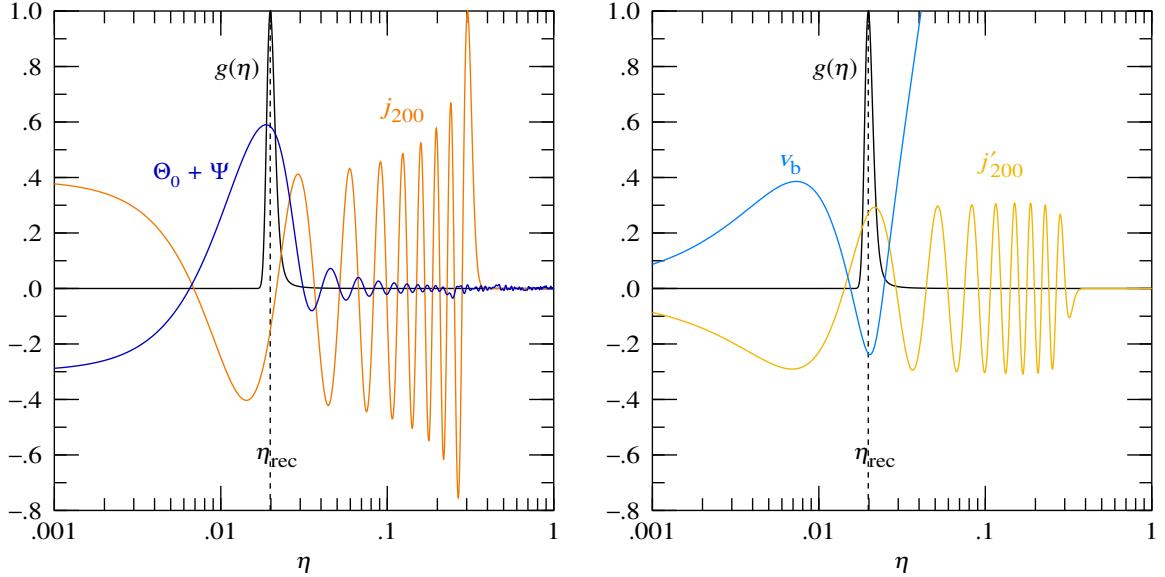


Figure 33.2 Illustrative example of the factors that go into the (left) monopole ( $n = 0$ ) and (right) dipole ( $n = 1$ ) contributions to the integrand of the solution (33.15) of the radiative transfer equation, as a function of conformal time  $\eta$ , in units  $\eta_0 = 1$ . The example is for 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 (33.7), scattering source terms  $S_n(\eta, \mathbf{k})$ , equations (33.4), and modified spherical Bessel functions  $j_{\ell n}[k(\eta_0 - \eta)]$ , equations (33.13). The visibility function  $g(\eta)$  has been scaled to 1 at its peak, and the monopole and dipole spherical Bessel functions  $j_\ell$  and  $j'_\ell$  have been scaled so  $j_\ell$  equals 1 at its (first) peak. The cosmological model is as given in §31.3. The dashed vertical line marks the time  $\eta_{\text{rec}}$  of recombination, where the Thomson optical depth is one.

### 33.2.1 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  $\Theta_\ell^{\text{obs}}$  expanded with respect to the direction  $\hat{\mathbf{n}}$  of observation are obtained from  $\Theta_\ell$  by flipping the sign of the photon direction,  $\Theta_\ell^{\text{obs}}(\eta, \mathbf{k}, \mu_\gamma) = \Theta_\ell(\eta, \mathbf{k}, -\mu_\gamma)$ . Flipping the sign of  $\hat{\mathbf{p}}$  is equivalent to flipping the sign of odd parity fluctuations,

$$\Theta_\ell^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^\ell \Theta_\ell(\eta_0, \mathbf{k}) . \quad (33.16)$$

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 (33.15) and simultaneously to flip the sign of the odd source functions  $S_\ell$ , namely  $S_1 = v_b \rightarrow -v_b$ . The CMB power spectrum involves products of pairs of  $\Theta_\ell^{\text{obs}}$  with the same  $\ell$ , and is unaffected by the sign flip  $\hat{\mathbf{n}} = -\hat{\mathbf{p}}_\gamma$  in the solution (33.15) of the radiative transfer equation.

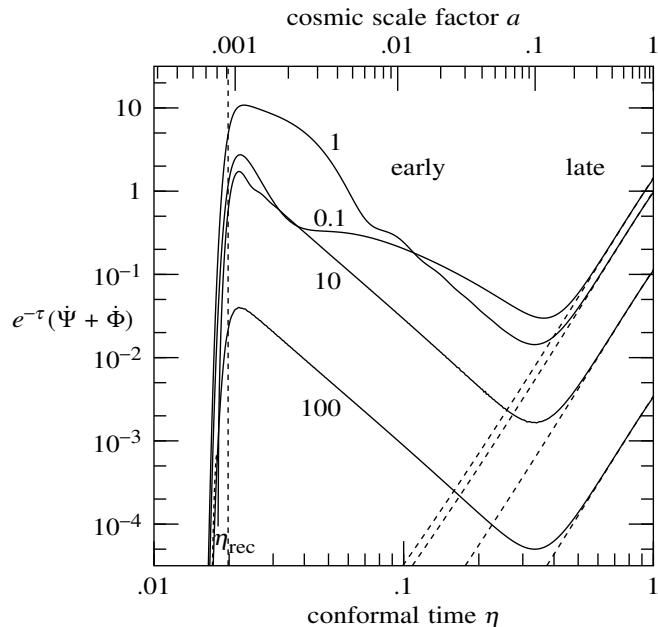


Figure 33.3 ISW integrand  $e^{-\tau}(\dot{\Psi} + \dot{\Phi})$  in equation (33.15) as a function of conformal time  $\eta$ , in units  $\eta_0 = 1$ , for the standard flat  $\Lambda$ CDM model of §31.3. Curves are labelled with the wavenumber  $k/(a_{\text{eq}} H_{\text{eq}})$  in units of the Hubble 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.

### 33.2.2 Integrated Sachs-Wolfe (ISW) effect

The first line of the solution (33.15) 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  $\Psi$  and  $\Phi$  (which are equal at late times) evolve in proportion to the growth factor  $g(a)$ , equation (29.125) (not to be confused with the visibility function  $g(\eta)$ ). The ISW integrand splits accordingly into early and late contributions,

The early and late time contributions to the ISW term are illustrated in Figure 33.3.

In addition to the early and late ISW effects, there is a “nonlinear” ISW effect. Nonlinear gravitational clustering causes the potential  $\Phi$  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 33.3 does not include the nonlinear ISW effect.

### 33.2.3 CMB transfer function in Fourier space

As seen in Chapters 29–32, 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

$$T_\ell(\eta, k) \equiv \frac{\Theta_\ell(\eta, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta, \mathbf{k})}{\zeta(\mathbf{k})}. \quad (33.18)$$

By isotropy, the CMB transfer function  $T_\ell(\eta, k)$  is a function only of the magnitude  $k$  of the wavevector  $\mathbf{k}$ .

Figure 33.4 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 (33.15) of the radiative transfer equation. The CMB transfer functions shown in Figure 33.4 are from a Boltzmann computation including photon and neutrino multipoles up to  $\ell_{\max} = 15$ .

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 33.4. The transfer functions are small for  $k(\eta_0 - \eta_{\text{rec}}) \ll \ell$ , peak at

$$k(\eta_0 - \eta_{\text{rec}}) \sim \ell + \ell^{1/3}, \quad (33.19)$$

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 33.4,

$$T_\ell(\eta_0, \mathbf{k})|_{\text{envelope}} \propto \exp(-k\eta_0/600). \quad (33.20)$$

The exponential decline is caused in part by dissipative processes around the time of recombination, §31.7 and §31.8, and in part by the finite width of recombination.

Besides the total, Figure 33.4 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 (33.15). The Figure therefore shows also the envelope of the oscillatory contributions. The envelope is computed as the absolute value of the complex integral in which the Bessel factor  $j_{\ell n}(y)$  in the integrand is replaced by the complex Hankel factor  $h_{\ell n}(y)$ ,

$$h_{\ell n}(y) \equiv j_{\ell n}(y) + \begin{cases} 0 & |y| < l + \frac{1}{2} + (l + \frac{1}{2})^{1/3} \\ i y_{\ell n}(y) & |y| \geq l + \frac{1}{2} + (l + \frac{1}{2})^{1/3} \end{cases}, \quad (33.21)$$

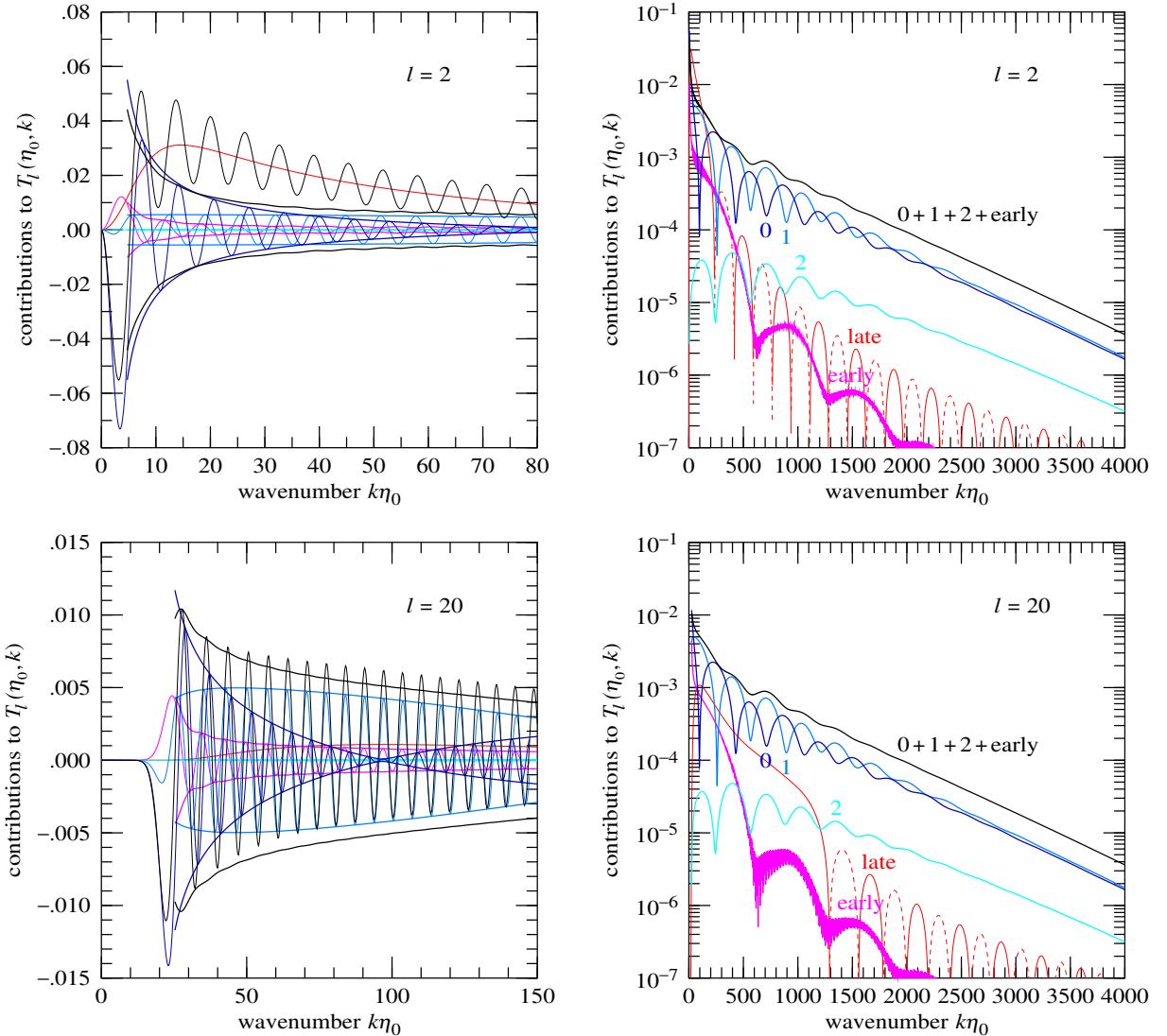


Figure 33.4 (Continued on the next page.) CMB transfer functions  $T_\ell(\eta_0, k)$  for a selection of harmonics  $\ell$ , plotted (left) linearly, showing the oscillating functions and their envelopes, and (right) logarithmically over a broader range of wavenumber  $k$ , showing only the envelopes of the underlying oscillating functions. The total (black) is a sum of the various contributions in equation (33.15): 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  $\ell$  and small  $k$ ). The cosmological model is the flat  $\Lambda$ CDM concordance model of §31.3. The computation is a Boltzmann computation including photon and neutrino multipoles up to  $\ell_{\max} = 15$ .

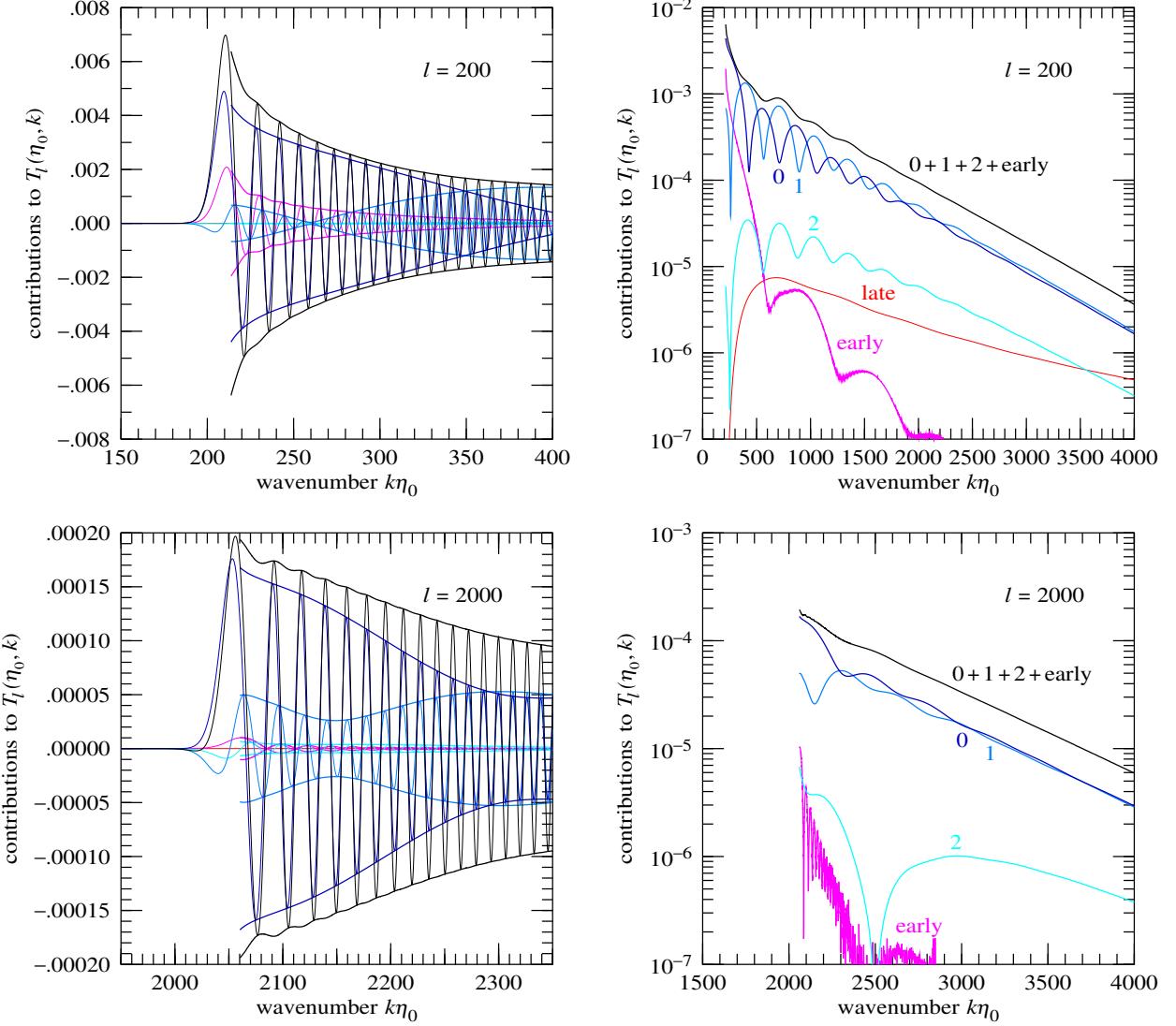


Figure 33.4 continued.

with  $y_{ln}(y)$  the modified spherical Bessel function of the second kind (whereas  $j_{ln}(y)$  is the modified spherical Bessel function of the first kind). The cut at  $y = l + \frac{1}{2} + (l + \frac{1}{2})^{1/3}$ , which is roughly the location of the first zero of  $y_{ln}(y)$ , is introduced to prevent the diverging behaviour of  $y_{ln}(y)$  as  $y \rightarrow 0$  from dominating the integral.

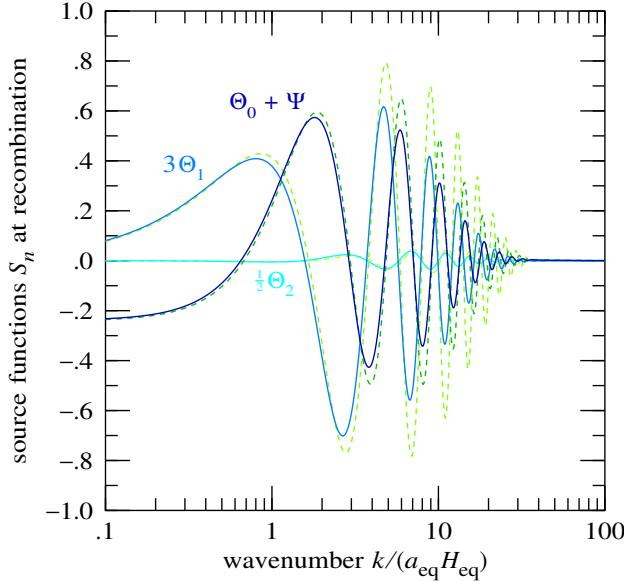


Figure 33.5 Thomson scattering source functions  $S_n$ , equation (33.4), at recombination as a function of wavenumber  $k/(a_{\text{eq}}H_{\text{eq}})$ . The solid (bluish) lines are the values  $\bar{S}_n(\mathbf{k})$  averaged over the visibility function  $g(\eta)$ , while the dashed (greenish) lines are the instantaneous values  $S_n(\eta_{\text{rec}}, \mathbf{k})$  at recombination, where the Thomson optical depth is one. The 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  $\ell_{\text{max}} = 1$ ), since this turns out to yield a better rapid recombination approximation to the CMB power spectrum than a full Boltzmann treatment, Figure 33.7. 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  $\Lambda$ CDM model described in §31.3.

### 33.2.4 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 terms, the multipoles  $\Theta_\ell(\eta_0, \mathbf{k})$  of the temperature fluctuation today are

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0}\Psi(\eta_0, \mathbf{k}) \approx \sum_{n=0}^2 \bar{S}_n(\mathbf{k}) j_{\ell n} [k(\eta_{\text{rec}} - \eta_0)] . \quad (33.22)$$

A better approximation that works also at larger  $k$  is the **rapid recombination approximation**, which replaces the source functions by their averages  $\bar{S}_n$  over recombination,

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0}\Psi(\eta_0, \mathbf{k}) \approx \sum_{n=0}^2 \bar{S}_n(\mathbf{k}) j_{\ell n} [k(\eta_{\text{rec}} - \eta_0)] , \quad \bar{S}_n(\mathbf{k}) \equiv \int_0^{\eta_0} g(\eta) S_n(\eta, \mathbf{k}) d\eta . \quad (33.23)$$

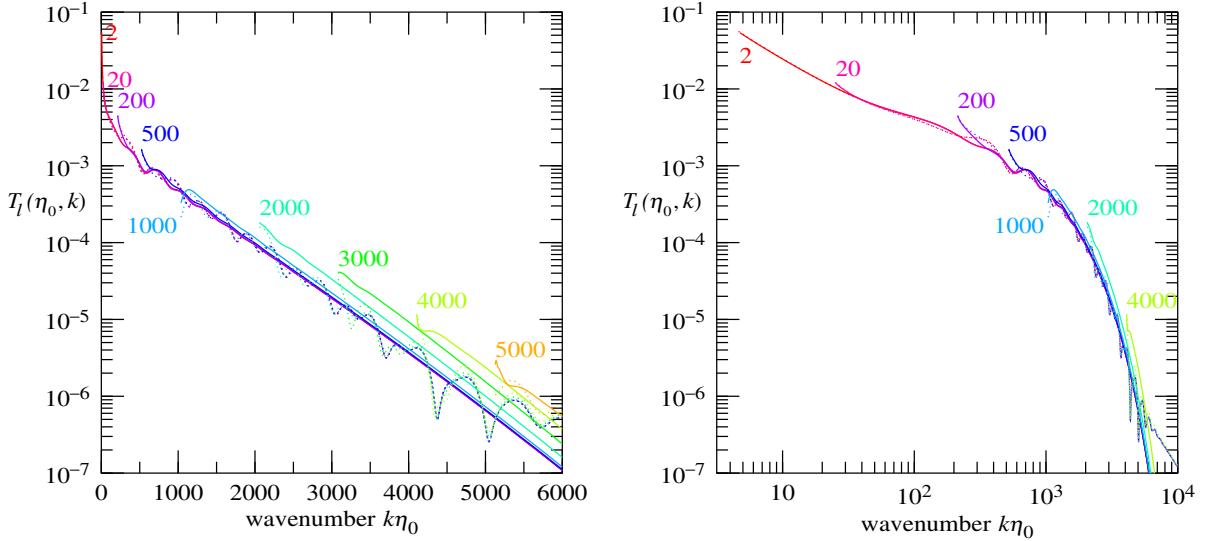


Figure 33.6 CMB transfer functions computed from a Boltzmann treatment (solid lines) including photon and neutrino multipoles up to  $\ell_{\max} = 15$ , compared to their values in the rapid recombination approximation (dotted lines) with the source functions taken from the hydrodynamic approximation, for a selection of harmonics  $\ell$ , 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, and quadrupole contributions, but exclude the ISW (early and late) contributions. 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  $\Lambda$ CDM model described in §31.3.

The instantaneous and fast approximations agree at small wavenumbers, where the source terms are slowly varying over the visibility function. The fast approximation works also at larger wavenumbers, where the source functions  $S_n(\eta, \mathbf{k})$  change significantly over the course of recombination.

Figure 33.5 shows the monopole, dipole, and quadrupole source terms  $S_n$  that go into the instantaneous approximation (dashed lines) and the rapid recombination approximation (solid lines). Averaging over recombination reduces the 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 31.1 or 32.1. 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 Figure 33.5 and 33.6.

Figure 33.6 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 (33.23), with source functions computed in the hydrodynamic approximation, for a selection of harmonics  $\ell$ . Whereas

the envelopes computed in the Boltzmann treatment are rather smooth at large wavenumber  $k$  (and exponentially declining, equation (33.20)), 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 (33.33) 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 33.7.

### 33.2.5 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  $\Theta_\ell(\eta, \mathbf{k})$ ,

$$\frac{\delta_{\ell'\ell}}{4\pi} (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) C_\ell(\eta, k) \equiv \langle [\Theta_{\ell'}(\eta, \mathbf{k}') + \delta_{\ell 0}\Psi(\eta, \mathbf{k}')] [\Theta_\ell(\eta, \mathbf{k}) + \delta_{\ell 0}\Psi(\eta, \mathbf{k})] \rangle . \quad (33.24)$$

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 (33.15) of the radiative transfer equation.

In terms of the CMB transfer function (33.18) and the primordial power spectrum  $P_\zeta(k)$  defined by equation (29.129), the CMB power spectrum  $C_\ell(\eta, k)$  is

$$C_\ell(\eta, k) = 4\pi |T_\ell(\eta, k)|^2 P_\zeta(k) . \quad (33.25)$$

## 33.3 CMB in real space

### 33.3.1 CMB harmonics in real space

The solution (33.15) 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  $\eta$  and comoving position  $\mathbf{x}$  in observed direction  $\hat{\mathbf{n}}$  on the sky is related to the Fourier-space temperature fluctuation by

$$\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}}) = \int e^{-i\mathbf{k}\cdot\mathbf{x}} \Theta(\eta, \mathbf{k}, \hat{\mathbf{n}}) \frac{d^3k}{(2\pi)^3} . \quad (33.26)$$

Astronomers observe the temperature fluctuation  $\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}})$  now, at time  $\eta_0$ , 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,

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}) = \int \Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}) \frac{d^3k}{(2\pi)^3} . \quad (33.27)$$

The spherical harmonic expansion of the observed real-space temperature fluctuation today is, with a conventional choice of normalization of harmonics  $\Theta_{\ell m}$ ,

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \Theta_{\ell m}(\eta_0, \mathbf{x}_0) Y_{\ell m}^*(\hat{\mathbf{n}}) . \quad (33.28)$$

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 (32.46) of the Fourier-space temperature fluctuation may be written, in view of the relation (32.103) between Legendre polynomials and spherical harmonics

$$\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} (-i)^\ell 4\pi \Theta_\ell(\eta, \mathbf{k}) Y_{\ell m}(\hat{\mathbf{k}}) Y_{\ell m}^*(\hat{\mathbf{n}}) . \quad (33.29)$$

From equations (33.27)–(33.29) it follows that the real-space photon harmonics are

$$\Theta_{\ell m}(\eta_0, \mathbf{x}_0) = 4\pi(-i)^\ell \int \Theta_\ell(\eta_0, \mathbf{k}) Y_{\ell m}(\hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3} . \quad (33.30)$$

### 33.3.2 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  $\Theta_{\ell m}(\eta_0, \mathbf{x}_0)$ ,

$$\delta_{\ell' \ell} \delta_{m' m} C_\ell(\eta_0) \equiv \langle [\Theta_{\ell' m'}^*(\eta_0, \mathbf{x}_0) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{x}_0)] [\Theta_{\ell m}(\eta_0, \mathbf{x}_0) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{x}_0)] \rangle . \quad (33.31)$$

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 (33.30) for the real-space harmonics in terms of Fourier-space harmonics, together with the power spectrum (33.24) of the Fourier-space harmonics, it follows that the power spectrum  $C_\ell(\eta_0)$  of real-space harmonics of the CMB today is

$$C_\ell(\eta_0) = \int C_\ell(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (33.32)$$

In terms of the CMB transfer function  $T_\ell$  and primordial curvature power spectrum  $P_\zeta$  or its dimensionless equivalent  $\Delta_\zeta^2$ , equation (29.131), the power spectrum  $C_\ell(\eta_0)$  is, from equation (33.25),

$$C_\ell(\eta_0) = 4\pi \int |T_\ell(\eta_0, k)|^2 P_\zeta(k) \frac{4\pi k^2 dk}{(2\pi)^3} = 4\pi \int |T_\ell(\eta_0, k)|^2 \Delta_\zeta^2(k) \frac{dk}{k}. \quad (33.33)$$

If the primordial power spectrum  $\Delta_\zeta^2$  is a power-law with tilt  $n$ , equation (29.134), then the CMB power spectrum today is

$$C_\ell(\eta_0) = 4\pi \Delta_\zeta^2(k_p) \int_0^\infty |T_\ell(\eta_0, k)|^2 \left(\frac{k}{k_p}\right)^{n-1} \frac{dk}{k}. \quad (33.34)$$

As discussed in §33.2.3, 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  $\ell$ , equation (33.19)), and then oscillate with an exponentially declining envelope, equation (33.20). Thus the power spectrum  $C_\ell(\eta_0)$  (33.33) at harmonic  $\ell$  principally probes comoving scales  $1/k$  that are  $1/\ell$  times the comoving distance  $\eta_0 - \eta_{\text{rec}}$  to recombination today,

$$\frac{1}{k} \sim \frac{\eta_0 - \eta_{\text{rec}}}{\ell}. \quad (33.35)$$

Physically, harmonic number  $\ell$  probes angular scale  $\theta \sim \pi/\ell$  on the sky, and the power spectrum at harmonic number  $\ell$  probes comoving scale  $\pi/k \sim (\eta_0 - \eta_{\text{rec}})\theta$  on the CMB sky.

### 33.3.3 Instantaneous and rapid recombination approximations 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 instantaneous or rapid recombination approximations offers a short-cut to computing the CMB power spectrum that at least captures qualitative features. The instantaneous or rapid recombination approximations effectively sidestep step 5 of the numerical computation outlined in §29.8.

---

**Exercise 33.1 CMB power spectrum in the instantaneous and rapid recombination approximation.** Compute the CMB power spectrum  $C_\ell(\eta_0)$  today in: (1) the instantaneous recombination approximation, equation (33.22), with source functions calculated in the simple approximation, §29.7; and (2) the rapid recombination approximation, equation (33.23), with source functions calculated in the hydrodynamic approximation, §31.2. Comment.

**Solution.** See Figures 33.7 and 33.8. I used the standard flat  $\Lambda$ CDM cosmological parameters given in §31.3, and the normalization of the power spectrum measured from Planck, equation (29.135) I used Mathematica to

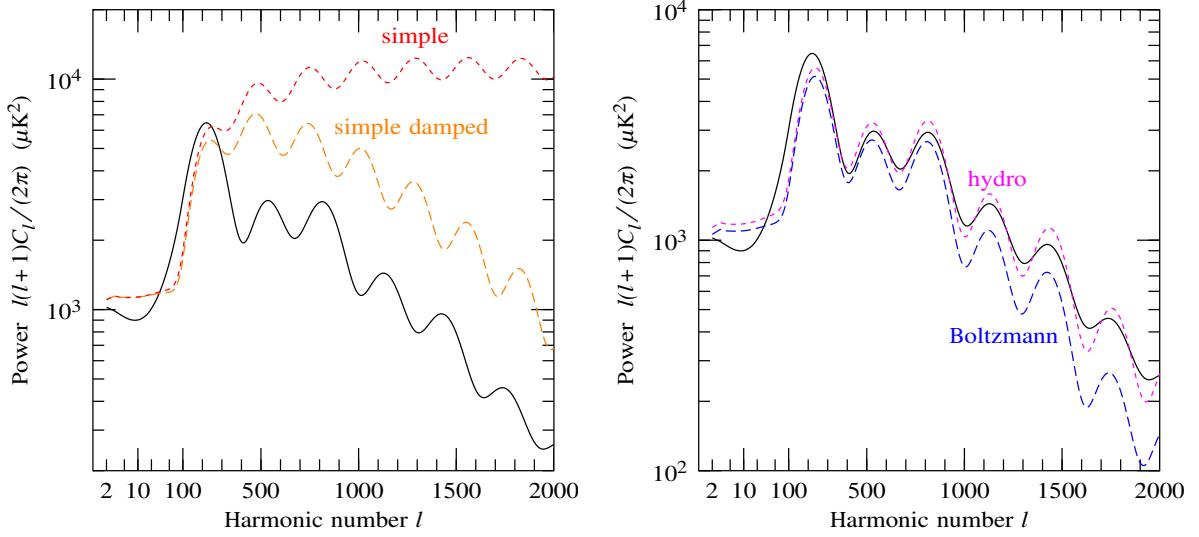


Figure 33.7 Model CMB power spectra computed (left) in the instantaneous approximation, equation (33.22), with source functions computed in the simple approximation, §29.7, without (short dashed line) and with (long dashed line) artificial damping, equations (29.56) and (29.57) with  $\epsilon = 10^{-3}$ ; and (right) in the rapid recombination approximation, equation (33.23), with source functions computed in the hydrodynamic approximation (§31.2, short dashed line), and in a full Boltzmann treatment (§32.1, long dashed line) including photon and neutrino multipoles up to  $\ell_{\max} = 15$ . The solid (black) lines are a reference model power spectrum computed with CAMB. The instantaneous and rapid recombination approximations exclude (early and late) ISW contributions to the power spectrum. The CAMB spectrum is similar to that shown in Figure 10.2, but without refinements from reionization and lensing.

solve the evolutionary equations in the simple and hydrodynamic approximations. But for the integral (33.33), I abandoned fighting Mathematica, and resorted to a publicly available implementation of Bessel functions (Amos, 1986), and a cubic spline integration implemented in fortran. In Figure 33.8 (but not Figure 33.7) 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 (33.15) and (33.17), 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 (31.46). 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

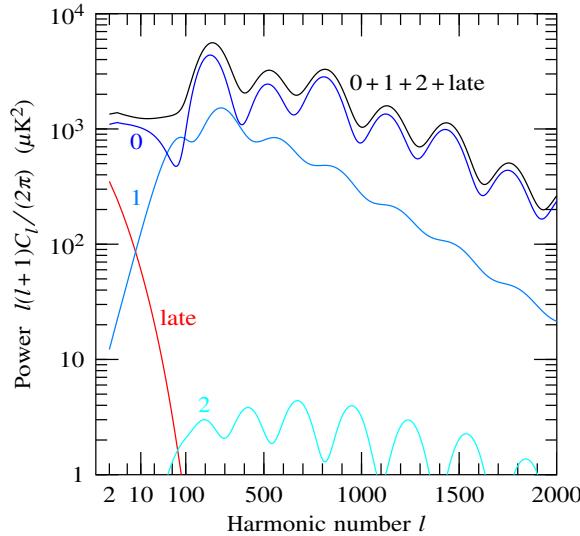


Figure 33.8 Monopole, dipole, and quadrupole contributions to the CMB power spectrum in the rapid approximation with source functions computed in the hydrodynamic approximation (top model in the right panel of Figure 33.7). The late ISW contribution is computed from integration of the derivative of the growth function over the line of sight, equations (33.15) and (33.17). 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). 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. The late ISW effect produces a slight turn-up at the smallest harmonics.

observed by astronomers on Earth today. The geometry depends on various cosmological parameters, notably the curvature  $\Omega_k$  and the Hubble parameter  $H_0$  today.

3. The power spectrum is roughly constant and dominated by the monopole at the largest scales,  $\ell \lesssim 20$ . This is the Sachs-Wolfe plateau, §33.5, a signature of a near-scale-invariant primordial power spectrum. The late ISW effect produces a slight upturn 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, §31.10, 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  $\ell$ . The decline arises partly from dissipative processes around the time of recombination, §31.7 and §31.8, and in part from the finite width of recombination.

**Exercise 33.2 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?
- 

### 33.4 Observing CMB power

The power spectrum  $C_\ell(\eta_0)$ , equation (33.32), gives an expectation value for the variance (33.31) 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  $\ell$  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  $\zeta$ . This in turn implies a Gaussian distribution of temperature fluctuations  $\Theta$  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  $C_\ell$ .

For each harmonic number  $\ell$ , there are  $2\ell + 1$  harmonics  $\Theta_{\ell m}$  with the same  $\ell$  but different  $m$ . Isotropy predicts that the expected variance is the same,  $C_\ell$ , for each  $m$ . Thus one way to estimate the variance  $C_\ell$  is to take

$$C_\ell(\text{est}) = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |\Theta_{\ell m}|^2. \quad (33.36)$$

The finite number  $2\ell + 1$  of modes at each  $\ell$  places a fundamental fractional uncertainty of  $\approx 1/\sqrt{2\ell}$  on the accuracy with which  $C_\ell$  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  $C_\ell$ , 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.

### 33.5 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 §33.2.2.

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 (33.15),

$$\Theta_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = [\Theta_0(\eta_{\text{rec}}, \mathbf{k}) + \Psi(\eta_{\text{rec}}, \mathbf{k})] j_\ell [k(\eta_0 - \eta_{\text{rec}})] . \quad (33.37)$$

At superhorizon scales  $k\eta_{\text{rec}} \ll 1$ , the radiation monopole at the time  $\eta_{\text{rec}}$  of recombination is given by the superhorizon solution  $\Theta_0 - \Phi = \zeta_\gamma$ , equation (29.23), so

$$\Theta_0(\eta_{\text{rec}}, \mathbf{k}) + \Psi(\eta_{\text{rec}}, \mathbf{k}) = \Psi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \Phi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \zeta_\gamma(\mathbf{k}) . \quad (33.38)$$

The CMB transfer function  $T_\ell(\eta, k)$  is conventionally normalized to the primordial curvature fluctuation  $\zeta(\mathbf{k})$ , equation (33.18). For adiabatic fluctuations  $\zeta$  is the same for all species; more generally,  $\zeta$  could be different for different species. For definiteness, take the simple two-component matter plus radiation model of Chapter 29, where the superhorizon potential in the late matter-dominated regime is  $\Phi(\text{late}) = -\frac{3}{5}\zeta_c$ , equation (29.65), 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  $\zeta_c$ , is

$$T_0(\eta_{\text{rec}}, k) = \frac{\Psi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \Phi_{\text{super}}(\eta_{\text{rec}}, \mathbf{k}) + \zeta_\gamma(\mathbf{k})}{\zeta_c(\mathbf{k})} \approx -\frac{6}{5} + \frac{\zeta_\gamma(\mathbf{k})}{\zeta_c(\mathbf{k})} = \begin{cases} -\frac{1}{5} & \text{adiabatic ,} \\ -\frac{6}{5} & \text{isocurvature .} \end{cases} \quad (33.39)$$

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  $\ell$ 'th harmonic today is, from equation (33.37),

$$T_\ell(\eta_0, k) = T_0(\eta_{\text{rec}}, k) j_\ell [k(\eta_0 - \eta_{\text{rec}})] . \quad (33.40)$$

The resulting CMB angular power spectrum at superhorizon scales  $k\eta_{s,\text{rec}} \ll 1$  is

$$C_\ell(\eta_0) = 4\pi T_0(\eta_{\text{rec}}, k)^2 \int_0^\infty j_\ell [k(\eta_0 - \eta_{\text{rec}})]^2 \Delta_\zeta^2(k) \frac{dk}{k} , \quad (33.41)$$

where  $T_0(\eta_{\text{rec}}, k)$ , being approximately constant, equation (33.39), has been taken outside the integral. If the primordial curvature power spectrum  $\Delta_\zeta^2(k)$  is a power law with tilt  $n$ , equation (29.134), then the integral over the squared Bessel function can be done analytically, equation (33.54b), yielding

$$C_\ell(\eta_0) = 4\pi T_0(\eta_{\text{rec}}, k)^2 \Delta_\zeta^2 [1/(\eta_0 - \eta_{\text{rec}})] U_{\ell;\ell}(n-1) . \quad (33.42)$$

For the particular case of a scale-invariant primordial power spectrum,  $n = 1$ , the CMB power spectrum  $C_\ell$

at large scales today is given by

$$\frac{\ell(\ell+1)C_\ell(\eta_0)}{2\pi} = T_0(\eta_{\text{rec}}, k)^2 \Delta_\zeta^2 [1/(\eta_0 - \eta_{\text{rec}})] \quad \text{if } n = 1 . \quad (33.43)$$

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  $C_\ell$ .

### 33.6 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 the neutrinos streamed freely. The horizon distance  $\eta_\nu$  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 32.2. Prior to recombination, the neutrino quadrupole provided the dominant source for the difference  $\Psi - \Phi$  between the scalar potentials, Figure 32.4. In Exercise 32.5 you discovered that the neutrino quadrupole causes a finite difference  $\Psi - \Phi$  even in the superhorizon initial conditions, equation (32.96).

Observationally accessible scales in the CMB or in the clustering of matter are large compared to the horizon distance  $\eta_\nu$  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 (33.15) of the radiative transfer equation is then

$$\begin{aligned} \mathcal{N}_\ell(\eta, \mathbf{k}) + \delta_{\ell 0}\Psi(\eta, \mathbf{k}) &= \int_0^\eta [\dot{\Psi}(\eta', \mathbf{k}) + \dot{\Phi}(\eta', \mathbf{k})] j_\ell[k(\eta' - \eta)] d\eta' \quad \text{ISW} \\ &\quad + [\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k})] j_\ell(-k\eta) \quad \text{monopole ,} \end{aligned} \quad (33.44)$$

which contains only Integrated Sachs-Wolfe and monopole terms. In equation (33.44), the time  $\eta_\nu$  of neutrino decoupling has been replaced by zero, since it is negligibly small. The optical depth factor has been omitted from the ISW term, and the Gaussian suppression factor from the monopole factor, since cosmological scales are so much larger than the neutrino decoupling scale  $\eta_\nu$  that neither factor is relevant.

Equation (33.44) holds at any time  $\eta$  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 non-relativistic 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.

### 33.6.1 Truncating the neutrino Boltzmann hierarchy

The integral solution (33.44) 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

$$j_\ell(y) - \frac{2\ell+3}{y} j_{\ell+1}(y) + j_{\ell+2}(y) = 0. \quad (33.45)$$

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 (33.44) vanishes. The ISW term on the first line of equation (33.44) gives a non-vanishing contribution to the combination, which is, with the identity  $1/y = y'/(y' - y) - 1/(y' - y)$ ,

$$\mathcal{N}_\ell + \delta_{\ell 0}\Psi + \frac{2\ell+3}{y}\mathcal{N}_{\ell+1} + \mathcal{N}_{\ell+2} = (2\ell+3) \int_0^y \frac{\partial[\Psi(y') + \Phi(y')]}{\partial y'} \frac{y'}{y} \frac{j_{\ell+1}(y' - y)}{(y' - y)} dy'. \quad (33.46)$$

The integrand on the right hand side of equation (33.46) is everywhere finite, and for  $y \gg y'$  is of order  $y'/y^2$  times the integrand of the ISW integral in equation (33.44). In the actual case,  $\Psi + \Phi$  varies rapidly at horizon-crossing  $y' \sim 1$ , but subsequently varies slowly. In this case the integral on the right hand side of equation (33.48) 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  $\ell$ .

If the integral on the right hand side of equation (33.46) is neglected, which becomes an increasingly good approximation at higher  $\ell$ , then

$$\mathcal{N}_{\ell+2} \approx -(\mathcal{N}_\ell + \delta_{\ell 0}\Psi) - \frac{2\ell+3}{k\eta}\mathcal{N}_{\ell+1}. \quad (33.47)$$

Ma and Bertschinger (1995) proposed truncating the neutrino Boltzmann hierarchy by using the approximation (33.47) at some suitably high harmonic number  $\ell$ . The approximation is worst around epochs where  $\Psi + \Phi$  varies rapidly, such as around horizon-crossing,  $k\eta \sim 1$ .

### 33.6.2 Approximate neutrino quadrupole

The neutrino quadrupole  $\mathcal{N}_2$  is of special interest because it is a principle source for the difference  $\Psi - \Phi$  in scalar potentials. For the quadrupole, the approximation (33.47) is

$$\mathcal{N}_2 \approx -(\mathcal{N}_0 + \Psi) - \frac{3\mathcal{N}_1}{k\eta}. \quad (33.48)$$

The approximation (33.48) is not adequate for precision modelling, but it provides the basis for the approximation of neutrinos as an imperfect fluid, equation (31.11). It is a better approximation than simply setting the neutrino quadrupole to zero,  $\mathcal{N}_2 = 0$ . The approximation (33.48) leads to a second order differential equation for the neutrino monopole, equation (31.91), which allows the behaviour of neutrinos to be explored qualitatively, Exercise 31.7.

### Exercise 33.3 Cosmic Neutrino Background.

1. Is there a Cosmic Neutrino Background? Think about whether neutrinos are relativistic or non-relativistic 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 non-relativistic 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 (33.44) with  $\eta = \eta_0$ . Without the ISW term, only the monopole term remains,

$$\mathcal{N}_\ell(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) = [\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k})] j_\ell(k\eta_0) . \quad (33.49)$$

The initial value is the superhorizon result

$$\mathcal{N}_0(0, \mathbf{k}) + \Psi(0, \mathbf{k}) = \Psi(0) + \Phi(0) + \zeta_\nu . \quad (33.50)$$

The neutrino power spectrum is proportional to the photon Sachs-Wolfe power spectrum (??), with constant of proportionality

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} = \left( \frac{\Psi(0) + \Phi(0) + \zeta_\nu}{\Psi_{\text{super}}(\eta_{\text{rec}}) + \Phi_{\text{super}}(\eta_{\text{rec}}) + \zeta_\gamma} \right)^2 . \quad (33.51)$$

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 (29.65), so

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} \approx \left( \frac{-\frac{3}{5}\zeta_r + \zeta_\nu}{-\frac{6}{5}\zeta_c + \zeta_\gamma} \right)^2 . \quad (33.52)$$

Inflation generically predicts adiabatic fluctuations with  $\zeta$ 's of all species the same, in which case

$$\frac{C_\ell^{(\nu)}}{C_\ell^{\text{SW}}} \approx \left(\frac{5}{3}\right)^2. \quad (33.53)$$

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 (33.44), 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.
- 

### 33.7 Appendix: Integrals over spherical Bessel functions

Two useful integrals over spherical Bessel functions are

$$U_\ell(z) \equiv \int_0^\infty j_\ell(y) y^z \frac{dy}{y} = \frac{2^{z-2} \sqrt{\pi} \Gamma\left[\frac{1}{2}(\ell+z)\right]}{\Gamma\left[\frac{1}{2}(\ell+3-z)\right]}, \quad (33.54a)$$

$$U_{\ell;\ell'}(z) \equiv \int_0^\infty j_\ell(y) j_{\ell'}(y) y^z \frac{dy}{y} = \frac{2^{z-3} \pi \Gamma(2-z) \Gamma\left[\frac{1}{2}(\ell+\ell'+z)\right]}{\Gamma\left[\frac{1}{2}(\ell+\ell'+4-z)\right] \Gamma\left[\frac{1}{2}(\ell-\ell'+3-z)\right] \Gamma\left[\frac{1}{2}(\ell'-\ell+3-z)\right]}, \quad (33.54b)$$

where  $\Gamma(z)$  is the Gamma function. The integrals satisfy the recurrence relations

$$\begin{aligned} U_\ell(z) &= (\ell - 2 + z) U_{\ell-1}(z-1) \\ &= \frac{\ell - 2 + z}{\ell + 1 - z} U_{\ell-2}(z), \end{aligned} \quad (33.55a)$$

$$\begin{aligned} U_{\ell;\ell'}(z) &= \frac{\ell + \ell' - 2 + z}{\ell + \ell' + 2 - z} U_{\ell-1;\ell'-1}(z) \\ &= \frac{(\ell + \ell' - 2 + z)(\ell - \ell' - 3 + z)}{2(z-2)} U_{\ell-1;\ell'}(z-1) \\ &= \frac{(\ell + \ell' - 2 + z)(\ell - \ell' - 3 + z)}{(\ell + \ell' + 2 - z)(\ell' - \ell - 1 + z)} U_{\ell-2;\ell'}(z). \end{aligned} \quad (33.55b)$$

# 34

---

## Polarization of the Cosmic Microwave Background

Well before recombination, frequent collisions drive photons into thermodynamic equilibrium. In thermodynamic equilibrium, the photon distribution is unpolarized. But, as will be seen in §34.9, 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, §34.6.2. 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, 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 (Ade, 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.

### 34.1 Photon polarization

Photons have spin one. They have two distinct spin eigenstates, right- and left-handed, in which the spin is respectively aligned and anti-aligned with the photon's direction of motion, the direction  $\hat{\mathbf{p}}$  of the photon's momentum  $\mathbf{p}$ . The general eigenstate of a photon is a complex linear combination of right- and left-handed eigenstates. In a Newman-Penrose tetrad with the 3-axis (*z*-axis) taken along the direction  $\hat{\mathbf{p}}$  of motion of the photon, the spin eigenstate of a photon is described by the complex **polarization vector**

$$\boldsymbol{\varepsilon} = \varepsilon^a \boldsymbol{\gamma}_a = \varepsilon^+ \boldsymbol{\gamma}_+ + \varepsilon^- \boldsymbol{\gamma}_- . \quad (34.1)$$

The polarization vector  $\boldsymbol{\varepsilon}$  is transverse, that is, it is orthogonal to the photon's direction of motion,

$$\hat{\mathbf{p}} \cdot \boldsymbol{\varepsilon} = 0 . \quad (34.2)$$

According to the usual rules of quantum mechanics, the squared amplitudes  $|\varepsilon^+|^2$  and  $|\varepsilon^-|^2$  of the polarization vector (34.1) represent the probabilities of observing the photon to have polarization  $\gamma_+$  or  $\gamma_-$ . For example, if a photon with polarization  $\varepsilon$  is sent through a right circularly polarized filter, then the photon will be transmitted with probability  $|\varepsilon^+|^2$ , and the transmitted photon will then be 100% right circularly polarized. The total probability of the spin states is one,

$$|\varepsilon^+|^2 + |\varepsilon^-|^2 = 1 . \quad (34.3)$$

The complex conjugate of the polarization vector is  $\varepsilon^* = \varepsilon^{+*}\gamma_- + \varepsilon^{-*}\gamma_+$  (note that complex conjugation flips  $\gamma_+ \leftrightarrow \gamma_-$ ), and the normalization condition (34.3) can be written

$$\varepsilon^* \cdot \varepsilon = 1 . \quad (34.4)$$

More generally, if a photon has polarization  $\varepsilon$ , then the probability  $P_{\varepsilon'}$  of observing it to have polarization  $\varepsilon'$  is, by the usual rules of quantum mechanics,

$$P_{\varepsilon'} = |\varepsilon'^* \cdot \varepsilon|^2 . \quad (34.5)$$

A photon in a pure  $\gamma_+$  eigenstate (i.e. with polarization vector  $e^{-i\phi}\gamma_+$  where  $e^{-i\phi}$  is some arbitrary phase factor) is right circularly polarized, while a photon in a pure  $\gamma_-$  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  $\varepsilon_1 = e^{-i\phi}(\gamma_+ + \gamma_-)/\sqrt{2}$  is linearly polarized in the 1-direction ( $x$ -direction), while a photon with polarization vector

$$\varepsilon_\chi = e^{-i\phi} \frac{e^{-i\chi}\gamma_+ + e^{i\chi}\gamma_-}{\sqrt{2}} \quad (34.6)$$

is linearly polarized along a direction rotated right-handedly by angle  $\chi$  from the 1-axis. A polarization angle of  $\chi = \pi$  flips the sign of  $\varepsilon_\chi$ , equivalent to changing its phase, so the polarization angle is determined only modulo  $\pi$ . A photon that is a superposition of unequal non-zero magnitudes of right and left polarizations is said to be elliptically polarized.

## 34.2 Spin sign convention

The standard physics convention is that a particle is said to have spin  $s$  if it changes by  $e^{-is\chi}$  under a right-handed rotation by angle  $\chi$  about a preferred direction (cf. §13.7). In quantum mechanics, the preferred direction is set by an observer who measures spin along a certain direction. In the photon-baryon fluid under consideration, photons are “observed” by successive electrons off which they scatter.

The polarization vector  $\varepsilon$  of a photon is always transverse to its direction of motion  $\hat{p}$ , equation (34.2). In a Newman-Penrose tetrad with the 3-axis ( $z$ -axis) along the direction  $\hat{p}$  of motion, the vectors  $\gamma_\pm$  have spin-weights  $\pm 1$ , because they change by a phase  $e^{\mp i\chi}$  under a right-handed rotation by angle  $\chi$  about the 3-axis. The contravariant coefficients  $\varepsilon^\pm$  (raised indices) change as  $e^{\pm i\chi}$ , oppositely to  $\gamma_\pm$  (lowered indices),

ensuring that the abstract polarization vector  $\boldsymbol{\varepsilon}$  is a tetrad scalar, invariant under rotations. The covariant coefficients  $\varepsilon_{\pm}$  (lowered indices) are

$$\varepsilon_{\pm} = \boldsymbol{\gamma}_{\pm} \cdot \boldsymbol{\varepsilon} = \varepsilon^{\mp}, \quad (34.7)$$

which have spin-weight  $\pm 1$  opposite to that of the contravariant coefficients  $\varepsilon^{\pm}$ . It is convenient to work with the covariant components  $\varepsilon_{\pm}$  of the polarization vector, because then the sign of the index agrees with the sign of the spin.

### 34.3 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 (34.6) 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  $\boldsymbol{\varepsilon}$  with occupation numbers  $f(\boldsymbol{\varepsilon})$ . Then the density matrix  $\mathbf{f}$  may be defined by the tensor

$$\mathbf{f} \equiv \sum_{\boldsymbol{\varepsilon}} f(\boldsymbol{\varepsilon}) \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon}^*. \quad (34.8)$$

In Newman-Penrose components, the density matrix  $f_{ab}$  is the  $2 \times 2$  complex matrix

$$f_{ab} \equiv \sum_{\boldsymbol{\varepsilon}} f(\boldsymbol{\varepsilon}) \varepsilon_a \varepsilon_b^*, \quad (34.9)$$

where the bar on the index  $\bar{b}$  indicates that the complex conjugate index is to be taken (that is,  $\bar{+} = -$  and  $\bar{-} = +$ ), because complex conjugation flips spinor indices,  $\boldsymbol{\gamma}_{\pm}^* = \boldsymbol{\gamma}_{\mp}$ . In terms of its components, the density matrix is  $\mathbf{f} = f_{ab} \boldsymbol{\gamma}^a \otimes \boldsymbol{\gamma}^b = f_{ab} \boldsymbol{\gamma}_{\bar{a}} \otimes \boldsymbol{\gamma}_{\bar{b}}$ . The density matrix is Hermitian,

$$f_{ab}^* = f_{\bar{b}\bar{a}}. \quad (34.10)$$

If the system of photons is measured along the polarization direction  $\boldsymbol{\varepsilon}$ , then the occupation number of photons with that polarization will be found to be, in accordance with equation (34.5),

$$f(\boldsymbol{\varepsilon}) = \varepsilon^{\bar{a}*} \varepsilon^b f_{ab}. \quad (34.11)$$

#### 34.3.1 Physical interpretation of the photon density matrix

Since the complex  $2 \times 2$  photon density matrix  $f_{ab}$  is Hermitian, equation (34.10), it is diagonalizable with 2 real eigenvalues. The form (34.8) 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.

The Newman-Penrose components of the  $2 \times 2$  density matrix  $f_{ab}$  comprise a complex scalar (spin 0)  $f_{+-}$  and a complex tensor (spin 2)  $f_{++}$ . Complex conjugation flips the indices  $+ \leftrightarrow -$ ,

$$f_{+-}^* = f_{-+}, \quad f_{++}^* = f_{--}. \quad (34.12)$$

Since  $\boldsymbol{\varepsilon}^* \cdot \boldsymbol{\varepsilon} = 1$  for each photon, the trace of the density matrix (34.9) counts the total number of photons. The unpolarized scalar occupation number  $f$  defined earlier, equation (30.28), equals one when there is one photon in each of the two polarization states, so the trace equals twice the unpolarized occupation number  $f$ ,

$$2f \equiv f_a^a = \sum_{\gamma} f(\boldsymbol{\varepsilon}_{\gamma}). \quad (34.13)$$

The complex spin-0 component  $f_{+-}$  ( $= f^{-+}$ , the coefficient of  $\boldsymbol{\gamma}_- \otimes \boldsymbol{\gamma}_+$ ) of the density matrix measures the intensity of left circularly polarized light, while its complex conjugate  $f_{-+}$  ( $= f^{+-}$ , the coefficient of  $\boldsymbol{\gamma}_+ \otimes \boldsymbol{\gamma}_-$ ) measures the intensity of right circularly polarized light. The sum  $f_{+-} + f_{-+}$ , which is twice the real part of  $f_{+-}$ , measures the total intensity of light in both polarizations, while the difference  $f_{+-} - f_{-+}$ , which is twice the imaginary part of  $f_{+-}$ , measures the net circularly polarized intensity, the excess of left over right 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  $\chi$ , equation (34.6), contributes a density matrix

$$\boldsymbol{\varepsilon}_{\chi} \otimes \boldsymbol{\varepsilon}_{\chi}^* = \frac{1}{2} (\boldsymbol{\gamma}_+ \otimes \boldsymbol{\gamma}_- + \boldsymbol{\gamma}_- \otimes \boldsymbol{\gamma}_+) + \frac{1}{2} e^{-2i\chi} \boldsymbol{\gamma}_+ \otimes \boldsymbol{\gamma}_+ + \frac{1}{2} e^{2i\chi} \boldsymbol{\gamma}_- \otimes \boldsymbol{\gamma}_-. \quad (34.14)$$

The first terms on the right hand side of equation (34.14) contribute a trace of one, which is as it should be for a single photon. The remaining terms imply  $f_{++} = \frac{1}{2} e^{2i\chi}$  ( $= f^{--}$ , the coefficient of  $\boldsymbol{\gamma}_- \otimes \boldsymbol{\gamma}_-$ ), with complex conjugate  $f_{--} = \frac{1}{2} e^{-2i\chi}$  ( $= f^{++}$ , the coefficient of  $\boldsymbol{\gamma}_+ \otimes \boldsymbol{\gamma}_+$ ). Twice the amplitude of  $f_{++}$  gives the degree of linear polarization, which here is one (100% linearly polarized), while the phase  $2\chi$  of  $f_{++}$  measures the angle  $\chi$  by which the direction of polarization is rotated right-handedly from the 1-axis ( $x$ -axis).

**Concept question 34.1 Elliptically polarized light.** Can a beam of elliptically polarized light be distinguished from a sum of beams of linearly polarized and circularly polarized light?

### 34.3.2 Relation to Stokes parameters

The 4 components of the polarization density matrix  $f_{ab}$  are related to the 4 conventional real Stokes parameters  $I$ ,  $Q$ ,  $U$ , and  $V$  by

$$2f_{+-} = I + iV , \quad (34.15a)$$

$$2f_{++} = Q + iU . \quad (34.15b)$$

The Stokes parameter  $I$  is the total intensity, the trace of the density matrix,  $I = f_{+-} + f_{-+} = 2 \operatorname{Re} f_{+-} = 2f$ .

## 34.4 Temperature fluctuation for polarized photons

Previously, the perturbation to the unpolarized scalar occupation number  $f = \operatorname{Re} f_{+-}$  was expressed in terms of the temperature fluctuation,  $\Theta \equiv \delta T/T$ , equation (??). The temperature fluctuation  $\Theta \equiv \Theta_{ab} \boldsymbol{\gamma}^a \otimes \boldsymbol{\gamma}^b$  including polarization can be defined similarly in terms of the density matrix  $f_{ab}$ , equation (34.9), which is the generalization of the occupation number to include polarization,

$$f_{ab} \equiv \frac{\partial \overset{0}{f}_\gamma}{\partial \ln T} \Theta_{ab} . \quad (34.16)$$

The trace of  $\Theta_{ab}$  is twice the scalar temperature fluctuation,  $\Theta_a^a = 2\Theta$ . The trace-free part of  $\Theta_{ab}$  describes the polarized temperature fluctuation.

Often it is convenient to abbreviate the  $ab$  indices to a single spin index  $s$  running over 0, 2,

$${}_0\Theta \equiv \Theta_{+-} , \quad {}_2\Theta \equiv \Theta_{++} . \quad (34.17)$$

The spin subscript  $s$  is positioned on the left to distinguish it from harmonic indices  $\ell m$ . The spin components  ${}_s\Theta$  are complex. Their complex conjugates can be denoted with a negative index,

$${}_{-0}\Theta \equiv {}_0\Theta^* = \Theta_{-+} , \quad {}_{-2}\Theta \equiv {}_2\Theta^* = \Theta_{--} . \quad (34.18)$$

It will be seen in §34.9 that Thomson scattering generates linear polarization but not circular polarization. In this case the circularly polarized component vanishes,  $\operatorname{Im} \Theta_{+-} = 0$  (Stokes parameter  $V = 0$ ), and the spin-0 component  ${}_0\Theta$  is real and equal to the unpolarized temperature fluctuation  $\Theta$ ,

$${}_0\Theta = \Theta . \quad (34.19)$$

The polarized temperature fluctuation  ${}_s\Theta(\eta, \mathbf{x}, \hat{\mathbf{p}}, \chi)$  depends not only on conformal time  $\eta$ , comoving position  $\mathbf{x}$ , and photon direction  $\hat{\mathbf{p}}$ , but also on the angle  $\chi$  of rotation about the photon direction  $\hat{\mathbf{p}}$ . The spin- $s$  component of the temperature fluctuation varies as  ${}_s\Theta(\eta, \mathbf{x}, \hat{\mathbf{p}}, \chi) \propto e^{-is\chi}$  under a right-handed rotation by angle  $\chi$  about  $\hat{\mathbf{p}}$ .

### 34.5 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  $\partial_m$  on the left hand side of the Boltzmann equation (32.8) should therefore be replaced by the covariant derivative  $D_m f_{ab}$ ,

$$D_m f_{ab} = \partial_m f_{ab} - \Gamma_{am}^c f_{cb} - \Gamma_{bm}^c f_{ac} . \quad (34.20)$$

However, the polarized (trace-free) part of  $f_{ab}$  is of linear order, and the tetrad-frame connections  $\Gamma_{abm}$  with  $a, b$  both spatial are all of linear order, equation (28.23) (in any gauge), so the connection terms on the right hand side of equation (34.20) 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 (??) in conformal Newtonian gauge. The gravitational  $G$  term in this equation arises, equation (32.21), from the redshifting of photons in the unperturbed photon distribution  $\overset{0}{f}$ . 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

$$\boxed{\dot{\Theta} - ik\mu \Theta - G = C[\Theta]} , \quad (34.21a)$$

$$\boxed{{}_2\dot{\Theta} - ik\mu {}_2\Theta = C[{}_2\Theta]} . \quad (34.21b)$$

The collision terms  $C[s\Theta]$  that arise from non-relativistic electron-photon (Thomson) scattering are calculated in §34.9. In conformal Newtonian gauge, and including not only scalar ( $m = 0$ ) but also vector ( $|m| = 1$ ) and tensor ( $|m| = 2$ ) potentials from Exercise 32.3, equation (32.26), the gravitational redshift term  $G$  in the unpolarized Boltzmann equation (34.21a) is

$$G(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \dot{\Phi} + ik\mu\Psi + ik\mu\hat{\mathbf{p}} \cdot \mathbf{W} + \hat{p}^a \hat{p}^b \dot{h}_{ab} . \quad (34.22)$$

### 34.6 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  $_sY_{\ell m}$ , §34.15 (Seljak and Zaldarriaga, 1997; Zaldarriaga and Seljak, 1997; Hu and White, 1997). With the normalization conventional in CMB studies, consistent with the normalization (32.46) of the scalar ( $m = 0$ ) fluctuation  $\Theta$ , the harmonic expansion of the polarized temperature fluctuation  $_s\Theta$  is, with  $\hat{\mathbf{k}}$  along the

3-direction ( $z$ -direction),

$$\begin{aligned} {}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) &= \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} \sqrt{4\pi(2\ell+1)} {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}Y_{\ell m}^*(\hat{\mathbf{p}}, \chi) \\ &= \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} (2\ell+1) {}_s\Theta_{\ell m}(\eta, \mathbf{k}) D_{\ell m s}(\phi, \theta, \chi), \end{aligned} \quad (34.23)$$

where  ${}_sY_{\ell m}(\hat{\mathbf{p}}, \chi)$  are the spin-weighted spherical harmonics defined by equation (34.86), and  $D_{\ell m s}(\phi, \theta, \chi)$  are the Wigner rotation matrices discussed in §34.15.1. 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 expansion (34.23) 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 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 (34.95), ensuring reality conditions (34.35) on the harmonic coefficients that match those on the Newman-Penrose components of the gravitational potentials. The factor of  $e^{-is\chi}$  makes explicit the spin factor that Hu and White (1997) absorb into basis vectors  $\boldsymbol{\gamma}_a \otimes \boldsymbol{\gamma}_b$  of the polarization matrix.

### 34.6.1 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 (34.110) 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 (34.21), common to all spins  $s$  and all harmonics  $\ell m$ , is

$$({}_s\dot{\Theta} - ik\mu_s\Theta)_{\ell m} = {}_s\dot{\Theta}_{\ell m} + k \left[ \frac{\kappa_{\ell m s}}{2\ell+1} {}_s\Theta_{\ell-1,m} + \frac{ism}{\ell(\ell+1)} {}_s\Theta_{\ell m} - \frac{\kappa_{\ell+1,m s}}{2\ell+1} {}_s\Theta_{\ell+1,m} \right], \quad (34.24)$$

where the coefficients  $\kappa_{\ell m n}$  are given by equation (34.111). The harmonic expansion of the gravitational term  $G$ , equation (34.22), in the unpolarized Boltzmann equation is

$$G(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \sum_{\ell=0}^2 \sum_{m=-\ell}^{\ell} (-i)^{\ell+m} (2\ell+1) G_{\ell m}(\eta, \mathbf{k}) D_{\ell m 0}(\phi, \theta), \quad (34.25)$$

with

$$G_{00} = \dot{\Phi} , \quad (34.26a)$$

$$G_{10} = -\frac{k}{3}\Psi , \quad (34.26b)$$

$$G_{2,\pm 1} = -\frac{k}{5\sqrt{3}}W_{\pm} , \quad (34.26c)$$

$$G_{2,\pm 2} = \frac{\sqrt{2}}{5\sqrt{3}}h_{\pm\pm} , \quad (34.26d)$$

where  $W_{\pm} \equiv \frac{1}{\sqrt{2}}(W_x \pm iW_y)$  are the spin-weight  $\pm 1$  components of the vector perturbation  $W_a$ , equation (26.22), and  $h_{\pm\pm} \equiv h_{xx} \pm i h_{xy}$  are the spin-weight  $\pm 2$  components of the tensor perturbation  $h_{ab}$ , equation (26.23).

### 34.6.2 Electric and magnetic parts of the polarized photon distribution

The rotation matrices  $D_{\ell ms}$  transform as (34.96) under a parity transformation. Parity eigenstates of the rotation matrices are

$$\begin{aligned} (1 \pm P)D_{\ell ms}(\phi, \theta, \chi) &= D_{\ell ms}(\phi, \theta, \chi) \pm D_{\ell ms}(\phi + \pi, \pi - \theta, -\chi) \\ &= D_{\ell ms}(\phi, \theta, \chi) \pm (-)^{\ell}D_{\ell m, -s}(\phi, \theta, \chi) . \end{aligned} \quad (34.27)$$

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  $(-)^{\ell+1}$ ,

$${}_{\pm s}\Theta_{\ell m} = {}_sE_{\ell m} \pm i {}_sB_{\ell m} . \quad (34.28)$$

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. The magnetic spin 0 fluctuation vanishes,  ${}_0B_{\ell m} = 0$ , because there is no circular polarization, so only the spin  $\pm 2$  temperature fluctuation has a magnetic part. There being no ambiguity, the spin index  $s$  is dropped on  ${}_sE$  and  ${}_sB$  for the spin  $\pm 2$  fluctuation,

$${}_{\pm 2}\Theta_{\ell m} = E_{\ell m} \pm iB_{\ell m} . \quad (34.29)$$

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 (34.35). Because of the reality conditions, the split (34.29) of the polarized temperature fluctuation into electric and magnetic components is essentially equivalent to a Stokes split  ${}_{\pm 2}\Theta = Q \pm iU$  with real  $Q$  and  $U$ .

Resolved into parity eigenstates, the Boltzmann equations (34.21) for the unpolarized and polarized temperature fluctuations are

$$(\dot{\Theta} - ik\mu\Theta)_{\ell m} = \dot{\Theta}_{\ell m} + k \left( \frac{\kappa_{\ell m 0}}{2\ell + 1} \Theta_{\ell-1, m} - \frac{\kappa_{\ell+1, m 0}}{2\ell + 1} \Theta_{\ell+1, m} \right) = G_{\ell m} + C[\Theta_{\ell m}] , \quad (34.30a)$$

$$(\dot{E} - ik\mu E)_{\ell m} = \dot{E}_{\ell m} + k \left( \frac{\kappa_{\ell m 2}}{2\ell + 1} E_{\ell-1, m} - \frac{2m}{\ell(\ell + 1)} B_{\ell m} - \frac{\kappa_{\ell+1, m 2}}{2\ell + 1} E_{\ell+1, m} \right) = C[E_{\ell m}] , \quad (34.30b)$$

$$(\dot{B} - ik\mu B)_{\ell m} = \dot{B}_{\ell m} + k \left( \frac{\kappa_{\ell m 2}}{2\ell + 1} B_{\ell-1, m} + \frac{2m}{\ell(\ell + 1)} E_{\ell m} - \frac{\kappa_{\ell+1, m 2}}{2\ell + 1} B_{\ell+1, m} \right) = C[B_{\ell m}] . \quad (34.30c)$$

The equations for  $m < 0$  are the same as those for  $m > 0$  with  $B$  flipped in sign.

### 34.7 Vector and tensor Einstein equations

The photon energy-momentum depends only on the unpolarized photon distribution  $\Theta$ . The vector components of the photon energy-momenta  $T_{\gamma}^{mn}$  are given in terms of unpolarized multipole moments  $\Theta_{\ell m}$  by, from equation (??),

$$T_{\gamma}^{0\pm} = 4\bar{\rho}_{\gamma}\Theta_{1,\pm 1} , \quad (34.31a)$$

$$T_{\gamma}^{3\pm} = -i\frac{4}{\sqrt{3}}\bar{\rho}_{\gamma}\Theta_{2,\pm 1} , \quad (34.31b)$$

while the tensor components are

$$T_{\gamma}^{\pm\pm} = \frac{4\sqrt{2}}{\sqrt{3}}\bar{\rho}_{\gamma}\Theta_{2,\pm 2} . \quad (34.32)$$

The vector Einstein equations are, equations (28.41),

$$-k^2 W_{\pm} = -16\pi Ga^2 (\bar{\rho}_c v_{c,\pm} + \bar{\rho}_b v_{b,\pm} + 4\bar{\rho}_{\gamma}\Theta_{1,\pm 1} + 4\bar{\rho}_{\nu}\mathcal{N}_{1,\pm 1}) , \quad (34.33a)$$

$$k \left( \frac{\partial}{\partial\eta} + 2\frac{\dot{a}}{a} \right) W_{\pm} = \frac{64}{\sqrt{3}}\pi Ga^2 (\bar{\rho}_{\gamma}\Theta_{2,\pm 1} + \bar{\rho}_{\nu}\mathcal{N}_{2,\pm 1}) , \quad (34.33b)$$

where  $v_{\pm} \equiv \frac{1}{\sqrt{2}}(v_x \pm i v_y)$  are the spin-weight  $\pm 1$  components of the bulk velocity of a species. Only one of the two equations (34.33) is needed; the other is satisfied automatically as long as total energy-momentum is conserved. The tensor Einstein equations are, equation (28.43),

$$\left( \frac{\partial^2}{\partial\eta^2} + 2\frac{\dot{a}}{a}\frac{\partial}{\partial\eta} - \nabla^2 \right) h_{\pm\pm} = -\frac{32\sqrt{2}}{\sqrt{3}}\pi Ga^2 (\bar{\rho}_{\gamma}\Theta_{2,\pm 2} + \bar{\rho}_{\nu}\mathcal{N}_{2,\pm 2}) . \quad (34.34)$$

### 34.8 Reality conditions on the polarized photon distribution

The initial photon distribution well before recombination is in thermodynamic equilibrium and therefore unpolarized. The Einstein scalar (32.7), vector (34.33), and tensor (34.34) equations show that the scalar  $\Psi$

and  $\Phi$ , vector  $W_{\pm}$ , and tensor  $h_{\pm\pm}$  gravitational potentials are sourced by unpolarized 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 (34.26). 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  $\Theta_{2m}$ , equation (34.49d). The polarized Boltzmann hierarchy (34.30) 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  $\Psi$  and  $\Phi$  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

$$\Theta_{\ell m}^* = \Theta_{\ell, -m}, \quad E_{\ell m}^* = E_{\ell, -m}, \quad B_{\ell m}^* = -B_{\ell, -m}. \quad (34.35)$$

In particular, all scalar ( $m = 0$ ) fluctuations are real. The scalar magnetic fluctuation vanishes,  $B_{\ell 0} = 0$ .

**Concept question 34.2 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 ( $m = 2$ ) modes, and these source photon or neutrino multipoles of the same  $m$ , equations (34.26). Thomson scattering sources polarized fluctuations, but leaves the azimuthal mode  $m$  remains unchanged.

## 34.9 Polarized Thomson scattering

The squared invariant amplitude  $|\mathcal{M}|^2$  for electron-photon scattering by non-relativistic electrons with random spins, in which the initial photon polarization state is  $\boldsymbol{\varepsilon}$  and the final polarization state  $\boldsymbol{\varepsilon}'$ , is, generalizing the unpolarized expression (32.54),

$$|\mathcal{M}|^2 = (8\pi\alpha)^2 |\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2, \quad (34.36)$$

where  $\alpha \equiv e^2/(\hbar c)$  is the fine-structure constant. The differential cross-section  $d\sigma_T/d\omega'$  for polarized Thomson scattering is related to the squared invariant amplitude  $|\mathcal{M}|^2$  by, in units  $c = \hbar = 1$ ,

$$\frac{d\sigma_T}{d\omega'} = \frac{|\mathcal{M}|^2}{(8\pi m_e)^2} = \frac{\alpha^2}{m_e^2} |\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2 = r_e^2 |\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2 = \frac{3}{8\pi} \sigma_T |\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2, \quad (34.37)$$

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.

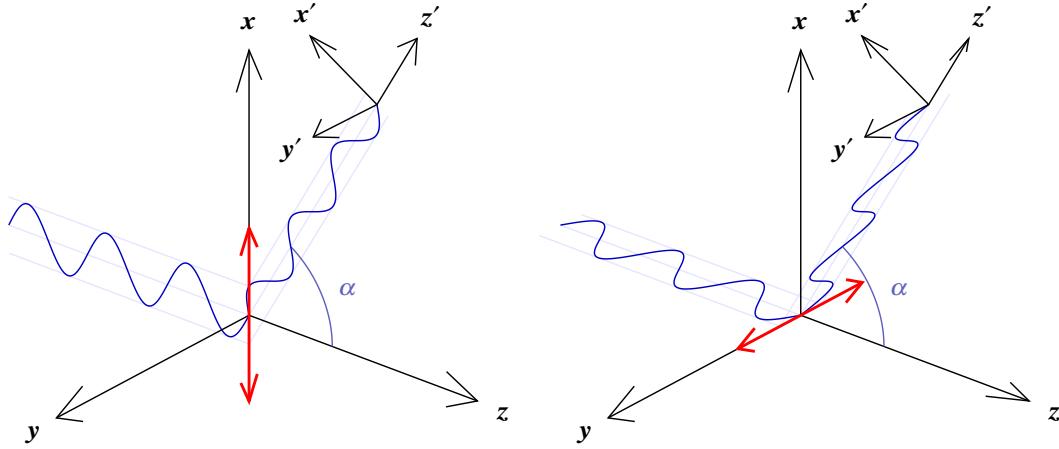


Figure 34.1 Polarized light incident in the  $z$  direction (wiggly blue line) on an electron causes the electron to oscillate in the direction  $\boldsymbol{\epsilon}$  of polarization (red arrow). The oscillating electron emits scattered light at the same frequency (wiggly blue line). For incident light with polarization vector  $\boldsymbol{\epsilon}_x$  in the scattering plane (left), the polarization vector  $\boldsymbol{\epsilon}_{x'}$  of the scattered light is rotated by the scattering angle  $\alpha$ , and reduced in amplitude by a factor  $\cos \alpha$ , so that  $\boldsymbol{\epsilon}_x \cdot \boldsymbol{\epsilon}_{x'} = \cos \alpha$ . On the other hand, for incident light with polarization vector  $\boldsymbol{\epsilon}_y$  orthogonal to the scattering plane (right), the polarization vector  $\boldsymbol{\epsilon}_{y'}$  of the scattered light is the same as that of the incident light, so that  $\boldsymbol{\epsilon}_y \cdot \boldsymbol{\epsilon}_{y'} = 1$ .

The collision integral  $C[\Theta]$  for unpolarized Thomson scattering was given previously by equation (32.73). The same equation holds for polarized scattering, except that the scalar temperature fluctuation  $\Theta$  is replaced by the set of 4 quantities  $\Theta_{ab}$ , and  $|\mathcal{M}|^2$  becomes a  $4 \times 4$  matrix (which reduces to a  $3 \times 3$  matrix if there is no circular polarization, as will be found to be the case below). The collision integral (32.73) generalizes to

$$C[\Theta_{ab}(\hat{\mathbf{p}}, \chi)] = \frac{\bar{n}_e a}{16\pi m_e^2} \int [|\mathcal{M}|^2]_{ab}^{a'b'} [(\hat{\mathbf{p}} - \hat{\mathbf{p}}') \cdot \mathbf{v}_b - \Theta_{a'b'}(\hat{\mathbf{p}}, \chi) + \Theta_{a'b'}(\hat{\mathbf{p}}', \chi')] \frac{d\sigma' d\chi'}{4\pi 2\pi}, \quad (34.38)$$

the integration over  $d\chi'/2\pi$  enforcing orthogonality of spin states. 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 (32.64), so the integral over the baryon velocity term yields the same result as in the unpolarized case. The factor  $\Theta_{a'b'}(\hat{\mathbf{p}})$  is independent of  $\hat{\mathbf{p}}'$ , and can be taken outside the integral. The collision integral (34.38) thus reduces by the same manipulations (32.74)–(32.77) as in the unpolarized case to

$$C[\Theta_{ab}(\hat{\mathbf{p}}, \chi)] = |\dot{\tau}| \left\{ \hat{\mathbf{p}} \cdot \mathbf{v}_b \delta_{ab,0} - \Theta_{ab}(\hat{\mathbf{p}}, \chi) + \int \frac{3}{2} [|\boldsymbol{\epsilon}' \cdot \boldsymbol{\epsilon}|^2]_{ab}^{a'b'} \Theta_{a'b'}(\hat{\mathbf{p}}', \chi') \frac{d\sigma' d\chi'}{4\pi 2\pi} \right\}, \quad (34.39)$$

generalizing the unpolarized collision integral (32.77). The Kronecker delta  $\delta_{ab,0}$  is to be interpreted as equal to 1 for the unpolarized collision term  $C[\Theta]$ , and zero otherwise.

Choose axes such that 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  $x-z$  plane, as illustrated in Figure 34.1. If the scattering angle is  $\alpha$ , then

the scalar products of polarization vectors in the  $x$  and  $y$  directions are  $\boldsymbol{\varepsilon}'_x \cdot \boldsymbol{\varepsilon}_x = \cos \alpha$ ,  $\boldsymbol{\varepsilon}'_x \cdot \boldsymbol{\varepsilon}_y = 0$ ,  $\boldsymbol{\varepsilon}'_y \cdot \boldsymbol{\varepsilon}_y = 1$ . In this special frame aligned with the scattering plane, the integrand on the right hand side of the collision integral (34.39) is

$$\frac{3}{2} [|\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2]_{ab}^{a'b'} \Theta_{a'b'}(\hat{\mathbf{p}}') = \frac{3}{2} \begin{pmatrix} \cos^2 \alpha & 0 & 0 \\ 0 & \cos \alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Theta_{xx} \\ \Theta_{xy} \\ \Theta_{yy} \end{pmatrix}. \quad (34.40)$$

Since  $\Theta_{ab}$  is Hermitian,  $\Theta_{xx}$  and  $\Theta_{yy}$  are real, but  $\Theta_{xy}$  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 (34.40) shows that if light incident in a given direction is initially unpolarized ( $\Theta_{ab}$  isotropic), then the scattered light will be polarized ( $\Theta_{ab}$  anisotropic). But if  $\Theta_{xy}$  is initially real, it remains real after a scattering event. Since the imaginary part of  $\Theta_{xy}$  is associated with circular polarization, Thomson scattering generates linear polarization, but not circular polarization.

In Newman-Penrose components, the absence of circularly polarized light implies that  $\Theta_{+-} = \Theta_{-+} = \Theta$ , where  $\Theta$  is the unpolarized temperature fluctuation. In Newman-Penrose components, equation (34.40) becomes

$$\begin{aligned} \frac{3}{2} [|\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2]_{ab}^{a'b'} \Theta_{a'b'}(\hat{\mathbf{p}}') &= \frac{3}{4} \begin{pmatrix} 1 + \cos^2 \alpha & -\frac{1}{2} \sin^2 \alpha & -\frac{1}{2} \sin^2 \alpha \\ -\sin^2 \alpha & \frac{1}{2}(1 + \cos \alpha) & \frac{1}{2}(1 - \cos \alpha) \\ -\sin^2 \alpha & \frac{1}{2}(1 - \cos \alpha) & \frac{1}{2}(1 + \cos \alpha) \end{pmatrix} \begin{pmatrix} \Theta \\ \Theta_{++} \\ \Theta_{--} \end{pmatrix} \\ &= \frac{3}{2} \begin{pmatrix} \frac{2}{3} d_{000} + \frac{1}{3} d_{200} & -\sqrt{\frac{1}{6}} d_{202} & -\sqrt{\frac{1}{6}} d_{20-2} \\ -\sqrt{\frac{2}{3}} d_{220} & d_{222} & d_{22-2} \\ -\sqrt{\frac{2}{3}} d_{2-20} & d_{2-22} & d_{2-2-2} \end{pmatrix} \begin{pmatrix} \Theta \\ \Theta_{++} \\ \Theta_{--} \end{pmatrix}, \end{aligned} \quad (34.41)$$

where the functions  $d_{lmn}(\alpha)$  are the polar part of the Wigner rotation matrix, equation (34.92). Equation (34.41) can be written

$$\frac{3}{2} [|\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2]_s^{s'} \Theta = \Theta \delta_{s0} + \sum_{s'} (-i)^{s'-s} c_{ss'} d_{2ss'}(\alpha) \Theta, \quad (34.42)$$

with  $s'$  summed over  $0, 2, -2$ . The first term on the right hand side of equation (34.42) is the unpolarized contribution, while the remainder is the polarized contribution. The coefficients  $c_{ss'}$  encapsulate the polarization structure of Thomson scattering,

$$c_{ss'} = \begin{pmatrix} \frac{1}{2} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} \\ \sqrt{\frac{3}{2}} & \frac{3}{2} & \frac{3}{2} \\ \sqrt{\frac{3}{2}} & \frac{3}{2} & \frac{3}{2} \end{pmatrix}. \quad (34.43)$$

The coefficients depend only on the absolute value of the spins,  $c_{ss'} = c_{|s||s'|}$ .

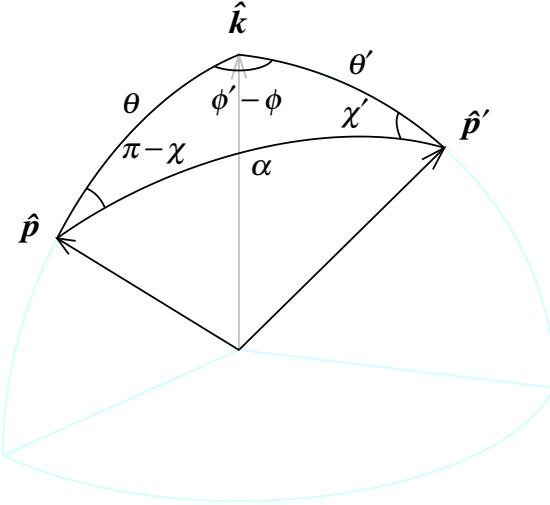


Figure 34.2 Angles between photon momentum  $\hat{p}$ , scattered photon momentum  $\hat{p}'$ , and wavevector  $\hat{k}$ .

The addition theorem (34.114) allows the rotation matrix  $d_{2ss'}(\alpha)$  from the  $\hat{p}'$  frame into the  $\hat{p}$  frame to be written as a product of rotation matrices from the  $\hat{p}'$  frame into the  $\hat{k}$  frame into the  $\hat{p}$  frame,

$$\frac{3}{2} [|\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2]_{s s'}^{s'} \Theta(\hat{p}', \chi') = \Theta(\hat{p}') \delta_{s0} + \sum_{s'} (-i)^{s'-s} c_{ss' s'} \Theta(\hat{p}', \chi') \sum_{m=-2}^2 D_{2ms}(\phi, \theta, \chi) D_{2ms'}^*(\phi', \theta', \chi') . \quad (34.44)$$

Figure 34.2 illustrates the various angles involved in transforming from the scattering frame to a frame where the wavevector  $\hat{k}$  is along the  $z$ -axis.

When  ${}_{s'} \Theta(\hat{p}', \chi')$  in equation (34.44) is expanded in rotation matrices, equation (34.23), the orthogonality of the rotation matrices, equation (34.107), makes the integration over directions  $\hat{p}'$  and  $\chi'$  straightforward, yielding

$$\int \frac{3}{2} [|\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2]_{s s'}^{s'} \Theta(\hat{p}', \chi') \frac{d\Omega'}{4\pi} \frac{d\chi'}{2\pi} = \Theta_{00} \delta_{s0} + \sum_{m=-2}^2 (-i)^{2+m-s} D_{2ms}(\phi, \theta, \chi) \sum_{s'} c_{ss' s'} \Theta_{2m} . \quad (34.45)$$

The sum over  $c_{ss'}$  in equation (34.45) is

$$\sum_{s'} c_{ss' s'} \Theta_{2m} = c_{s0} \Theta_{2m} + 2c_{s2} E_{2m} = c_{s0} (\Theta_{2m} + \sqrt{6} E_{2m}) . \quad (34.46)$$

The collision integral (34.39) is then

$$C[s\Theta(\hat{\mathbf{p}}, \chi)] = |\dot{\tau}| \left[ \hat{\mathbf{p}} \cdot \mathbf{v}_b \delta_{s0} - {}_s\Theta(\hat{\mathbf{p}}) + \Theta_{00} \delta_{s0} + c_{s0} \sum_{m=-2}^2 (-i)^{2+m-s} D_{2ms}(\phi, \theta, \chi) (\Theta_{2m} + \sqrt{6} E_{2m}) \right]. \quad (34.47)$$

Expanded in harmonics, the collision integral is

$$C[s\Theta] = \sum_{\ell=|s|}^{\infty} \sum_{m=-\ell}^{\ell} (-i)^{\ell+m-s} (2\ell+1) C[s\Theta_{\ell m}] D_{\ell m s}(\phi, \theta, \chi), \quad (34.48)$$

with collision terms for the individual harmonics being

$$C[\Theta_{00}] = 0, \quad (34.49a)$$

$$C[\Theta_{1m}] = |\dot{\tau}| \left( -\Theta_{1m} + \frac{1}{3} v_{b,m} \right), \quad (34.49b)$$

$$C[\Theta_{2m}] = |\dot{\tau}| \left[ -\Theta_{2m} + \frac{1}{10} (\Theta_{2m} + \sqrt{6} E_{2m}) \right], \quad (34.49c)$$

$$C[E_{2m}] = |\dot{\tau}| \left[ -E_{2m} + \frac{\sqrt{6}}{10} (\Theta_{2m} + \sqrt{6} E_{2m}) \right], \quad (34.49d)$$

$$C[B_{2m}] = -|\dot{\tau}| B_{2m}, \quad (34.49e)$$

$$C[s\Theta_{\ell m}] = -|\dot{\tau}| {}_s\Theta_{\ell m} \quad (\ell > 2). \quad (34.49f)$$

Scalar, vector, and tensor modes correspond to those with respectively  $m = 0$ ,  $\pm 1$ , and  $\pm 2$ .

**Exercise 34.3 Photon diffusion including polarization.** A diffusion approximation for the photon quadrupole fluctuation  $\Theta_2$  is obtained by neglecting time derivatives,  $\dot{\Theta}_2 = 0$ , and higher order multipoles,  $\Theta_3 = 0$ , in the Boltzmann equation for  $\Theta_2$ . Without polarization, this led to the quadrupole (31.67) in the unpolarized Boltzmann equation (32.81c). Derive the diffusion approximation for the photon quadrupole  $\Theta_2$  taking into account polarization.

**Solution.** With polarization, the Boltzmann equations for the scalar ( $m = 0$ ) quadrupole ( $\ell = 2$ ) unpolarized and polarized fluctuations  $\Theta_2$  and  $E_2$  are coupled to each other by Thomson-scattering collision terms, equations (34.49c) and (34.49d). The Boltzmann equations are

$$\dot{\Theta}_2 + \frac{k}{5} (2\Theta_1 - 3\Theta_3) = |\dot{\tau}| \left[ -\Theta_2 + \frac{1}{10} (\Theta_2 + \sqrt{6} E_2) \right], \quad (34.50a)$$

$$\dot{E}_2 - \frac{k}{\sqrt{5}} E_3 = |\dot{\tau}| \left[ -E_2 + \frac{\sqrt{6}}{10} (\Theta_2 + \sqrt{6} E_2) \right]. \quad (34.50b)$$

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 (34.50) to

$$\frac{2k}{5}\Theta_1 = |\dot{\tau}| \left[ -\Theta_2 + \frac{1}{10}(\Theta_2 + \sqrt{6}E_2) \right], \quad (34.51a)$$

$$0 = |\dot{\tau}| \left[ -E_2 + \frac{\sqrt{6}}{10}(\Theta_2 + \sqrt{6}E_2) \right]. \quad (34.51b)$$

The equation (34.51b) for  $E_2$  implies that

$$E_2 = \sqrt{\frac{3}{8}}\Theta_2. \quad (34.52)$$

Inserting this into the equation (34.51a) for  $\Theta_2$  implies

$$\Theta_2 = -\frac{8k}{15|\dot{\tau}|}\Theta_1. \quad (34.53)$$

This looks like the earlier unpolarized estimate (31.67), 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 (31.74)–(31.76).

---

### 34.10 Summary of Boltzmann equations for polarized photons

In summary, the Boltzmann equations for polarized photons, generalizing the Boltzmann equations (32.81) for unpolarized photons, are given by equations (34.30), with gravitational redshift source terms  $G_{\ell m}$  given by equations (34.26), and Thomson-scattering collision terms  $C[\Theta_{\ell m}]$ ,  $C[E_{\ell m}]$ , and  $C[B_{\ell m}]$  given by equations (34.49).

---

**Exercise 34.4 Boltzmann code including polarization.** Upgrade the code you wrote in Exercise 32.1 to implement polarization.

---

### 34.11 Radiative transfer of the polarized CMB

The Boltzmann, or radiative transfer, equation for unpolarized photons was given previously by equation (33.1). For the polarized photon distribution, the radiative transfer equations are

$$\left( \frac{\partial}{\partial\eta} - ik\mu - \dot{\tau} \right) (\Theta + \Psi + \hat{\mathbf{p}} \cdot \mathbf{W}) = I - \dot{\tau} S, \quad (34.54a)$$

$$\left( \frac{\partial}{\partial\eta} - ik\mu - \dot{\tau} \right) {}_2\Theta = -\dot{\tau} {}_2S. \quad (34.54b)$$

The  $I$  in the unpolarized radiative transfer equation (34.54a) is the ISW contribution, a sum of harmonics

$$I(\eta, \mathbf{k}, \hat{\mathbf{p}}) \equiv \dot{\Psi} + \dot{\Phi} + \hat{\mathbf{p}} \cdot \dot{\mathbf{W}} + \hat{p}^a \hat{p}^b \dot{h}_{ab} = \sum_{n=0}^2 \sum_{m=-n}^n (-i)^{n-m} I_{nm}(\eta, \mathbf{k}) D_{nm0}(\phi, \theta), \quad (34.55)$$

with

$$I_{00} \equiv \dot{\Psi} + \dot{\Phi}, \quad (34.56a)$$

$$I_{1,\pm 1} \equiv \dot{W}_\pm, \quad (34.56b)$$

$$I_{2,\pm 2} \equiv \sqrt{\frac{2}{3}} \dot{h}_{\pm\pm}. \quad (34.56c)$$

The  $_s S$  in equations (34.54) are Thomson-scattering source terms,

$$S(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \Psi + \hat{\mathbf{p}} \cdot \mathbf{W} + \hat{\mathbf{p}} \cdot \mathbf{v}_b + \Theta_{00} + \frac{1}{2} \sum_{m=-2}^2 (-i)^{2+m} (\Theta_{2m} + \sqrt{6} E_{2m}) D_{2m0}, \quad (34.57a)$$

$${}_2 S(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \sqrt{\frac{3}{2}} \sum_{m=-2}^2 (-i)^{2+m-2} (\Theta_{2m} + \sqrt{6} E_{2m}) D_{2m2}. \quad (34.57b)$$

The harmonic components  ${}_s S_{nm}$  of  $_s S$  defined by

$${}_s S(\eta, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \sum_{n=|s|}^2 \sum_{m=-n}^n (-i)^{n+m-s} {}_s S_{nm}(\eta, \mathbf{k}) D_{nms}(\phi, \theta, \chi) \quad (34.58)$$

are, generalizing equations (33.4),

$$S_{00} \equiv \Theta_{00} + \Psi, \quad (34.59a)$$

$$S_{10} \equiv v_b, \quad (34.59b)$$

$$S_{1,\pm 1} \equiv v_{b,\pm} + W_\pm, \quad (34.59c)$$

$$S_{2m} \equiv \frac{1}{2} (\Theta_{2m} + \sqrt{6} E_{2m}) \quad (-2 \leq m \leq 2), \quad (34.59d)$$

$${}_2 S_{2m} \equiv \sqrt{\frac{3}{2}} (\Theta_{2m} + \sqrt{6} E_{2m}) \quad (-2 \leq m \leq 2). \quad (34.59e)$$

The solution of the radiative transfer equations (34.54) is, generalizing the unpolarized solution (33.6),

$$\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{p}}) + \Psi(\eta_0, \mathbf{k}) + \hat{\mathbf{p}} \cdot \mathbf{W}(\eta_0, \mathbf{k}) = \int_0^{\eta_0} [e^{-\tau} I(\eta, \mathbf{k}) + g(\eta) S(\eta, \mathbf{k})] e^{-ik\mu(\eta-\eta_0)} d\eta, \quad (34.60a)$$

$${}_2 \Theta(\eta_0, \mathbf{k}, \hat{\mathbf{p}}, \chi) = \int_0^{\eta_0} g(\eta) {}_2 S(\eta, \mathbf{k}) e^{-ik\mu(\eta-\eta_0)} d\eta, \quad (34.60b)$$

where  $g(\eta)$  is the visibility function, equation (33.7).

### 34.12 Harmonics of the polarized CMB photon distribution

The spherical harmonics of the solution (34.60) can be found, as previously, by expanding the exponential  $e^{-iy\mu}$  in spherical Bessel functions, equation (33.10). Spin-weighted modified spherical Bessel functions  $j_{\ell nms}(y)$  with  $\ell, n \geq \max(|m|, |s|)$  can be defined by a generalization of equation (33.11),

$$(-i)^{n+m-s} D_{nms}(\phi, \theta, \chi) e^{-iy \cos \theta} = \sum_{\ell=\max(|m|, |s|)}^{\infty} (-i)^{\ell+m-s} (2\ell+1) D_{\ell ms}(\phi, \theta, \chi) j_{\ell nms}(y) . \quad (34.61)$$

The spin index is dropped for brevity from the spin 0 modified Bessel functions,  $j_{\ell nm0}(y) = j_{\ell nm}(y)$ . The spin spherical Bessel functions  $j_{\ell nms}$  are symmetric in  $ms$ ,

$$j_{\ell nms} = j_{\ell nsm} . \quad (34.62)$$

Complex conjugation of  $j_{\ell nms}$  flips the sign of  $m$  or  $s$

$$j_{\ell nms}^* = j_{\ell n, -m, s} = j_{\ell n, m, -s} . \quad (34.63)$$

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  ${}_s\beta_{\ell nm}$ ,

$$j_{\ell nm, \pm s} = {}_s\epsilon_{\ell nm} \pm i {}_s\beta_{\ell nm} . \quad (34.64)$$

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,

$$j_{\ell nm, \pm 2} = \epsilon_{\ell nm} \pm i \beta_{\ell nm} . \quad (34.65)$$

Under  $m \rightarrow -m$ , the electric components are unchanged, while the magnetic components change sign,

$$j_{\ell n, -m} = j_{\ell nm} , \quad \epsilon_{\ell n, -m} = \epsilon_{\ell nm} , \quad \beta_{\ell n, -m} = -\beta_{\ell nm} . \quad (34.66)$$

In all, the spin spherical Bessel functions of relevance are, from equations (34.117) for  $j_{\ell nn s}$  with  $n = m$ ,

and from the recurrence (34.110) for  $j_{\ell nm s}$  for  $n > m$ ,

$$\begin{aligned}
 j_{\ell 00} &= j_\ell , & j_{\ell 10} &= \frac{d j_\ell}{dy} , & j_{\ell 20} &= \frac{1}{2} \left( 1 + 3 \frac{d^2}{dy^2} \right) j_\ell , \\
 j_{\ell 11} &= \sqrt{\frac{\ell(\ell+1)}{2}} \frac{j_\ell}{y} , & j_{\ell 21} &= \sqrt{\frac{3\ell(\ell+1)}{2}} \frac{d(j_\ell/y)}{dy} , \\
 \epsilon_{\ell 20} &= \sqrt{\frac{3(\ell+2)!}{8(\ell-2)!}} \frac{j_\ell}{y^2} , & \epsilon_{\ell 21} &= \frac{\sqrt{(\ell-1)(\ell+2)}}{2} \frac{1}{y^2} \frac{d(y j_\ell)}{dy} , & \epsilon_{\ell 22} &= \frac{1}{4y^2} \left( \frac{d^2}{dy^2} - 1 \right) (y^2 j_\ell) , \\
 \beta_{\ell 20} &= 0 , & \beta_{\ell 21} &= -\frac{\sqrt{(\ell-1)(\ell+2)}}{2} \frac{j_\ell}{y} , & \beta_{\ell 22} &= -\frac{1}{2y^2} \frac{d(y^2 j_\ell)}{dy} .
 \end{aligned} \tag{34.67}$$

Expanding the solution (34.60) in spherical harmonics using equation (34.61) yields the harmonics of the CMB photon distribution today including polarization, generalizing equation (33.15),

$$\begin{aligned}
 \Theta_{\ell 0}(\eta_0, \mathbf{k}) + \delta_{\ell 0} \Psi(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \left[ \dot{\Psi}(\eta_0, \mathbf{k}) + \dot{\Phi}(\eta_0, \mathbf{k}) \right] j_{\ell 00} [k(\eta - \eta_0)] d\eta \\
 &\quad + |g(k)| \sum_{n=0}^2 S_{n 0}(\eta_{\text{rec}}, \mathbf{k}) j_{\ell n 0} [k(\eta_{\text{rec}} - \eta_0)] ,
 \end{aligned} \tag{34.68a}$$

$$\begin{aligned}
 \Theta_{\ell, \pm 1}(\eta_0, \mathbf{k}) + \frac{1}{3} \delta_{\ell 1} W_{\pm}(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \dot{W}_{\pm}(\eta_0, \mathbf{k}) j_{\ell 11} [k(\eta - \eta_0)] d\eta \\
 &\quad + |g(k)| \sum_{n=1}^2 S_{n, \pm 1}(\eta_{\text{rec}}, \mathbf{k}) j_{\ell n 1} [k(\eta_{\text{rec}} - \eta_0)] ,
 \end{aligned} \tag{34.68b}$$

$$\begin{aligned}
 \Theta_{\ell, \pm 2}(\eta_0, \mathbf{k}) &= \int_0^{\eta_0} e^{-\tau} \sqrt{\frac{2}{3}} \dot{h}_{\pm \pm}(\eta_0, \mathbf{k}) j_{\ell 22} [k(\eta - \eta_0)] d\eta \\
 &\quad + |g(k)| S_{2, \pm 2}(\eta_{\text{rec}}, \mathbf{k}) j_{\ell 22} [k(\eta_{\text{rec}} - \eta_0)] ,
 \end{aligned} \tag{34.68c}$$

$$E_{\ell m}(\eta_0, \mathbf{k}) = |g(k)| {}_2 S_{2m}(\eta_{\text{rec}}, \mathbf{k}) \epsilon_{\ell 2m} [k(\eta_{\text{rec}} - \eta_0)] \quad (-2 \leq m \leq 2) , \tag{34.68d}$$

$$B_{\ell m}(\eta_0, \mathbf{k}) = |g(k)| {}_2 S_{2m}(\eta_{\text{rec}}, \mathbf{k}) \beta_{\ell 2m} [k(\eta_{\text{rec}} - \eta_0)] \quad (-2 \leq m \leq 2) . \tag{34.68e}$$

The Thomson-scattering source terms  ${}_s S_{nm}$  are given by equations (34.59), and the spin spherical Bessel functions are given by equations (34.67).

### 34.12.1 Harmonics of the polarized CMB with respect to observed photon directions

As with the unpolarized power spectrum, §33.2.1, 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  $\chi$  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  $\chi$  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 (33.16),

$$\Theta_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^\ell \Theta_{\ell m}(\eta_0, \mathbf{k}), \quad E_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^\ell E_{\ell m}(\eta_0, \mathbf{k}), \quad B_{\ell m}^{\text{obs}}(\eta_0, \mathbf{k}) \equiv (-)^{\ell+1} B_{\ell m}(\eta_0, \mathbf{k}). \quad (34.69)$$

Equivalently, as in the unpolarized equation (33.15), multipoles with respect to the observed direction  $\hat{\mathbf{n}}$  to the CMB are obtained from equations (34.68) 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  ${}_s S_{n m}$  with odd  $n$ , namely  $S_{1 m}$ ,

$$k(\eta - \eta_0) \rightarrow k(\eta_0 - \eta), \quad S_{1 m} \rightarrow -S_{1 m}. \quad (34.70)$$

The sign flips do not affect power spectra, which involve products of fluctuations with the same  $\ell$  and parity.

## 34.13 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  $\eta$  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 (33.26),

$${}_s\Theta(\eta, \mathbf{x}, \hat{\mathbf{n}}, \chi) = \int e^{-i\mathbf{k}\cdot\mathbf{x}} {}_s\Theta(\eta, \mathbf{k}, \hat{\mathbf{n}}, \chi) \frac{d^3 k}{(2\pi)^3}. \quad (34.71)$$

Astronomers observe the temperature fluctuation  ${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi)$  now, at time  $\eta_0$ , 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,

$${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi) = \int \Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}, \chi) \frac{d^3 k}{(2\pi)^3}, \quad (34.72)$$

which generalizes the unpolarized expression (33.27).

The spherical harmonic expansion of the observed real-space temperature fluctuation today is, with conventional normalization of harmonics  ${}_s\Theta_{\ell m}$ ,

$${}_s\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}, \chi) = \sum_{\ell=|s|}^{\infty} \sum_{m=-\ell}^{\ell} {}_s\Theta_{\ell m}(\eta_0, \mathbf{x}_0) {}_{-s}Y_{\ell m}^*(\hat{\mathbf{n}}, \chi), \quad (34.73)$$

which generalizes equation (33.28). 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 (34.23), the coefficient  ${}_s\Theta_{\ell m}$  then has spin weight  $s$  and  $m$  as

opposed to  $s$  and  $-m$ . The spherical harmonic expansion (34.23) of the Fourier-space temperature fluctuation may be written

$${}_s\Theta(\eta_0, \mathbf{k}, \hat{\mathbf{n}}, \chi) = \sum_{\ell=|s|}^{\infty} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \sum_{n=-\ell}^{\ell} (-i)^{\ell+m-s} \sqrt{4\pi(2\ell+1)} {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) {}_{-s}Y_{\ell n}^*(\hat{\mathbf{n}}, \chi), \quad (34.74)$$

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 (34.88) of Wigner rotation matrices  $D_{\ell m'm}(\hat{\mathbf{n}}', \hat{\mathbf{n}})$ ,

$$\delta_{\ell'\ell} {}_sD_{\ell m'm}(\hat{\mathbf{n}}', \hat{\mathbf{n}}) \equiv \int {}_sY_{\ell' m'}^*(\hat{\mathbf{n}}') {}_sY_{\ell m}(\hat{\mathbf{n}}) do. \quad (34.75)$$

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 (34.74) are with respect to a frame in which the  $z$ -axis is along a direction  $\hat{\mathbf{z}}$  fixed in space.

From equations (34.72)–(34.74) it follows that the real-space harmonics are

$${}_s\Theta_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m-s} \int {}_s\Theta_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}. \quad (34.76)$$

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 (34.23) and (34.73) 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 (34.76) are

$$\Theta_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} (-i)^{\ell+m} \int \Theta_{\ell m}(\eta, \mathbf{k}) D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}, \quad (34.77a)$$

$$E_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-2}^2 (-i)^{\ell+m-2} \int E_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}, \quad (34.77b)$$

$$B_{\ell n}(\eta_0, \mathbf{x}_0) = \sqrt{4\pi(2\ell+1)} \sum_{m=-2}^2 (-i)^{\ell+m-2} \int B_{\ell m}(\eta, \mathbf{k}) {}_{-s}D_{\ell nm}^*(\hat{\mathbf{z}}, \hat{\mathbf{k}}) \frac{d^3k}{(2\pi)^3}. \quad (34.77c)$$

### 34.14 Polarized CMB power spectra

#### 34.14.1 Polarized CMB power spectra in Fourier space

Power spectra  $C_\ell^{X'X}(\eta, k)$  with  $X'$  and  $X$  running over any of  $\Theta$ ,  $E$ , and  $B$  are defined by, analogously to the power spectrum  $C_\ell(\eta, k)$  of unpolarized temperature multipoles, equation (33.24),

$$\frac{\delta_{\ell'\ell}}{4\pi} (2\pi)^3 \delta_D(\mathbf{k}' + \mathbf{k}) C_\ell^{X'X}(\eta, k) \equiv \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \langle X'_{\ell'm}(\eta, \mathbf{k}') X_{\ell m}(\eta, \mathbf{k}) \rangle . \quad (34.78)$$

The reality conditions (34.35) 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 (34.78) the unpolarized monopole  $\Theta_{00}$  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 (34.68), 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 (33.18) (the contributions  $\Psi$  and  $\frac{1}{3}W_\pm$  to the unpolarized monopole and dipole are again omitted for brevity),

$$T_{\ell m}^X(\eta, k) \equiv \frac{X_{\ell m}(\eta, \mathbf{k})}{\zeta(\mathbf{k})} , \quad (34.79)$$

where  $\zeta(\mathbf{k})$  is the primordial curvature fluctuation. In terms of the transfer functions (34.79) and the primordial curvature power spectrum  $P_\zeta$ , equation (29.129), the power spectrum  $C_\ell^{X'X}(\eta, \mathbf{k})$  is

$$C_\ell^{X'X}(\eta, \mathbf{k}) = 4\pi \sum_{m=-2}^2 T_{\ell m}^{X'*}(\eta, k) T_{\ell m}^X(\eta, k) P_\zeta(k) . \quad (34.80)$$

#### 34.14.2 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

$$C_\ell^{\Theta\Theta} , \quad C_\ell^{\Theta E} , \quad C_\ell^{EE} , \quad C_\ell^{BB} . \quad (34.81)$$

### 34.14.3 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  $\Theta$ ,  $E$ , and  $B$  are defined such that, generalizing equation (33.31),

$$\delta_{\ell'\ell}\delta_{m'm}C_\ell^{X'X}(\eta_0) \equiv \langle X'_{\ell'm'}^*(\eta_0, \mathbf{x}_0) X_{\ell m}(\eta_0, \mathbf{x}_0) \rangle . \quad (34.82)$$

Once again, the redshift contribution  $\Psi$  to the unpolarized monopole  $\Theta_{00}$ , and the Doppler-shift contribution  $\frac{1}{3}W_m$  to the dipole  $\Theta_{1m}$  on the right hand side of equation (34.82) 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 (34.77) for the real-space harmonics in terms of Fourier-space harmonics, together with the power spectra (34.78) 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 (33.32),

$$C_\ell^{X'X}(\eta_0) = \int C_\ell^{X'X}(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (34.83)$$

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 (34.79) and the primordial curvature power spectrum  $P_\zeta$ , the power spectra  $C_\ell^{X'X}(\eta_0)$  are, from equation (34.80),

$$C_\ell^{X'X}(\eta_0) = 4\pi \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \int T_{\ell m}^{X'*}(\eta_0, k) T_{\ell m}^X(\eta_0, k) P_\zeta(k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (34.84)$$

**Concept question 34.5 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  $C_\ell^{BB}$  measures only vector and tensor modes. Mathematically, the real-space harmonics (34.77) 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 (34.78) without summing over  $m$ . But in real space, the CMB power spectrum  $C_\ell(\eta_0)$ , equation (34.84), is a sum over the scalar, vector, and tensor Fourier-space power spectra,

$$C_\ell^{X'X}(\eta_0) = \sum_{m=-\min(\ell,2)}^{\min(\ell,2)} \int C_{\ell m}^{X'X}(\eta_0, k) \frac{4\pi k^2 dk}{(2\pi)^3} . \quad (34.85)$$

**Exercise 34.6 CMB polarized power spectrum.** Generalize the CMB code you wrote in Exercise 33.1 to include polarization.

---

### 34.15 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  $\hat{z}$ . 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  ${}_s Y_{\ell m}(\theta, \phi, \chi)$ . The spin-weighted spherical harmonics  ${}_s Y_{\ell m}(\theta, \phi, \chi)$  are defined in terms of the Wigner rotation matrix  $D_{\ell mn}(\phi, \theta, \chi)$  discussed in §34.15.1 by

$${}_s Y_{\ell m}(\theta, \phi, \chi) \equiv \sqrt{\frac{2\ell+1}{4\pi}} D_{\ell, m, -s}^*(\phi, \theta, \chi). \quad (34.86)$$

The usual spherical harmonics equal the spin-weighted harmonics with zero spin,  $Y_{\ell m} = {}_0 Y_{\ell m}$ . The reason for complex conjugation and the sign flip of the spin index  $s$  on the right hand side of equation (34.86) is that conventionally  ${}_s Y_{\ell m} \propto e^{im\phi-is\chi}$  whereas  $D_{\ell ms} \propto e^{-im\phi-is\chi}$ . The convention for the Wigner matrix, which treats the angles  $\phi$  and  $\chi$  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 (34.23), 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

$${}s Y_{\ell m}(\theta, \phi) \equiv {}s Y_{\ell m}(\theta, \phi, 0). \quad (34.87)$$

#### 34.15.1 Wigner rotation matrix

The full 3-dimensional rotation group is the orthogonal group  $O(3)$ , or, when extended to objects of half-integral spin, its covering group  $SU(2)$ . The eigenfunctions of  $O(3)$  or  $SU(2)$  are the elements  $D_{\ell m' m}$  of the Wigner rotation matrix.

The **Wigner rotation matrix**  $D_{\ell m' m}(\chi', \alpha, \chi)$  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  $\chi', \alpha, \chi$ ,

$$\delta_{\ell' \ell} D_{\ell m' m}(\chi', \alpha, \chi) \equiv \int Y_{\ell' m'}^*(\mathbf{n}') Y_{\ell m}(\mathbf{n}) d\Omega. \quad (34.88)$$

Equivalently, the spherical harmonics in the unrotated and rotated frames are related by

$$Y_{\ell m}(\mathbf{n}) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\chi', \alpha, \chi) Y_{\ell m'}(\mathbf{n}'). \quad (34.89)$$

Notice that spherical harmonics rotate into linear combinations of harmonics of the same harmonic number  $\ell$ , which is true because rotation leaves the total angular momentum  $L^2$  unchanged. The Euler angles  $\chi'$ ,  $\alpha$ ,  $\chi$  in equation (34.89) correspond to a right-handed rotation of the unit vector  $\mathbf{n}'$  by angle  $\chi'$  about the  $z$ -axis, followed by a right-handed rotation by angle  $\alpha$  about the  $y$ -axis, followed by a right-handed rotation by angle  $\chi$  about the  $z$ -axis,

$$\begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \begin{pmatrix} \cos \chi & \sin \chi & 0 \\ -\sin \chi & \cos \chi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \alpha & 0 & -\sin \alpha \\ 0 & 1 & 0 \\ \sin \alpha & 0 & \cos \alpha \end{pmatrix} \begin{pmatrix} \cos \chi' & \sin \chi' & 0 \\ -\sin \chi' & \cos \chi' & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} n'_x \\ n'_y \\ n'_z \end{pmatrix}. \quad (34.90)$$

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  $\chi$  about direction  $\mathbf{n}$  is  $\exp(-i\chi \mathbf{n} \cdot \mathbf{L})$ . Thus the operator  $D(\chi', \alpha, \chi)$  that generates a rotation by the 3 Euler angles is

$$D(\chi', \alpha, \chi) = e^{-i\chi L_z} e^{-i\alpha L_y} e^{-i\chi' L_z}. \quad (34.91)$$

The spherical harmonic components of the rotation operator are correspondingly (no sum over  $m'$ ,  $m$ )

$$D_{\ell m' m}(\chi', \alpha, \chi) = e^{-im\chi} d_{\ell m' m}(\alpha) e^{-im'\chi'}. \quad (34.92)$$

The matrix  $d_{\ell m' m}(\alpha)$  is the polar part of the full rotation matrix  $D_{\ell m' m}(\chi', \alpha, \chi)$ . The polar rotation matrix  $d_{\ell m' m}(\alpha)$  is a real matrix, orthogonal with respect to  $m'm$ , with matrix inverse

$$d_{\ell m' m}(\alpha)^{-1} = d_{\ell m' m}(-\alpha) = d_{\ell m m'}(\alpha) = d_{\ell, -m', -m}(\alpha) = (-)^{m' - m} d_{\ell m' m}(\alpha). \quad (34.93)$$

The matrix inverse of the Wigner rotation matrix is its Hermitian conjugate, its complex conjugate transpose,

$$D_{\ell m' m}(\chi', \alpha, \chi)^{-1} = D_{\ell m' m}(-\chi, -\alpha, -\chi') = D_{\ell m m'}^*(\chi', \alpha, \chi). \quad (34.94)$$

Complex conjugation flips the signs of  $m'$  and  $m$ , and multiplies by  $(-)^{m' - m}$ ,

$$D_{\ell m' m}^*(\chi', \alpha, \chi) = (-)^{m' - m} D_{\ell, -m', -m}(\chi', \alpha, \chi). \quad (34.95)$$

A parity transformation  $\alpha \rightarrow \pi - \alpha$ ,  $\chi' \rightarrow \chi' + \pi$ ,  $\chi \rightarrow -\chi$  flips the sign of the spin  $s$  and multiplies by  $(-)^{\ell}$ ,

$$D_{\ell m s}(\chi' + \pi, \pi - \alpha, -\chi) = (-)^{\ell} D_{\ell, m, -s}(\chi', \alpha, \chi). \quad (34.96)$$

Particular examples of equation (34.89), illustrating how the signs work out, are

$$Y_{\ell m}(\theta, \phi) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\chi', 0, \chi) Y_{\ell m'}(\theta, \phi + \chi' + \chi), \quad (34.97a)$$

$$Y_{\ell m}(\theta, \phi) = \sum_{m'=-\ell}^{\ell} D_{\ell m' m}(\phi', \alpha, -\phi) Y_{\ell m'}(\theta + \alpha, \phi'). \quad (34.97b)$$

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,

$$Y_{\ell m}(\theta, \phi) = \sqrt{\frac{2\ell + 1}{4\pi}} D_{\ell 0m}(0, -\theta, -\phi) = \sqrt{\frac{2\ell + 1}{4\pi}} D_{\ell m 0}^*(\phi, \theta, 0) , \quad (34.98)$$

consistent with equation (34.86).

The explicit form of the Wigner rotation matrix elements  $D_{\ell m' m}(\chi', \alpha, \chi)$  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)

$$L_z = -i \frac{\partial}{\partial \chi} , \quad L_{\pm} \equiv \frac{e^{\pm i \chi}}{\sqrt{2}} \left( \pm \frac{\partial}{\partial \alpha} + i \frac{1}{\sin \alpha} \frac{\partial}{\partial \chi'} + i \frac{\cos \alpha}{\sin \alpha} \frac{\partial}{\partial \chi} \right) . \quad (34.99)$$

A similar set of equations holds for the total angular momentum operator  $\mathbf{L}'$  in the rotated (primed) frame, with  $\chi' \leftrightarrow \chi$ . The Newman-Penrose components  $L_{\pm}$  are Hermitian conjugates with respect to integration over Euler angles,

$$L_+^\dagger = L_- , \quad (34.100)$$

meaning that for any differentiable functions  $f(\chi', \alpha, \chi)$  and  $g(\chi', \alpha, \chi)$ ,

$$\int_0^{2\pi} \int_0^{2\pi} \int_0^\pi f(L_+ g) \sin \alpha d\alpha d\chi' d\chi = \int_0^{2\pi} \int_0^{2\pi} \int_0^\pi (L_- f) g \sin \alpha d\alpha d\chi' d\chi , \quad (34.101)$$

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

$$[L_+, L_-] = L_z , \quad [L_z, L_{\pm}] = \pm L_{\pm} . \quad (34.102)$$

The squared total angular momentum operator is

$$L^2 = L_+ L_- + L_- L_+ + L_z^2 . \quad (34.103)$$

The explicit form of the squared total angular momentum operator is

$$L^2 = - \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial}{\partial \theta} + \frac{1}{\sin^2 \theta} (L_{z'}^2 - 2 \cos \theta L_{z'} L_z + L_z^2) . \quad (34.104)$$

The Wigner rotation matrix elements  $D_{\ell m' m}(\chi', \alpha, \chi)$  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  $-m$ ,

$$L^2 D_{\ell m' m}(\chi', \alpha, \chi) = \ell(\ell + 1) D_{\ell m' m}(\chi', \alpha, \chi) , \quad (34.105a)$$

$$L_{z'} D_{\ell m' m}(\chi', \alpha, \chi) = -m' D_{\ell m' m}(\chi', \alpha, \chi) , \quad (34.105b)$$

$$L_z D_{\ell m' m}(\chi', \alpha, \chi) = -m D_{\ell m' m}(\chi', \alpha, \chi) . \quad (34.105c)$$

The quantum numbers  $\ell$ ,  $m'$ , and  $m$  must be either all integral or all half-integral, and  $\ell$  must exceed both  $|m'|$  and  $|m|$ ,

$$\ell \geq |m'|, |m| . \quad (34.106)$$

The Wigner rotation matrices are orthogonal with respect to integration over Euler angles,

$$\int_0^{2\pi} \int_0^{2\pi} \int_0^\pi D_{\ell'm'n'}^*(\chi', \alpha, \chi) D_{\ell'mn}(\chi', \alpha, \chi) \sin \alpha d\alpha d\chi' d\chi = \frac{8\pi^2}{2\ell+1} \delta_{\ell'\ell} \delta_{m'm} \delta_{n'n} . \quad (34.107)$$

It follows from the commutation rules (34.102) that the angular momentum operators  $L_\pm$  raise and lower by one unit the  $z$ -component  $L_z$  of the angular momentum,

$$L_\pm D_{\ell,m',-m}(\chi', \alpha, \chi) = \sqrt{\frac{(\ell \pm m)(\ell \mp m + 1)}{2}} D_{\ell,m',-(m \pm 1)}(\chi', \alpha, \chi) . \quad (34.108)$$

The functions  $D_{\ell'mn}(\chi', \alpha, \chi)$  and  $d_{\ell'mn}(\alpha)$  satisfy many recurrence relations. A set of 4 building-block recurrences connecting  $D_{\ell'mn}$  to  $D_{\ell \pm \frac{1}{2}, m \pm \frac{1}{2}, n \pm \frac{1}{2}}$  is

$$\begin{aligned} D_{\frac{1}{2}, \frac{p}{2}, \frac{q}{2}} D_{\ell'mn} = \\ \frac{1}{2\ell+1} \left( pq \sqrt{(\ell - pm)(\ell - qn)} D_{\ell - \frac{1}{2}, m + \frac{p}{2}, n + \frac{q}{2}} + \sqrt{(\ell + 1 + pm)(\ell + 1 + qn)} D_{\ell + \frac{1}{2}, m + \frac{p}{2}, n + \frac{q}{2}} \right) , \end{aligned} \quad (34.109)$$

with  $p = \pm 1$  and  $q = \pm 1$ . Equation (34.109) remains true with  $D$  replaced by  $d$  everywhere. Numerically the most useful recurrence relation, stable for increasing  $\ell$ , is a consequence of equation (34.109),

$$\kappa_{\ell+1,mn} D_{\ell+1,mn} = (2\ell+1) \left[ \cos \alpha - \frac{mn}{\ell(\ell+1)} \right] D_{\ell'mn} - \kappa_{\ell'mn} D_{\ell-1,mn} , \quad (34.110)$$

with

$$\kappa_{\ell'mn} \equiv \sqrt{\frac{(\ell^2 - m^2)(\ell^2 - n^2)}{\ell^2}} , \quad (34.111)$$

starting from  $D_{\ell'mn}(\chi', \alpha, \chi) \equiv e^{-im\chi'} e^{-in\chi} d_{\ell'mn}(\alpha)$  with  $m$  or  $n$  equal to  $\ell$ , and

$$(-)^{\ell-m} d_{\ell'mn} = d_{\ell'mn} = \sqrt{\frac{(2\ell)!}{(\ell+m)!(\ell-m)!}} \cos^{\ell+m} \left( \frac{\alpha}{2} \right) \sin^{\ell-m} \left( \frac{\alpha}{2} \right) . \quad (34.112)$$

Again, equation (34.110) remains true with  $D$  replaced by  $d$  everywhere. The rotation matrices  $D_{\ell'mn}$  for  $m = n = 0$  reduce to Legendre polynomials,

$$D_{\ell00}(\chi', \alpha, \chi) = d_{\ell00}(\alpha) = P_\ell(\cos \alpha) . \quad (34.113)$$

The analysis of polarization in §34.9 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  $\alpha$  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  $\phi, \theta, \chi$  from the  $\hat{\mathbf{k}}$  frame into the  $\hat{\mathbf{p}}$  frame. The various angles are illustrated in Figure 34.2. The equivalence implies the addition theorem

$$\begin{aligned} d_{\ell m' m}(\alpha) &= \sum_{n=-\ell}^{\ell} D_{\ell n m}(\phi, \theta, \chi) D_{\ell m' n}(-\chi', -\theta', -\phi') \\ &= \sum_{n=-\ell}^{\ell} D_{\ell n m}(\phi, \theta, \chi) D_{\ell n m'}^*(\phi', \theta', \chi') . \end{aligned} \quad (34.114)$$

### 34.15.2 Spin-weighted spherical Bessel functions

Spin-weighted spherical Bessel functions  $j_{\ell n m s}(y)$  are defined by equation (34.61). 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 (34.61) for  $j_{\ell n m s}$  implies, from the recurrence (34.109),

$$\begin{aligned} &\frac{1}{2n+1} \left[ \sqrt{(n+1+pm)(n+1+qs)} j_{\ell+\frac{1}{2}, n+\frac{1}{2}, m+\frac{p}{2}, s+\frac{q}{2}} - ipq \sqrt{(n-pm)(n-qs)} j_{\ell+\frac{1}{2}, n-\frac{1}{2}, m+\frac{p}{2}, s+\frac{q}{2}} \right] \\ &= \frac{1}{2l+2} \left[ \sqrt{(\ell+1+pm)(\ell+1+qs)} j_{\ell n m s} - ipq \sqrt{(\ell+1-pm)(\ell+1-qs)} j_{\ell+1, n m s} \right] . \end{aligned} \quad (34.115)$$

For  $pm = n$ , the recurrence (34.115) simplifies to

$$\begin{aligned} &\sqrt{\frac{n+1+qs}{2n+1}} j_{\ell+\frac{1}{2}, n+\frac{1}{2}, n+\frac{1}{2}, s+\frac{q}{2}} \\ &= \frac{1}{2\ell+2} \left[ \sqrt{(\ell+1+n)(\ell+1+qs)} j_{\ell n n s} - ipq \sqrt{(\ell+1-n)(\ell+1-qs)} j_{\ell+1, n n s} \right] . \end{aligned} \quad (34.116)$$

From the recurrence (34.116) it can be shown by induction that  $j_{\ell n m, \pm s}(y)$  with  $n = m$  and integral  $\ell \geq n \geq s \geq 0$  is

$$j_{\ell n n, \pm s}(y) = \sqrt{\frac{(2n)!}{(n+s)!(n-s)!}} \sqrt{\frac{(\ell+n)!(\ell-s)!}{(\ell-n)!(\ell+s)!}} \frac{1}{(2y)^n} \left( \frac{\partial}{\partial y} \mp i \right)^s [y^s j_\ell(y)] . \quad (34.117)$$

Finally, the recurrence relation (34.110) with  $\cos \theta$  replaced by  $i \partial/\partial y$  yields  $j_{\ell n m s}(y)$  in general.

## PART THREE

---

### SPINORS



# 35

---

## The super geometric algebra

The **super geometric algebra** generalizes the geometric algebra to include spinors, which are spin- $\frac{1}{2}$  objects.

For simplicity, this chapter is restricted to the super geometric algebra in 3 dimensions. The generalization of the super geometric algebra to 4-dimensional Minkowski space is presented in Chapter 36.

### 35.1 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  $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3\}$  (or  $\{\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y, \boldsymbol{\gamma}_z\}$ , if you prefer). Choose a pair of basis vectors, in three dimensions conventionally taken to be the pair  $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\}$ , and from them form the spin basis vectors  $\{\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$ , the complex combinations

$$\boldsymbol{\gamma}_+ \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_1 + i\boldsymbol{\gamma}_2) , \quad (35.1a)$$

$$\boldsymbol{\gamma}_- \equiv \frac{1}{\sqrt{2}}(\boldsymbol{\gamma}_1 - i\boldsymbol{\gamma}_2) . \quad (35.1b)$$

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  $\{\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-, \boldsymbol{\gamma}_3\}$  is

$$\gamma_{ab} \equiv \boldsymbol{\gamma}_a \cdot \boldsymbol{\gamma}_b = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \quad (35.2)$$

Notice that the spin basis vectors  $\{\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  are themselves null,  $\boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_+ = \boldsymbol{\gamma}_- \cdot \boldsymbol{\gamma}_- = 0$ , whereas their scalar product with each other is non-zero  $\boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_- = 1$ .

## 35.2 Spin weight

An object is defined to have **spin weight**  $s$  if it varies by

$$e^{-is\theta} \quad (35.3)$$

under a right-handed rotation by angle  $\theta$  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  $\theta$  in the  $\gamma_1-\gamma_2$  plane, the basis vectors  $\gamma_a$  transform as (13.49)

$$\begin{aligned} \gamma_1 &\rightarrow \cos \theta \gamma_1 + \sin \theta \gamma_2 , \\ \gamma_2 &\rightarrow \sin \theta \gamma_1 - \cos \theta \gamma_2 , \\ \gamma_3 &\rightarrow \gamma_3 . \end{aligned} \quad (35.4)$$

It follows that the spin basis vectors  $\gamma_+$  and  $\gamma_-$  transform under a right-handed rotation by angle  $\theta$  in the  $\gamma_1-\gamma_2$  plane

$$\gamma_{\pm} \rightarrow e^{\mp i\theta} \gamma_{\pm} . \quad (35.5)$$

The transformation (35.5) identifies the spin vectors  $\gamma_+$  and  $\gamma_-$  as having spin weight +1 and -1 respectively. The  $\gamma_3$  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:

$\boxed{\text{spin weight } s = \text{number of } + \text{ minus } - \text{ covariant indices}} .$

(35.6)

The spin properties of the components of a tensor are thus manifest when expressed in a spin basis.

## 35.3 Pauli representation of spin basis vectors

In the Pauli representation (13.97), the spin basis vectors  $\gamma_{\pm}$  are represented by the  $2 \times 2$  Pauli matrices

$$\gamma_+ = \sigma_+ \equiv \frac{1}{\sqrt{2}} (\sigma_1 + i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} , \quad \gamma_- = \sigma_- \equiv \frac{1}{\sqrt{2}} (\sigma_1 - i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} . \quad (35.7)$$

The basis vector  $\gamma_3$  is represented as usual by the Pauli matrix  $\sigma_3$ ,

$$\gamma_3 = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \quad (35.8)$$

### 35.4 Spinor basis elements

Introduce a dyad of spinor basis elements  $\gamma_a$  with the index  $a$  running over spin up  $\uparrow$  and spin down  $\downarrow$  (the braces in equation (35.9) signify a set of spinors, not anticommutation),

$$\boxed{\gamma_a \equiv \{\gamma_{\uparrow}, \gamma_{\downarrow}\}} . \quad (35.9)$$

The spinor basis elements  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  physically signify spin up and spin down eigenstates. A more conventional (Dirac) notation is

$$\gamma_{\uparrow} = |\uparrow\rangle , \quad \gamma_{\downarrow} = |\downarrow\rangle . \quad (35.10)$$

It will be seen in §35.12 that the basis spinors  $\gamma_a$  are related to the 3D basis vectors  $\boldsymbol{\gamma}_a$  through a super geometric algebra that is essentially the square root of the geometric algebra. Elements of the geometric algebra act by pre-multiplication on the basis spinors  $\gamma_a$ . Under a rotation by rotor  $R$ , the basis spinors  $\gamma_a$  are defined to transform in the same way as rotors,

$$R : \gamma_a \rightarrow R\gamma_a . \quad (35.11)$$

In the Pauli representation (13.97) the basis spinors  $\gamma_a$  are the column vectors

$$\gamma_{\uparrow} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \quad \gamma_{\downarrow} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \quad (35.12)$$

that are rotated by pre-multiplying by elements of the special unitary group  $SU(2)$ . Rotations transform the basis elements  $\gamma_a$  into linear combinations of each other.

The rotor  $R$  corresponding to a right-handed rotation by angle  $\theta$  in the  $\boldsymbol{\gamma}_1$ - $\boldsymbol{\gamma}_2$  plane is  $e^{-i_3\theta/2}$ , equation (13.91). In the Pauli representation (35.9), the action of  $i_3 = I_3\sigma_3$  on the basis spinors is  $i_3\gamma_{\uparrow} = i\gamma_{\uparrow}$  and  $i_3\gamma_{\downarrow} = -i\gamma_{\downarrow}$ . Under a right-handed rotation by angle  $\theta$  in the  $\boldsymbol{\gamma}_1$ - $\boldsymbol{\gamma}_2$  plane, the basis spinors  $\gamma_a$  therefore transform as

$$\gamma_{\uparrow} \rightarrow e^{-i\theta/2}\gamma_{\uparrow} , \quad \gamma_{\downarrow} \rightarrow e^{i\theta/2}\gamma_{\downarrow} . \quad (35.13)$$

The behaviour (35.13), along with the definition (35.3) of spin, shows that the basis spinors  $\gamma_{\uparrow}$  and  $\gamma_{\downarrow}$  have respective spin weights  $+\frac{1}{2}$  and  $-\frac{1}{2}$ . A rotation by  $\theta = 2\pi$  changes the sign of the basis spinors  $\gamma_a$ . A rotation by  $4\pi$  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 (35.6) generalizes to the statement that the spin weight of a spinor tensor is

$$\boxed{\text{spin weight } s = \frac{1}{2} (\text{number of } \uparrow \text{ minus } \downarrow \text{ covariant indices})} . \quad (35.14)$$

In any equality between vector and spinor tensors, the spin weights of the left and right hand sides must be equal. The rule (35.14) hold not only for column spinors  $\gamma_a$ , but also for row spinors  $\gamma_a \cdot$ , §35.7, and for inner and outer products of spinors, §§35.8 and 35.11.

### 35.5 Pauli spinor

A **Pauli spinor**  $\varphi$  is a complex (with respect to  $i$ ) linear combination of the basis spinors elements  $\gamma_a$ ,

$$\boxed{\varphi = \varphi^a \gamma_a} . \quad (35.15)$$

Just as a multivector  $a^a \boldsymbol{\gamma}_a$  is a vector in the geometric algebra, so also  $\varphi^a \gamma_a$  is a spinor in the super geometric algebra. Consistent with the convention for the geometric algebra, §13.8, a rotation of a spinor  $\varphi$  (35.15) is an active rotation, which rotates the basis spinors  $\gamma_a$  while keeping the coefficients  $\varphi^a$  fixed. Conceptually, the spinor is attached to the frame, and a rotation bodily rotates the frame and everything attached to it.

By construction, a Pauli spinor transforms under a spatial rotation by rotor  $R$  like the basis spinors, equation (35.11),

$$R : \varphi \rightarrow R\varphi . \quad (35.16)$$

A Pauli spinor  $\varphi$  is a spin- $\frac{1}{2}$  object, in the sense that a rotation by  $2\pi$  changes the sign of the spinor, and a rotation by  $4\pi$  is required to return the spinor to its original value.

### 35.6 Spinor metric

In a matrix representation, the tensor product of spinor basis elements  $\gamma_a$  and  $\gamma_b$  can be represented as the  $2 \times 2$  matrix  $\gamma_a \gamma_b^\top$ , a matrix product of the column spinor  $\gamma_a$  with the row spinor  $\gamma_b^\top$ . In accordance with the transformation rule (35.11), the tensor product of basis spinors rotates as

$$R : \gamma_a \gamma_b^\top \rightarrow R \gamma_a \gamma_b^\top R^\top . \quad (35.17)$$

Consider the spinor tensor  $\varepsilon$  with the defining property that for any rotor  $R$

$$\varepsilon R^\top = \bar{R} \varepsilon . \quad (35.18)$$

A necessary and sufficient condition for (35.18) to hold is that, up to a choice of  $\pm$  sign on the right hand side, the spatial tetrad basis vectors  $\boldsymbol{\gamma}_a$  transform as

$$\varepsilon \boldsymbol{\gamma}_a^\top = -\boldsymbol{\gamma}_a \varepsilon \quad \text{for } a = 1, 2, 3 . \quad (35.19)$$

The  $-$  sign on the right hand side of (35.19) is chosen because in the Pauli representation the equation with a  $+$  sign has no solution. In the Pauli representation (13.97), where  $\boldsymbol{\gamma}_a = \sigma_a$ , the condition (35.19) requires that  $\varepsilon$  commutes with  $\boldsymbol{\gamma}_2$ , and anticommutes with  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_3$ . The only basis element of the spacetime algebra with the required (anti)commutation properties is  $\boldsymbol{\gamma}_2$ , so the spinor tensor  $\varepsilon$  must equal  $\boldsymbol{\gamma}_2$  up to a possible scalar normalization,

$$\varepsilon \equiv i \boldsymbol{\gamma}_2 . \quad (35.20)$$

In the Pauli representation (13.97), the spinor tensor (35.20) is the antisymmetric matrix

$$\varepsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \quad (35.21)$$

The chosen normalization is such that  $\varepsilon$  is real and orthogonal, and its square is minus the unit matrix,

$$\varepsilon^{-1} = \varepsilon^\top, \quad \varepsilon^2 = -1. \quad (35.22)$$

Despite the equality of  $\varepsilon$  and  $i\gamma_2$  in the Pauli representation,  $\varepsilon$  is defined to transform as a spinor tensor under spatial rotations, not as an element of the geometric algebra.

The condition (35.18) implies that the spinor tensor  $\varepsilon$  is invariant under rotations,

$$R : \varepsilon \rightarrow R\varepsilon R^\top = R\bar{R}\varepsilon = \varepsilon. \quad (35.23)$$

The components of the spinor tensor  $\varepsilon$  define the antisymmetric spinor metric  $\varepsilon_{ab}$ ,

$$\gamma_a^\top \varepsilon \gamma_b = \varepsilon_{ab}. \quad (35.24)$$

### 35.7 Row basis spinors

It is convenient to use the symbol  $\gamma_a \cdot$  with a trailing dot, symbolic of the trailing  $\varepsilon$ , to denote the row spinor  $\gamma_a^\top \varepsilon$ ,

$$\boxed{\gamma_a \cdot \equiv \gamma_a^\top \varepsilon}. \quad (35.25)$$

The motivation for the trailing dot notation is equation (35.29) below. The two row spinors (the braces in equation (35.26) signify a set of spinors, not anticommutation)

$$\gamma_a \cdot = \{\gamma_{\uparrow} \cdot, \gamma_{\downarrow} \cdot\} \quad (35.26)$$

provide a basis for row spinors. The spin weights of the row basis spinors are in accord with their covariant indices:  $\gamma_{\uparrow} \cdot$  has spin weight  $+\frac{1}{2}$ , while  $\gamma_{\downarrow} \cdot$  has spin weight  $-\frac{1}{2}$ . The row spinors  $\gamma_a \cdot$  rotate as

$$R : \gamma_a \cdot \equiv \gamma_a^\top \varepsilon \rightarrow \gamma_a^\top R^\top \varepsilon = \gamma_a^\top \varepsilon \bar{R} = \gamma_a \cdot \bar{R}. \quad (35.27)$$

Thus row spinors  $\gamma_a \cdot$  transform like reverse rotors, just as column spinors  $\gamma_a$  transform like rotors. In the Pauli representation (13.97) the row basis spinors  $\gamma_a \cdot$  are the row spinors

$$\gamma_{\uparrow} \cdot = \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad \gamma_{\downarrow} \cdot = \begin{pmatrix} -1 & 0 \end{pmatrix}. \quad (35.28)$$

### 35.8 Inner products of spinor basis elements

The product of the row spinor  $\gamma_a \cdot$  with the column spinor  $\gamma_b$  defines their inner product, or scalar product, which equals the **spinor metric**  $\varepsilon_{ab}$  in accordance with equation (35.24),

$$\boxed{\gamma_a \cdot \gamma_b = \varepsilon_{ab}}. \quad (35.29)$$

Equation (35.29) motivates the trailing dot notation for the row spinor. The scalar product is antisymmetric,

$$\boxed{\gamma_a \cdot \gamma_b = -\gamma_b \cdot \gamma_a}. \quad (35.30)$$

The antisymmetry of the spinor scalar product contrasts with the symmetry of the usual vector scalar product. The scalar product (35.29) is a scalar,

$$R : \gamma_a \cdot \gamma_b \rightarrow \gamma_a \cdot \bar{R} R \gamma_b = \gamma_a \cdot \gamma_b . \quad (35.31)$$

Thus the spinor metric  $\varepsilon_{ab}$  is invariant under rotations, just like the Euclidean metric  $\delta_{ab}$ .

### 35.9 Lowering and raising spinor indices

The antisymmetric spinor metric  $\varepsilon_{ab}$  is given in the Pauli representation by equation (35.24). The inverse metric  $\varepsilon^{ab}$  is defined by  $\varepsilon^{ab}\varepsilon_{bc} = \delta_c^a$ . The spinor metric and its inverse satisfy

$$\varepsilon_{ab} = -\varepsilon_{ba} = -\varepsilon^{ab} = \varepsilon^{ba} . \quad (35.32)$$

Indices on a spinor tensor are lowered and raised by pre-multiplying by the metric  $\varepsilon_{ab}$  and its inverse  $\varepsilon^{ab}$ . The contravariant components  $\gamma^a$  of the column spinor basis elements, satisfying  $\gamma^a = \varepsilon^{ab}\gamma_b$ , are

$$\gamma^\uparrow = -\gamma_\downarrow , \quad \gamma^\downarrow = \gamma_\uparrow . \quad (35.33)$$

For example,  $\gamma^\uparrow = \varepsilon^{\uparrow\downarrow}\gamma_\downarrow = -\gamma_\downarrow$ . 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,  $\gamma^a = \varepsilon^{ab}\gamma_b = -\gamma_b\varepsilon^{ba}$ . The contravariant components  $\gamma^a \cdot$  of the row spinor basis elements satisfy the same relations (35.33) 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,

$$\gamma^a \cdot \gamma_b = -\gamma_b \cdot \gamma^a = \delta_b^a . \quad (35.34)$$

### 35.10 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

$$\boxed{\varphi \cdot \equiv \varphi^\top \varepsilon = \varphi^a \gamma_a \cdot} . \quad (35.35)$$

It rotates as

$$R : \varphi \cdot \rightarrow \varphi \cdot \bar{R} . \quad (35.36)$$

A row spinor  $\varphi \cdot$  transforms like a reverse rotor.

The product of a row Pauli spinor  $\varphi \cdot = \varphi^a \gamma_a \cdot$  with a column Pauli spinor  $\chi = \chi^a \gamma_a$  forms a scalar, which may be written variously

$$\varphi \cdot \chi = \varphi^\top \varepsilon \chi = \varphi^a \gamma_a \cdot \chi^b \gamma_b = \varepsilon_{ab} \varphi^a \chi^b = \varphi^a \chi_a = -\varphi_a \chi^a = -\varepsilon^{ab} \varphi_a \chi_b . \quad (35.37)$$

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  $\varphi^a$  of a column spinor  $\varphi$  can be projected out by pre-multiplying by the row basis spinor  $\gamma_a \cdot$ ,

$$\gamma^a \cdot \varphi = \gamma^a \cdot \varphi^b \gamma_b = \delta_b^a \varphi^b = \varphi^a . \quad (35.38)$$

The components  $\varphi^a$  of a row spinor  $\varphi \cdot$  can be projected out by post-multiplying by minus the column basis spinor  $\gamma^a$ ,

$$-\varphi \cdot \gamma^a = -\varphi^b \gamma_b \cdot \gamma^a = \delta_b^a \varphi^b = \varphi^a . \quad (35.39)$$

If the coefficients  $\varphi^a$  and  $\chi^b$  of Pauli spinors  $\varphi = \varphi^a \gamma_a$  and  $\chi = \chi^b \gamma_b$  are taken to be ordinary commuting complex numbers, then the Pauli scalar product is anticommuting

$$\boxed{\varphi \cdot \chi = -\chi \cdot \varphi} . \quad (35.40)$$

In quantum field theory spinor coefficients are sometimes taken to be anticommuting, in which case the scalar product would be commuting. A proof that the Pauli spinors anticommute (so their coefficients must be ordinary commuting complex numbers) is given later, equation (35.68).

### 35.11 Outer products of spinor basis elements

A row spinor  $\gamma_a \cdot$  multiplied by a column spinor  $\gamma_b$  yields their scalar product. In the opposite order, a column spinor  $\gamma_a$  multiplied by a row spinor  $\gamma_b \cdot$  yields their outer product. The outer product  $\gamma_a \gamma_b \cdot$  rotates like a multivector in the geometric algebra,

$$R : \gamma_a \gamma_b \cdot \equiv \gamma_a \gamma_b^\top \varepsilon \rightarrow R \gamma_a \gamma_b^\top R^\top \varepsilon = R \gamma_a \gamma_b^\top \varepsilon \bar{R} = R \gamma_a \gamma_b \cdot \bar{R} . \quad (35.41)$$

The trailing dot on the outer product  $\gamma_a \gamma_b \cdot$  is symbolic of the trailing  $\varepsilon$ , necessary to convert the spinor tensor  $\gamma_a \gamma_b^\top$  into an object that transforms like a multivector.

The products of the 2 column basis spinors  $\gamma_a$  with the 2 row basis spinors  $\gamma_b \cdot$  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 (13.97), 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,

$$[\gamma_\downarrow, \gamma_\uparrow] \cdot = 1 , \quad (35.42)$$

where the 1 on the right hand side denotes the unit element of the 3D geometric algebra, the  $2 \times 2$  identity matrix. The trailing dot on the commutator indicates that the right partner of each product is a row spinor,

$[\gamma_\uparrow, \gamma_\downarrow] \cdot = \gamma_\uparrow \gamma_\downarrow^\top \cdot - \gamma_\downarrow \gamma_\uparrow^\top \cdot$ . The combination (35.42) is familiar from quantum mechanics as the spin-0 singlet formed from a combination of two spin- $\frac{1}{2}$  particles, commonly written in Dirac notation

$$[\gamma_\downarrow, \gamma_\uparrow] = |\downarrow\rangle|\uparrow\rangle - |\uparrow\rangle|\downarrow\rangle. \quad (35.43)$$

The spin weight of the singlet (35.42) is zero according to the rule (35.14), as it should be for a scalar.

The symmetric outer products of spinors form a triplet,

$$\{\gamma_\uparrow, \gamma_\uparrow\} \cdot = \sqrt{2} \boldsymbol{\gamma}_+, \quad \{\gamma_\uparrow, \gamma_\downarrow\} \cdot = -\boldsymbol{\gamma}_3, \quad \{\gamma_\downarrow, \gamma_\downarrow\} \cdot = -\sqrt{2} \boldsymbol{\gamma}_-. \quad (35.44)$$

The combinations (35.44) of basis spinors are, modulo normalization factors, familiar from quantum mechanics as the three components of the spin-1 triplet formed from a combination of two spin- $\frac{1}{2}$  particles,

$$\{\gamma_\uparrow, \gamma_\uparrow\} = 2|\uparrow\rangle|\uparrow\rangle, \quad \{\gamma_\uparrow, \gamma_\downarrow\} = (|\uparrow\rangle|\downarrow\rangle + |\downarrow\rangle|\uparrow\rangle), \quad \{\gamma_\downarrow, \gamma_\downarrow\} = 2|\downarrow\rangle|\downarrow\rangle. \quad (35.45)$$

The spin weights of the triplet (35.44) are respectively +1, 0, -1 according to the rules (35.6) and (35.14). 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 (35.42) is the unit matrix, whose trace is 2),

$$\text{Tr } \gamma_a \gamma_b \cdot = \gamma_b \cdot \gamma_a = \varepsilon_{ba}. \quad (35.46)$$

### 35.12 The 3D super geometric algebra

The 3D super geometric 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  $1 \times 1$ ,  $1 \times 2$ ,  $2 \times 1$ , and  $2 \times 2$ . The true scalars are just complex numbers. A column spinor  $\varphi$  is a complex linear combination of column basis spinors  $\gamma_a$ ,

$$\varphi = \varphi^a \gamma_a, \quad (35.47)$$

while a row spinor  $\varphi \cdot$  is a complex linear combination of row basis spinors  $\gamma_a \cdot$ ,

$$\varphi \cdot = \varphi^a \gamma_a \cdot. \quad (35.48)$$

A multivector  $a$  is a complex linear combination of outer products of the column and row basis spinors,

$$a = a^{ab} \gamma_a \gamma_b \cdot. \quad (35.49)$$

Linearity and the transformation law (35.41) imply that the algebra of sums and products of outer products of spinors is isomorphic to the geometric algebra.

There are two distinct kinds of scalar in the super geometric algebra, true scalars that are just complex numbers, and multivector scalars that are proportional to the unit matrix in a matrix representation. See §36.6.2 for an explanation of this conundrum.

As seen in §35.8 and §35.11, a column spinor  $\varphi$  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.

In applications to quantum field theory, rather than prohibiting certain kinds of multiplication, it is convenient instead to assert that prohibited multiplications simply yield a true scalar value of zero. Thus

$$\varphi\chi = 0, \quad \varphi \cdot \chi \cdot = 0, \quad \varphi\mathbf{a} = 0, \quad \mathbf{a}\varphi \cdot = 0. \quad (35.50)$$

This allows all objects in the super geometric algebra to be added and multiplied, regardless of their species.

In general, a sequence of products of spinors yields a non-zero result provided that they alternate between column spinor and row spinor,

$$\varphi\chi \cdot \psi \quad \text{or} \quad \varphi \cdot \chi \psi \cdot . \quad (35.51)$$

Both product sequences are associative,

$$\varphi\chi \cdot \psi = (\varphi\chi \cdot)\psi = \varphi(\chi \cdot \psi), \quad (35.52)$$

and

$$\varphi \cdot \chi \psi \cdot = (\varphi \cdot \chi)\psi \cdot = \varphi \cdot (\chi \psi \cdot). \quad (35.53)$$

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  $\mathbf{a} = a^{ab}\gamma_a\gamma_b \cdot$  and  $\mathbf{b} = b^{ab}\gamma_a\gamma_b \cdot$  be two multivectors expressed as a sum of outer products of spinors. Their product is the multivector

$$\mathbf{ab} = a^{ab}\gamma_a\gamma_b \cdot b^{cd}\gamma_c\gamma_d \cdot = \gamma_a a^{ab}\varepsilon_{bc}b^{cd}\gamma_d \cdot = \gamma_a a^{ab}b_b^d\gamma_d \cdot . \quad (35.54)$$

A sequence such as  $\varphi \cdot \mathbf{a}\chi$  simplifies as

$$\varphi \cdot \mathbf{a}\chi = \varphi^a\gamma_a \cdot a^{bc}\gamma_b\gamma_c \cdot \chi^d\gamma_d = \varphi^a\varepsilon_{ab}a^{bc}\varepsilon_{cd}\chi^d = \varphi^a a_a^c \chi_c . \quad (35.55)$$

### 35.13 *C*-conjugate Pauli spinor

The 3D super geometric algebra possesses a discrete transformation called *C*-conjugation (or simply conjugation, when there is no ambiguity). The *C*-conjugate Pauli spinor  $\bar{\varphi}$  is defined, equation (35.57), such that (a) its components are complex conjugates (with respect to  $i$ ) of those of the parent spinor  $\varphi$ , and (b) it rotates according to the complex conjugate representation of the Pauli matrices.

The complex conjugate  $\varphi^*$  of a Pauli spinor  $\varphi = \varphi^a\gamma_a$  is defined to be the spinor with complex conjugate (with respect to  $i$ ) coefficients,

$$\varphi^* \equiv \varphi^{a*}\gamma_a . \quad (35.56)$$

The  $C$ -conjugate Pauli spinor  $\bar{\varphi}$  is defined by (despite the similar notation, the conjugate spinor  $\bar{\varphi}$  is not the reverse spinor  $\bar{\varphi}$  defined by equation (13.110); rather, the reverse spinor coincides with the row  $C$ -conjugate spinor  $\bar{\varphi} = \bar{\varphi}^\cdot$  defined by equation (35.63); note that the conjugate overbar  $\bar{\phantom{a}}$  is slightly smaller and thinner than the reverse overbar  $\bar{\phantom{a}}$ ; but in any case, it should be clear from the context whether the conjugate or reverse spinor is intended),

$$\boxed{\bar{\varphi} \equiv \varepsilon \varphi^*} . \quad (35.57)$$

The 3D spinor metric tensor  $\varepsilon$  was constructed earlier to have the property (35.19) that commutation with  $\varepsilon$  converts basis vectors  $\gamma_a$  of the geometric algebra to minus their transposes. The spinor metric tensor  $\varepsilon$  has the additional property that commutation with it converts even (but not odd) basis elements 1 and  $I_3\gamma_a$  of the geometric algebra to their complex conjugates (with respect to  $i$ ) in the Pauli representation (13.97). Consequently commutation with  $\varepsilon$  converts rotors  $R$ , which are real linear combinations of the even basis elements, to their complex conjugates,

$$\varepsilon R = R^* \varepsilon . \quad (35.58)$$

It follows that, by construction, the  $C$ -conjugate Pauli spinor  $\bar{\varphi}$  has the property that it rotates according to the complex conjugate representation of the Pauli matrices,

$$R : \bar{\varphi} \equiv \varepsilon \varphi^* \rightarrow \varepsilon R \varphi^* = R^* \varepsilon \varphi^* = R^* \bar{\varphi} . \quad (35.59)$$

Conjugating a Pauli spinor  $\varphi$  twice changes its sign,

$$\bar{\bar{\varphi}} = -\varphi . \quad (35.60)$$

The components  $\bar{\varphi}^a$  of the conjugate spinor are projected out by pre-multiplying by the row spinor  $\gamma^a \cdot$ , equation (36.59),

$$\bar{\varphi}^a = \gamma^a \cdot \bar{\varphi} . \quad (35.61)$$

In the Pauli representation, the components of the Pauli spinor and its conjugate are related by

$$\varphi^a = \begin{pmatrix} \varphi^\uparrow \\ \varphi^\downarrow \end{pmatrix} , \quad \bar{\varphi}^a = \begin{pmatrix} \varphi^{\downarrow*} \\ -\varphi^{\uparrow*} \end{pmatrix} . \quad (35.62)$$

Equations (35.62) show that  $C$ -conjugation flips spin: the conjugate of a spin-up spinor is a spin-down spinor, and vice versa.

### 35.14 Scalar products of spinors and conjugate spinors

The row  $C$ -conjugate Pauli spinor  $\bar{\varphi}^\cdot$  corresponding to the column  $C$ -conjugate spinor  $\bar{\varphi}$  coincides with the Hermitian conjugate spinor  $\varphi^\dagger$ , which in turn coincides with the reverse spinor  $\bar{\varphi}$ , equation (13.110),

$$\bar{\varphi}^\cdot \equiv \bar{\varphi}^\top \varepsilon = \varphi^\dagger \varepsilon^\top \varepsilon = \varphi^\dagger = \bar{\varphi} . \quad (35.63)$$

Despite the equality of the row conjugate spinor  $\bar{\varphi} \cdot$  and the reverse spinor  $\bar{\varphi}$ , the column conjugate spinor  $\bar{\varphi}$  defined by equation (35.57) is not the same as the reverse spinor, and the two should not be confused.

The scalar product of a row conjugate Pauli spinor  $\bar{\varphi} \cdot$  with a column Pauli spinor  $\chi$  coincides with the product of the Hermitian conjugate spinor  $\varphi^\dagger$  with the spinor  $\chi$ ,

$$\bar{\varphi} \cdot \chi = \varphi^\dagger \chi = (\begin{array}{cc} \varphi^{\uparrow *} & \varphi^{\downarrow *} \end{array}) \begin{pmatrix} \chi^\uparrow \\ \chi^\downarrow \end{pmatrix} = \varphi^{\uparrow *} \chi^\uparrow + \varphi^{\downarrow *} \chi^\downarrow . \quad (35.64)$$

In particular, the scalar product  $\bar{\varphi} \cdot \varphi$  of a spinor with its own conjugate is real and positive,

$$\bar{\varphi} \cdot \varphi = \varphi^\dagger \varphi . \quad (35.65)$$

The complex conjugate of the scalar product satisfies

$$(\bar{\varphi} \cdot \chi)^* \equiv ((\varepsilon \varphi^*)^\top \varepsilon \chi)^* = \varphi^\top \varepsilon^\top \bar{\chi} = -\varphi^\top \varepsilon \bar{\chi} = -\varphi \cdot \bar{\chi} . \quad (35.66)$$

The sign flip in the fourth expression occurs because the spinor metric tensor  $\varepsilon$  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

$$(\bar{\varphi} \cdot \varphi)^* = -\varphi \cdot \bar{\varphi} . \quad (35.67)$$

Equation (35.67), along with the condition that the scalar product is real,  $(\bar{\varphi} \cdot \varphi)^* = \bar{\varphi} \cdot \varphi$ , equation (35.65), requires that the scalar product  $\bar{\varphi} \cdot \varphi$  be anticommuting,

$$\bar{\varphi} \cdot \varphi = -\varphi \cdot \bar{\varphi} . \quad (35.68)$$

Equation (35.68) proves that the scalar product of Pauli spinors must be anticommuting, as asserted earlier, equation (35.40).

In non-relativistic quantum mechanics, the real positive scalar (35.65) is interpreted as the probability of the Pauli spinor  $\varphi$ . Since conjugating a Pauli spinor twice flips its sign, equation (35.60), the scalar product (35.65) is the same regardless of whether the spinor  $\varphi$  or its conjugate  $\bar{\varphi}$  is taken:

$$\bar{\bar{\varphi}} \cdot \bar{\varphi} = -\varphi \cdot \bar{\varphi} = \bar{\varphi} \cdot \varphi . \quad (35.69)$$

### 35.15 *C*-conjugate multivectors

DOUBLE-CHECK.

*C*-conjugate multivectors  $\bar{\mathbf{a}}$  in the super geometric algebra are defined, similarly to *C*-conjugate Pauli spinors, such that their components are complex conjugates of the parent multivector  $\mathbf{a}$ , and they rotate according to the complex conjugate representation of the Pauli matrices (the *C*-conjugate multivector  $\bar{\mathbf{a}}$  is not the same as the reverse multivector  $\bar{\mathbf{a}}$ ; note that the conjugate overbar  $\bar{\phantom{a}}$  is slightly smaller and thinner than the reverse overbar  $\bar{\phantom{a}}$ ).

Consistent with the definition (35.56) of the complex conjugate spinor  $\varphi^*$ , the complex conjugate multivector  $\mathbf{a}^*$  of a multivector  $\mathbf{a} \equiv a^A \gamma_A$  is defined to be

$$\mathbf{a}^* \equiv a^{A*} \gamma_A . \quad (35.70)$$

Complex conjugation commutes with the isomorphism between multivectors and outer products of spinors in the super geometric 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 (35.57) of the  $C$ -conjugate spinor  $\varphi^*$ , the  $C$ -conjugate multivector  $\bar{\mathbf{a}}$  is defined by

$$\bar{\mathbf{a}} \equiv \varepsilon \mathbf{a}^* \varepsilon^\top . \quad (35.71)$$

If the multivector is an outer product of spinors,  $\mathbf{a} = \varphi\chi\cdot$ , then the  $C$ -conjugate multivector is the outer product of the  $C$ -conjugate spinors,  $\bar{\mathbf{a}} = \bar{\varphi}\bar{\chi}\cdot$ . Like the  $C$ -conjugate spinor, equation (35.59), the  $C$ -conjugate multivector  $\bar{\mathbf{a}}$  rotates according to the complex conjugate representation of the Pauli matrices,

$$R : \bar{\mathbf{a}} \equiv \varepsilon \mathbf{a}^* \varepsilon^\top \rightarrow \varepsilon R \mathbf{a}^* \bar{R} \varepsilon^\top = R^* \varepsilon \mathbf{a}^* \varepsilon^\top \bar{R}^* = R^* \bar{\mathbf{a}} \bar{R}^* . \quad (35.72)$$

### 35.16 No self-conjugate Majorana Pauli spinor

A Pauli spinor satisfying the self-conjugate “Majorana” condition

$$\bar{\varphi} = \varphi . \quad (35.73)$$

cannot exist, since any self-conjugate Pauli spinor must vanish, because of the antisymmetry of the scalar product, equation (35.68).

# 36

---

## Super spacetime algebra

This chapter presents the **super spacetime algebra**, the generalization of the 4-dimensional spacetime algebra to include spinors.

### 36.1 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,  $\gamma_v$  (outgoing) and  $\gamma_u$  (ingoing), while the remaining two axes  $\gamma_+$  and  $\gamma_-$  are spin axes.

#### 36.1.1 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

$$\boxed{\gamma_v \equiv \frac{1}{\sqrt{2}}(\gamma_0 + \gamma_3)}, \quad (36.1a)$$

$$\boxed{\gamma_u \equiv \frac{1}{\sqrt{2}}(\gamma_0 - \gamma_3)}, \quad (36.1b)$$

$$\boxed{\gamma_+ \equiv \frac{1}{\sqrt{2}}(\gamma_1 + i\gamma_2)}, \quad (36.1c)$$

$$\boxed{\gamma_- \equiv \frac{1}{\sqrt{2}}(\gamma_1 - i\gamma_2)}, \quad (36.1d)$$

or in matrix form

$$\begin{pmatrix} \gamma_v \\ \gamma_u \\ \gamma_+ \\ \gamma_- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & i & 0 \\ 0 & 1 & -i & 0 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}. \quad (36.2)$$

All four tetrad axes are null

$$\boldsymbol{\gamma}_v \cdot \boldsymbol{\gamma}_v = \boldsymbol{\gamma}_u \cdot \boldsymbol{\gamma}_u = \boldsymbol{\gamma}_+ \cdot \boldsymbol{\gamma}_+ = \boldsymbol{\gamma}_- \cdot \boldsymbol{\gamma}_- = 0 . \quad (36.3)$$

The tetrad metric of the Newman-Penrose tetrad  $\{\boldsymbol{\gamma}_v, \boldsymbol{\gamma}_u, \boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-\}$  is

$$\gamma_{mn} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} . \quad (36.4)$$

### 36.1.2 Boost and spin weight

An object is defined to have **boost weight**  $n$  if it varies by

$$e^{n\theta} \quad (36.5)$$

under a boost by rapidity  $\theta$  along the positive 3-direction.

Under a boost by rapidity  $\theta$  in the 3-direction, the basis vectors  $\boldsymbol{\gamma}_m$  transform as (14.42)

$$\boldsymbol{\gamma}_0 \rightarrow \boldsymbol{\gamma}_0 \cosh \theta + \boldsymbol{\gamma}_3 \sinh \theta , \quad (36.6a)$$

$$\boldsymbol{\gamma}_3 \rightarrow \boldsymbol{\gamma}_3 \cosh \theta + \boldsymbol{\gamma}_0 \sinh \theta , \quad (36.6b)$$

$$\boldsymbol{\gamma}_a \rightarrow \boldsymbol{\gamma}_a \quad (a = 1, 2) . \quad (36.6c)$$

It follows that a boost by rapidity  $\theta$  in the 3-direction multiplies the outgoing and ingoing axes  $\boldsymbol{\gamma}_v$  and  $\boldsymbol{\gamma}_u$  by a blueshift factor  $e^\theta$  and its reciprocal,

$$\boldsymbol{\gamma}_v \rightarrow e^\theta \boldsymbol{\gamma}_v , \quad \boldsymbol{\gamma}_u \rightarrow e^{-\theta} \boldsymbol{\gamma}_u . \quad (36.7)$$

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

$$e^\theta = \left( \frac{1+v}{1-v} \right)^{1/2} . \quad (36.8)$$

Thus  $\boldsymbol{\gamma}_v$  has boost weight +1, and  $\boldsymbol{\gamma}_u$  has boost weight -1. The spin axes  $\boldsymbol{\gamma}_\pm$  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:

boost weight  $n = \text{number of } v \text{ minus } u \text{ covariant indices}$

. (36.9)

The operation of boosting along the 3-axis, which is the same as a rotation in the  $\boldsymbol{\gamma}_0$ - $\boldsymbol{\gamma}_3$  plane, commutes with the operation of rotating in the  $\boldsymbol{\gamma}_1$ - $\boldsymbol{\gamma}_2$  plane. The concept of spin weight presented in §35.2 holds unchanged. The outgoing and ingoing basis vectors  $\boldsymbol{\gamma}_v$  and  $\boldsymbol{\gamma}_u$  have spin weight zero, while  $\boldsymbol{\gamma}_+$  and  $\boldsymbol{\gamma}_-$  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 (35.14)):

$$\boxed{\text{spin weight } s = \text{number of } + \text{ minus } - \text{ covariant indices}} . \quad (36.10)$$

The boost and spin properties of the components of a tensor are thus manifest in a Newman-Penrose tetrad.

**Concept question 36.1 Boost and spin weight.** Consider the abstract vector

$$\mathbf{A} = A^m \boldsymbol{\gamma}_m . \quad (36.11)$$

Since the contravariant components  $A^m$  is the coefficient of  $\boldsymbol{\gamma}_m$ , isn't the boost and spin weight of  $A^m$  that of  $\boldsymbol{\gamma}_m$ ? **Answer.** No. It is the covariant components  $A_m$  that have the boost and spin weight of  $\boldsymbol{\gamma}_m$ , through

$$A_m = \boldsymbol{\gamma}_m \cdot \mathbf{A} . \quad (36.12)$$

The contravariant components  $A^m$  in a Newman-Penrose tetrad have boost and spin weights opposite to the covariant components  $A_m$ .

## 36.2 Chiral representation of $\gamma$ -matrices

The **chiral representation** of the Dirac  $\gamma$ -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 (14.91) by the transformation

$$X : \boldsymbol{\gamma}_m \rightarrow X \boldsymbol{\gamma}_m X^{-1} , \quad (36.13)$$

where  $X$  is the real unitary ( $X^{-1} = X^\top$ ) matrix

$$X \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} . \quad (36.14)$$

The  $\gamma$ -matrices in the chiral representation are

$$\boldsymbol{\gamma}_0 = i \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} , \quad \boldsymbol{\gamma}_a = i \begin{pmatrix} 0 & \sigma_a \\ -\sigma_a & 0 \end{pmatrix} . \quad (36.15)$$

The bivectors  $\boldsymbol{\sigma}_a$  and  $I\boldsymbol{\sigma}_a$  and the pseudoscalar  $I$  are

$$\boldsymbol{\gamma}_0 \boldsymbol{\gamma}_a = \boldsymbol{\sigma}_a = \begin{pmatrix} \sigma_a & 0 \\ 0 & -\sigma_a \end{pmatrix} , \quad \frac{1}{2} \varepsilon_{abc} \boldsymbol{\gamma}_b \boldsymbol{\gamma}_c = I \boldsymbol{\sigma}_a = i \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_a \end{pmatrix} , \quad I = i \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \quad (36.16)$$

The chiral matrix  $\gamma_5$  is

$$\gamma_5 \equiv -iI = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (36.17)$$

By construction, the chiral matrix  $\gamma_5$  is diagonal in the chiral representation.

The Newman-Penrose basis vectors in the chiral representation are the imaginary matrices

$$\boldsymbol{\gamma}_v = i \begin{pmatrix} 0 & \sigma_v \\ \sigma_u & 0 \end{pmatrix}, \quad \boldsymbol{\gamma}_u = i \begin{pmatrix} 0 & \sigma_u \\ \sigma_v & 0 \end{pmatrix}, \quad \boldsymbol{\gamma}_+ = i \begin{pmatrix} 0 & \sigma_+ \\ -\sigma_+ & 0 \end{pmatrix}, \quad \boldsymbol{\gamma}_- = i \begin{pmatrix} 0 & \sigma_- \\ -\sigma_- & 0 \end{pmatrix}, \quad (36.18)$$

where  $\sigma_m$  are the Newman-Penrose Pauli matrices

$$\sigma_v \equiv \frac{1}{\sqrt{2}}(1 + \sigma_3) = \sqrt{2} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \sigma_u \equiv \frac{1}{\sqrt{2}}(1 - \sigma_3) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad (36.19a)$$

$$\sigma_+ \equiv \frac{1}{\sqrt{2}}(\sigma_1 + i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_- \equiv \frac{1}{\sqrt{2}}(\sigma_1 - i\sigma_2) = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \quad (36.19b)$$

The Newman-Penrose bivectors form 6 real matrices that group into three right-handed bivectors (notation  $\boldsymbol{\gamma}_{mn} \equiv \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n$ ),

$$\boldsymbol{\gamma}_{v+} = \sqrt{2} \begin{pmatrix} \sigma_+ & 0 \\ 0 & 0 \end{pmatrix}, \quad \frac{1}{2}(\boldsymbol{\gamma}_{vu} - \boldsymbol{\gamma}_{+-}) = \begin{pmatrix} -\sigma_3 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{\gamma}_{u-} = \sqrt{2} \begin{pmatrix} \sigma_- & 0 \\ 0 & 0 \end{pmatrix}, \quad (36.20)$$

and three left-handed bivectors,

$$\boldsymbol{\gamma}_{u+} = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & -\sigma_+ \end{pmatrix}, \quad \frac{1}{2}(\boldsymbol{\gamma}_{vu} + \boldsymbol{\gamma}_{+-}) = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_3 \end{pmatrix}, \quad \boldsymbol{\gamma}_{v-} = \sqrt{2} \begin{pmatrix} 0 & 0 \\ 0 & -\sigma_- \end{pmatrix}. \quad (36.21)$$

### 36.3 Spinor basis elements

Introduce a tetrad of spinor basis elements  $\gamma_a$  (the convention of this book is to write spinors  $\gamma_a$  in normal face, not bold face),

$$\gamma_a \equiv \{\gamma_{V\uparrow}, \gamma_{U\downarrow}, \gamma_{U\uparrow}, \gamma_{V\downarrow}\}. \quad (36.22)$$

The indices  $\{V\uparrow, U\downarrow, U\uparrow, V\downarrow\}$  signify the transformation properties of the basis spinors:  $V$  and  $U$  signify boost weight  $+\frac{1}{2}$  and  $-\frac{1}{2}$ , while  $\uparrow$  and  $\downarrow$  signify spin weight  $+\frac{1}{2}$  and  $-\frac{1}{2}$ . The index notation, while non-standard, fits naturally with the Newman-Penrose  $\{v, u, +, -\}$  index notation. Under a Lorentz transformation, the basis spinors  $\gamma_a$  are defined to transform in the same way as rotors,

$$R : \gamma_a \rightarrow R\gamma_a. \quad (36.23)$$

In the chiral representation (36.15) the basis spinors  $\gamma_a$  are the column spinors

$$\gamma_{V\uparrow} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_{U\downarrow} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_{U\uparrow} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \gamma_{V\downarrow} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad (36.24)$$

which are Lorentz-transformed by pre-multiplying by rotors expressed in the chiral representation. The basis spinors  $\gamma_a$  in the chiral representation may be obtained from those in the Dirac representation by the transformation

$$X : \gamma_a \rightarrow X\gamma_a, \quad (36.25)$$

where the matrix  $X$  is defined by equation (36.14).

The basis spinors  $\gamma_a$  are eigenvectors of the chirality operator  $\gamma_5$ , equation (36.17), with eigenvalues  $\pm 1$ . 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,

$$\gamma_5\gamma_{V\uparrow} = \gamma_{V\uparrow}, \quad \gamma_5\gamma_{U\downarrow} = \gamma_{U\downarrow}, \quad \gamma_5\gamma_{U\uparrow} = -\gamma_{U\uparrow}, \quad \gamma_5\gamma_{V\downarrow} = -\gamma_{V\downarrow}. \quad (36.26)$$

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 (36.16). The right-handed basis spinors  $\gamma_{V\uparrow}$  and  $\gamma_{U\downarrow}$  are called right-handed because the boost axis and the spin axis point in the same direction (along the 3-direction for  $\gamma_{V\uparrow}$ , and along the negative 3-direction for  $\gamma_{U\downarrow}$ ). Conversely, the left-handed basis spinors  $\gamma_{U\uparrow}$  and  $\gamma_{V\downarrow}$  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  $\theta$  along the spin axis (3-axis) multiplies the basis spinors  $\gamma_a$  by  $e^{\pm\theta/2}$  according to

$$\gamma_{V\uparrow} \rightarrow e^{\theta/2}\gamma_{V\uparrow}, \quad \gamma_{U\downarrow} \rightarrow e^{-\theta/2}\gamma_{U\downarrow}, \quad \gamma_{U\uparrow} \rightarrow e^{-\theta/2}\gamma_{U\uparrow}, \quad \gamma_{V\downarrow} \rightarrow e^{\theta/2}\gamma_{V\downarrow}. \quad (36.27)$$

The transformations (36.27) confirm that the basis spinors with a  $V$  index have boost weight  $+\frac{1}{2}$ , while the basis spinors with a  $U$  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  $\theta$  about the spin axis (3-axis) multiplies the basis spinors  $\gamma_a$  by  $e^{\pm i\theta/2}$  according to

$$\gamma_{V\uparrow} \rightarrow e^{-i\theta/2}\gamma_{V\uparrow}, \quad \gamma_{U\downarrow} \rightarrow e^{i\theta/2}\gamma_{U\downarrow}, \quad \gamma_{U\uparrow} \rightarrow e^{-i\theta/2}\gamma_{U\uparrow}, \quad \gamma_{V\downarrow} \rightarrow e^{i\theta/2}\gamma_{V\downarrow}. \quad (36.28)$$

The transformations (36.28) 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

$$\boxed{\text{boost weight } n = \frac{1}{2} (\text{number of } V \text{ minus } U \text{ covariant indices})}, \quad (36.29a)$$

$$\boxed{\text{spin weight } s = \frac{1}{2} (\text{number of } \uparrow \text{ minus } \downarrow \text{ covariant indices})}, \quad (36.29b)$$

which generalize the rules (36.9) and (36.10). The rules (36.29) hold not only for column spinors  $\gamma_a$ , but also for row spinors  $\gamma_a \cdot$ , §36.7.3, and for inner and outer products of spinors, §36.5.3 and §36.6.1.

### 36.4 Dirac, Weyl, and Majorana spinors

A **Dirac spinor**  $\psi$  is a complex (with respect to *i*) linear combination of the 4 spinor basis elements  $\gamma_a$ ,

$$\psi = \psi^a \gamma_a . \quad (36.30)$$

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  $\psi^a \gamma_a$  is a spinor in the super spacetime algebra. Consistent with the convention for the geometric algebra, §13.8, a rotation of a spinor  $\psi$  (36.30) is an active rotation, which rotates the basis spinors  $\gamma_a$  while keeping the coefficients  $\psi^a$  fixed. Conceptually, the spinor is attached to the frame, and a rotation bodily rotates the frame and everything attached to it.

A Dirac spinor  $\psi$  Lorentz transforms as

$$R : \psi \rightarrow R\psi . \quad (36.31)$$

A Dirac spinor  $\psi$  is a spin- $\frac{1}{2}$  object, in the sense that a rotation by  $2\pi$  changes the sign of the spinor, and a rotation by  $4\pi$  is required to return the spinor to its original value.

**Concept question 36.2 Lorentz transformation of the phase of a spinor.** Should not a Lorentz transformation also change the phase of a spinor  $\psi$  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. Moreover any position-dependent phase  $e^{ip_\mu x^\mu}$  is a Lorentz scalar, unchanged by a Lorentz transformation.

#### 36.4.1 Weyl decomposition of a Dirac spinor

A Dirac spinor  $\psi$  can be decomposed into a sum of right- and left-handed chiral **Weyl spinors**  $\psi_R$  and  $\psi_L$

$$\psi = \psi_R + \psi_L , \quad (36.32)$$

that are right- and left-handed eigenvectors of the chiral operator  $\gamma_5$ ,

$$\gamma_5 \psi_L = \pm \psi_L . \quad (36.33)$$

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  $\psi$ ,

$$\psi_R = \frac{1}{2}(1 \pm \gamma_5)\psi . \quad (36.34)$$

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. For a Dirac spinor, the right- and left-handed components cannot be rotated into each other by any Lorentz transformation.

### 36.4.2 Majorana spinor

A **Majorana spinor** Lorentz transforms similarly to a Dirac spinor, but differs fundamentally from a Dirac spinor in that it satisfies a different scalar product, as will be seen in the next section, §36.5.

## 36.5 Spinor scalar product

### 36.5.1 Spinor metric tensor

In a matrix representation, the tensor product of spinor basis elements  $\gamma_a$  and  $\gamma_b$  can be represented as the matrix  $\gamma_a \gamma_b^\top$ , a matrix product of the column spinor  $\gamma_a$  with the row spinor  $\gamma_b^\top$ . In accordance with the transformation rule (36.23), the tensor product of basis spinors Lorentz transforms as

$$R : \gamma_a \gamma_b^\top \rightarrow R \gamma_a \gamma_b^\top R^\top . \quad (36.35)$$

Consider the spinor metric tensor  $\varepsilon$  with the defining property that for any Lorentz rotor  $R$

$$\varepsilon R^\top = \bar{R} \varepsilon . \quad (36.36)$$

That  $\varepsilon$  defines a Lorentz-invariant spinor metric will be seen in §36.5.3. A necessary and sufficient condition for (36.36) to hold is that, up to a choice of  $\mp$  sign on the right hand side, the tetrad basis vectors  $\gamma_m$  transform as

$$\varepsilon \gamma_m^\top = \begin{cases} -\gamma_m \varepsilon & \text{Dirac ,} \\ \gamma_m \varepsilon & \text{Majorana .} \end{cases} \quad (36.37)$$

The two different choices of sign lead to two different species of spinor, Dirac and Majorana spinors, whose dynamics are described by two different Lagrangians, Dirac and Majorana Lagrangians. In either the Dirac representation (14.92) or the chiral representation (36.15), the Dirac (Majorana) condition (36.37) requires that  $\varepsilon$  anticommutes (commutes) with  $\gamma_0$  and  $\gamma_2$ , and commutes (anticommutes) with  $\gamma_1$  and  $\gamma_3$ . The only basis element of the spacetime algebra with the required (anti)commutation properties is  $\sigma_2$  ( $I\sigma_2$ ) so the spinor tensor  $\varepsilon$  must equal  $\sigma_2$  ( $I\sigma_2$ ) up to a possible scalar normalization,

$$\varepsilon \equiv \begin{cases} i\sigma_2 & \text{Dirac ,} \\ I\sigma_2 & \text{Majorana .} \end{cases} \quad (36.38)$$

The normalization is chosen such that  $\varepsilon$  is real and orthogonal. Its square is minus the unit matrix,

$$\varepsilon^{-1} = \varepsilon^\top , \quad \varepsilon^2 = -1 . \quad (36.39)$$

The Dirac and Majorana spinor metric tensors  $\varepsilon_D$  and  $\varepsilon_M$  are related by

$$\varepsilon_D = \varepsilon_M \gamma_5 , \quad \varepsilon_M = \varepsilon_D \gamma_5 . \quad (36.40)$$

In the chiral representation (36.16), the spinor tensor (36.38) is the antisymmetric matrix

$$\varepsilon = \begin{cases} \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} & \text{Dirac ,} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} & \text{Majorana .} \end{cases} \quad (36.41)$$

Despite the equality of  $\varepsilon$  and  $i\sigma_2$  or  $I\sigma_2$  in the chiral and Dirac representations,  $\varepsilon$  is defined to transform as a spinor tensor under Lorentz transformations, not as an element of the spacetime algebra.

The condition (36.36) implies that the spinor tensor  $\varepsilon$  is invariant under Lorentz transformations,

$$R : \varepsilon \rightarrow R\varepsilon R^\top = R\bar{R}\varepsilon = \varepsilon . \quad (36.42)$$

The components of the spinor tensor  $\varepsilon$  define the antisymmetric spinor metric  $\varepsilon_{ab}$ ,

$$\gamma_a^\top \varepsilon \gamma_b = \varepsilon_{ab} . \quad (36.43)$$

Comment: the spinor metric tensor  $\varepsilon$  is sometimes called the charge conjugation operator in the physics literature. From the present perspective, the reasons are convoluted: one takes the adjoint spinor  $\bar{\psi}$ , here identified as the row conjugate spinor  $\bar{\psi}^\cdot$  (see §36.7.3), re-transposes it, and then pre-multiplies by  $\varepsilon^\top$  to recover the conjugate spinor  $\bar{\psi}$ . The role of  $\varepsilon$  as defining the spinor metric is more fundamental.

### 36.5.2 Row basis spinors

It is convenient to use the symbol  $\gamma_a \cdot$  with a trailing dot, symbolic of the trailing  $\varepsilon$ , to denote the row spinor  $\gamma_a^\top \varepsilon$ ,

$$\gamma_a \cdot \equiv \gamma_a^\top \varepsilon . \quad (36.44)$$

The motivation for the trailing dot notation is equation (36.48) below. The four row spinors

$$\gamma_a \cdot = \{\gamma_{V\uparrow} \cdot, \gamma_{U\downarrow} \cdot, \gamma_{U\uparrow} \cdot, \gamma_{V\downarrow} \cdot\} \quad (36.45)$$

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  $V$  index have boost weight  $+\frac{1}{2}$ , while basis spinors with a  $U$  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  $\gamma_a \cdot$  Lorentz transform as

$$R : \gamma_a \cdot \equiv \gamma_a^\top \varepsilon \rightarrow \gamma_a^\top R^\top \varepsilon = \gamma_a^\top \varepsilon \bar{R} = \gamma_a \cdot \bar{R} . \quad (36.46)$$

In the chiral representation (36.15) the row basis spinors  $\gamma_a \cdot$  are the row spinors

$$\gamma_{V\uparrow} \cdot = (0 \ 1 \ 0 \ 0), \quad \gamma_{U\downarrow} \cdot = (-1 \ 0 \ 0 \ 0), \quad \gamma_{U\uparrow} \cdot = (0 \ 0 \ 0 \mp 1), \quad \gamma_{V\downarrow} \cdot = (0 \ 0 \pm 1 \ 0), \quad (36.47)$$

where for the second pair the upper sign is Dirac, the lower sign Majorana.

### 36.5.3 Inner products of spinor basis elements

The product of the row spinor  $\gamma_a \cdot$  with the column spinor  $\gamma_b$  defines their inner product, or scalar product, which equals the **spinor metric**  $\varepsilon_{ab}$  in accordance with equation (36.43),

$$\boxed{\gamma_a \cdot \gamma_b = \varepsilon_{ab}}. \quad (36.48)$$

Equation (36.48) motivates the trailing dot notation for the row spinor. The scalar product is antisymmetric,

$$\gamma_a \cdot \gamma_b = -\gamma_b \cdot \gamma_a. \quad (36.49)$$

The scalar product (36.48) is a Lorentz scalar,

$$R : \gamma_a \cdot \gamma_b \rightarrow \gamma_a \cdot \bar{R}R \gamma_b = \gamma_a \cdot \gamma_b. \quad (36.50)$$

Thus the spinor metric  $\varepsilon_{ab}$  is Lorentz-invariant, just like the Minkowski metric  $\eta_{mn}$ .

### 36.5.4 Lowering and raising spinor indices

The antisymmetric spinor metric  $\varepsilon_{ab}$  is given in the chiral representation by equation (36.41). The inverse metric  $\varepsilon^{ab}$  is defined by  $\varepsilon^{ab}\varepsilon_{bc} = \delta_c^a$ . The spinor metric and its inverse satisfy

$$\varepsilon_{ab} = -\varepsilon_{ba} = -\varepsilon^{ab} = \varepsilon^{ba}. \quad (36.51)$$

Indices on a spinor tensor are lowered and raised by pre-multiplying by the metric  $\varepsilon_{ab}$  and its inverse  $\varepsilon^{ab}$ . The contravariant components  $\gamma^a$  of the column spinor basis elements, satisfying  $\gamma^a = \varepsilon^{ab}\gamma_b$ , are

$$\gamma^{V\uparrow} = -\gamma_{U\downarrow}, \quad \gamma^{U\downarrow} = \gamma_{V\uparrow}, \quad \gamma^{U\uparrow} = \pm\gamma_{V\downarrow}, \quad \gamma^{V\downarrow} = \mp\gamma_{U\uparrow}, \quad (36.52)$$

where for the second pair the upper sign is Dirac, the lower sign Majorana. For example,  $\gamma^{V\uparrow} = \varepsilon^{V\uparrow U\downarrow}\gamma_{U\downarrow} = -\gamma_{U\downarrow}$ . Post-multiplying by the metric or its inverse yields a result of opposite sign,  $\gamma^a = \varepsilon^{ab}\gamma_b = -\gamma_b\varepsilon^{ba}$ . The contravariant components  $\gamma^a \cdot$  of the row spinor basis elements satisfy the same relations (36.52) with a trailing dot appended on left and right hand sides. The scalar products of contravariant with covariant basis spinors form the unit matrix,

$$\gamma^a \cdot \gamma_b = -\gamma_b \cdot \gamma^a = \delta_b^a. \quad (36.53)$$

The scalar products of contravariant basis spinors are

$$\gamma^a \cdot \gamma^b = -\gamma^b \cdot \gamma^a = -\varepsilon^{ab}. \quad (36.54)$$

### 36.5.5 Scalar products of spinors

A general row Dirac or Majorana spinor  $\psi \cdot$  is a complex (with respect to  $i$ ) linear combination of the 4 row basis spinors

$$\boxed{\psi \cdot \equiv \psi^\top \varepsilon = \psi^a \gamma_a \cdot} . \quad (36.55)$$

It Lorentz transforms as

$$R : \psi \cdot \rightarrow \psi \cdot \bar{R} . \quad (36.56)$$

A row spinor  $\psi \cdot$  transforms like a reverse rotor.

The scalar product of a row spinor  $\psi \cdot = \psi^a \gamma_a \cdot$  with a column spinor  $\chi = \chi^a \gamma_a$  may be written variously

$$\psi \cdot \chi = \psi^\top \varepsilon \chi = \psi^a \gamma_a \cdot \chi^b \gamma_b = \varepsilon_{ab} \psi^a \chi^b = \psi^a \chi_a = -\psi_a \chi^a = -\varepsilon^{ab} \psi_a \chi_b . \quad (36.57)$$

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  $\psi^a \gamma_a$  of a spinor in components similarly changes the sign,

$$\psi = \psi^a \gamma_a = \psi^a \varepsilon_{ab} \gamma^b = -\psi_a \gamma^a . \quad (36.58)$$

The components  $\psi^a$  of a column spinor  $\psi$  can be projected out by pre-multiplying by the row basis spinor  $\gamma_a \cdot$ ,

$$\gamma^a \cdot \psi = \gamma^a \cdot \psi^b \gamma_b = \delta_b^a \psi^b = \psi^a . \quad (36.59)$$

The components  $\psi^a$  of a row spinor  $\psi \cdot$  can be projected out by post-multiplying by minus the column basis spinor  $\gamma^a$ ,

$$-\psi \cdot \gamma^a = -\psi^b \gamma_b \cdot \gamma^a = \delta_b^a \psi^b = \psi^a . \quad (36.60)$$

## 36.6 Super spacetime algebra

### 36.6.1 Outer products of spinor basis elements

A row spinor  $\gamma_a \cdot$  multiplied by a column spinor  $\gamma_b$  yields their scalar product. In the opposite order, a column spinor  $\gamma_a$  multiplied by a row spinor  $\gamma_b \cdot$  yields their outer product. The outer product  $\gamma_a \gamma_b \cdot$  Lorentz transforms like a multivector in the spacetime algebra,

$$R : \gamma_a \gamma_b \cdot \equiv \gamma_a \gamma_b^\top \varepsilon \rightarrow R \gamma_a \gamma_b^\top R^\top \varepsilon = R \gamma_a \gamma_b^\top \varepsilon \bar{R} = R \gamma_a \gamma_b \cdot \bar{R} . \quad (36.61)$$

The trailing dot on the outer product  $\gamma_a \gamma_b \cdot$  is symbolic of the trailing  $\varepsilon$ , necessary to convert the spinor tensor  $\gamma_a \gamma_b^\top$  into an object that transforms like a multivector.

The products of the 4 column basis spinors  $\gamma_a$  with the 4 row basis spinors  $\gamma_b \cdot$  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 Dirac and Majorana super spacetime algebras differ because the Dirac and Majorana spinor metric tensors  $\varepsilon$  differ. The algebras differ by various sign flips in the relations between multivectors and outer products of spinors. A notable difference is that the basis vectors  $\gamma_m$  are given by symmetric outer products of spinors for Dirac, but by antisymmetric outer products of spinors for Majorana, equations (36.68) and (36.69).

In the chiral representation (36.15), the 16 outer products of basis spinors map to elements of the spacetime algebra as follows. The boost and spin weights of the left and right hand sides of each of equations (36.62)–(36.71) below match, as they should. The antisymmetric outer products of right-handed spinors form a right-handed scalar singlet,

$$[\gamma_{U\downarrow}, \gamma_{V\uparrow}] \cdot = \frac{1}{2}(1 + \gamma_5) . \quad (36.62)$$

The trailing dot on the commutator indicates that the right partner of each product is a row spinor,  $[\gamma_{V\uparrow}, \gamma_{U\downarrow}] \cdot = \gamma_{V\uparrow} \gamma_{U\downarrow}^\top \cdot - \gamma_{U\downarrow} \gamma_{V\uparrow}^\top \cdot$ . Similarly the antisymmetric outer products of left-handed spinors form a left-handed scalar singlet,

$$[\gamma_{U\uparrow}, \gamma_{V\downarrow}] \cdot = \pm \frac{1}{2}(1 - \gamma_5) , \quad (36.63)$$

where the upper (+) sign is for Dirac, the lower (−) sign for Majorana. The symmetric outer products of right-handed spinors form a triplet of right-handed bivectors,

$$\{\gamma_{V\uparrow}, \gamma_{V\uparrow}\} \cdot = \gamma_{v+} , \quad \{\gamma_{V\uparrow}, \gamma_{U\downarrow}\} \cdot = \frac{1}{2}(\gamma_{vu} - \gamma_{+-}) , \quad \{\gamma_{U\downarrow}, \gamma_{U\downarrow}\} \cdot = -\gamma_{u-} . \quad (36.64)$$

The symmetric outer products of left-handed spinors form a triplet of left-handed bivectors,

$$\{\gamma_{U\uparrow}, \gamma_{U\uparrow}\} \cdot = \pm \gamma_{u+} , \quad \{\gamma_{U\uparrow}, \gamma_{V\downarrow}\} \cdot = \pm \frac{1}{2}(\gamma_{vu} + \gamma_{+-}) , \quad \{\gamma_{V\downarrow}, \gamma_{V\downarrow}\} \cdot = \mp \gamma_{v-} , \quad (36.65)$$

where the upper sign is for Dirac, the lower sign for Majorana. The 4 outer products of right- with left-handed spinors yield the 4 right-handed vectors

$$\begin{aligned} \gamma_{V\uparrow} \gamma_{V\downarrow} \cdot &= \mp \frac{i}{2\sqrt{2}}(1 + \gamma_5) \gamma_v , & \gamma_{U\downarrow} \gamma_{U\uparrow} \cdot &= \pm \frac{i}{2\sqrt{2}}(1 + \gamma_5) \gamma_u , \\ \gamma_{V\uparrow} \gamma_{U\uparrow} \cdot &= \pm \frac{i}{2\sqrt{2}}(1 + \gamma_5) \gamma_+ , & \gamma_{U\downarrow} \gamma_{V\downarrow} \cdot &= \mp \frac{i}{2\sqrt{2}}(1 + \gamma_5) \gamma_- , \end{aligned} \quad (36.66)$$

where the upper sign is for Dirac, the lower sign for Majorana. The 4 outer products of left- with right-handed spinors yield the 4 left-handed vectors

$$\begin{aligned} \gamma_{V\downarrow} \gamma_{V\uparrow} \cdot &= -\frac{i}{2\sqrt{2}}(1 - \gamma_5) \gamma_v , & \gamma_{U\uparrow} \gamma_{U\downarrow} \cdot &= \frac{i}{2\sqrt{2}}(1 - \gamma_5) \gamma_u , \\ \gamma_{U\uparrow} \gamma_{V\uparrow} \cdot &= \frac{i}{2\sqrt{2}}(1 - \gamma_5) \gamma_+ , & \gamma_{V\downarrow} \gamma_{U\downarrow} \cdot &= -\frac{i}{2\sqrt{2}}(1 - \gamma_5) \gamma_- . \end{aligned} \quad (36.67)$$

The chiral vectors combine in symmetric and antisymmetric combinations to form the 4 basis vectors and 4 basis pseudovectors. For Dirac the 4 symmetric outer products of right- with left-handed spinors yield the 4 basis vectors,

$$\{\gamma_{V\uparrow}, \gamma_{V\downarrow}\} \cdot = -\frac{i}{\sqrt{2}} \gamma_v , \quad \{\gamma_{U\downarrow}, \gamma_{U\uparrow}\} \cdot = \frac{i}{\sqrt{2}} \gamma_u , \quad \{\gamma_{V\uparrow}, \gamma_{U\uparrow}\} \cdot = \frac{i}{\sqrt{2}} \gamma_+ , \quad \{\gamma_{U\downarrow}, \gamma_{V\downarrow}\} \cdot = -\frac{i}{\sqrt{2}} \gamma_- , \quad (36.68)$$

whereas for Majorana the 4 antisymmetric outer products of right- with left-handed spinors yield the 4 basis vectors,

$$[\gamma_{V\uparrow}, \gamma_{V\downarrow}] \cdot = \frac{i}{\sqrt{2}} \boldsymbol{\gamma}_v, \quad [\gamma_{U\downarrow}, \gamma_{U\uparrow}] \cdot = -\frac{i}{\sqrt{2}} \boldsymbol{\gamma}_u, \quad [\gamma_{V\uparrow}, \gamma_{U\uparrow}] \cdot = -\frac{i}{\sqrt{2}} \boldsymbol{\gamma}_+, \quad [\gamma_{U\downarrow}, \gamma_{V\downarrow}] \cdot = \frac{i}{\sqrt{2}} \boldsymbol{\gamma}_-. \quad (36.69)$$

Conversely, for Dirac the 4 antisymmetric outer products of right- with left-handed spinors yield the 4 basis pseudovectors,

$$[\gamma_{V\uparrow}, \gamma_{V\downarrow}] \cdot = -\frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_v, \quad [\gamma_{U\downarrow}, \gamma_{U\uparrow}] \cdot = \frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_u, \quad [\gamma_{V\uparrow}, \gamma_{U\uparrow}] \cdot = \frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_+, \quad [\gamma_{U\downarrow}, \gamma_{V\downarrow}] \cdot = -\frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_-. \quad (36.70)$$

whereas for Majorana the 4 symmetric outer products of right- with left-handed spinors yield the 4 basis pseudovectors,

$$\{\gamma_{V\uparrow}, \gamma_{V\downarrow}\} \cdot = \frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_v, \quad \{\gamma_{U\downarrow}, \gamma_{U\uparrow}\} \cdot = -\frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_u, \quad \{\gamma_{V\uparrow}, \gamma_{U\uparrow}\} \cdot = -\frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_+, \quad \{\gamma_{U\downarrow}, \gamma_{V\downarrow}\} \cdot = \frac{i}{\sqrt{2}} \gamma_5 \boldsymbol{\gamma}_-. \quad (36.71)$$

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 (36.62) and (36.63) is the unit matrix, whose trace is 4),

$$\text{Tr } \gamma_a \gamma_b \cdot = \gamma_b \cdot \gamma_a = \varepsilon_{ba}. \quad (36.72)$$

The expansion of the 16 outer products  $\gamma_a \gamma_b \cdot$  of spinors in terms of the 16 basis elements  $\boldsymbol{\gamma}_A$  of the spacetime algebra, and vice versa, define the matrix of coefficients  $\gamma_{ab}^A$  and its inverse  $\gamma_A^{ab}$ ,

$$\gamma_a \gamma_b \cdot = \gamma_{ab}^A \boldsymbol{\gamma}_A, \quad \boldsymbol{\gamma}_A = \gamma_A^{ab} \gamma_a \gamma_b \cdot. \quad (36.73)$$

The coefficients  $\gamma_{ab}^A$  in the chiral representation can be read off from equations (36.62)–(36.71). In the chiral representation, the coefficients for odd elements of the spacetime algebra (vectors and pseudovectors) are all imaginary, while coefficients for even elements are all real.

**Concept question 36.3 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

$$\varphi_R = \frac{1}{2}(1 + \gamma_5)\varphi, \quad (36.74)$$

where  $\varphi$  is a complex scalar. The chiral operator  $\gamma_5$  is not a scalar, but rather a totally antisymmetric tensor of rank 4. The Newman-Penrose expression (36.117) for  $\gamma_5$  shows that it has boost and spin weight zero.

### 36.6.2 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  $1 \times 1$ ,  $1 \times 4$ ,  $4 \times 1$ , and  $4 \times 4$ . The super spacetime algebra consists of arbitrary sums and products of all 4 species.

The true scalars are just complex numbers. A column spinor  $\psi$  is a complex linear combination of column basis spinors  $\gamma_a$ ,

$$\psi = \psi^a \gamma_a , \quad (36.75)$$

while a row spinor  $\psi^\cdot$  is a complex linear combination of row basis spinors  $\gamma_a^\cdot$ ,

$$\psi^\cdot = \psi^a \gamma_a^\cdot . \quad (36.76)$$

A multivector  $a$  is a complex linear combination of outer products of the column and row basis spinors,

$$a = a^{ab} \gamma_a \gamma_b^\cdot . \quad (36.77)$$

Linearity and the transformation law (36.61) imply that the algebra of sums and products of outer products of spinors is isomorphic to the spacetime algebra.

It might seem puzzling that there are two distinct kinds of scalar in the super spacetime algebra, true scalars, and multivector scalars proportional to the unit matrix in a matrix representation. In fact both species of scalar appear in physics. An example of a true scalar is the action to which the principle of least action is applied. An example of a multivector scalar is a scalar field. In §37.4 it will be seen that the covariant spacetime derivative of a true scalar is another true scalar, whereas the covariant spacetime derivative of a multivector scalar is a vector.

As seen in §36.5.3 and §36.6.1, a column spinor  $\psi$  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 facilitates interpretation of the column and row spinors as creation and annihilation operators) 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,

$$\psi \chi = 0 , \quad \psi^\cdot \chi^\cdot = 0 . \quad (36.78)$$

Similarly, a multivector  $a$  can only pre-multiply a column spinor  $\psi$ , and can only post-multiply a row spinor  $\psi^\cdot$ , as is again manifestly true in a matrix representation. Thus a multivector  $a$  post-multiplying a column spinor or pre-multiplying a row spinor yields zero,

$$\psi a = 0 , \quad a \psi^\cdot = 0 . \quad (36.79)$$

In general, a sequence of products of spinors yields a non-zero result provided that they alternate between column spinor and row spinor,

$$\psi \chi^\cdot \varphi \quad \text{or} \quad \psi^\cdot \chi \varphi^\cdot . \quad (36.80)$$

Both product sequences are associative,

$$\psi \chi^\cdot \varphi = (\psi \chi^\cdot) \varphi = \psi (\chi^\cdot \varphi) , \quad (36.81)$$

and

$$\psi^\cdot \chi \varphi^\cdot = (\psi^\cdot \chi) \varphi^\cdot = \psi^\cdot (\chi \varphi^\cdot) . \quad (36.82)$$

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

$$\psi \cdot \mathbf{a} \chi \quad (36.83)$$

are admissible, and in general non-vanishing.

The trace of an outer product of spinors is a true scalar

$$\text{Tr } \psi \chi \cdot = \psi^a \chi^b \varepsilon_{ba} = -\psi \cdot \chi . \quad (36.84)$$

### 36.6.3 Fierz identities

The associative law and the scalar product make it straightforward to simplify long strings of products of spinors and multivectors, CHECK a process known in quantum field theory as Fierz rearrangement.

Let  $\mathbf{a} = a^{ab} \gamma_a \gamma_b \cdot$  and  $\mathbf{b} = b^{ab} \gamma_a \gamma_b \cdot$  be two multivectors expressed as a sum of outer products of spinors. Their product is the multivector

$$\mathbf{a} \mathbf{b} = a^{ab} \gamma_a \gamma_b \cdot b^{cd} \gamma_c \gamma_d \cdot = \gamma_a a^{ab} \varepsilon_{bc} b^{cd} \gamma_d \cdot = \gamma_a a^{ab} b_b^d \gamma_d \cdot . \quad (36.85)$$

A sequence of multivectors sandwiched by spinors simplifies as

$$\psi \cdot \mathbf{a} \mathbf{b} \chi = \psi^a \gamma_a \cdot a^{bc} \gamma_b \gamma_c \cdot b^{de} \gamma_d \gamma_e \cdot \chi^f \gamma_f = \psi^a \varepsilon_{ab} a^{bc} \varepsilon_{cd} b^{de} \varepsilon_{ef} \chi^f = \psi^a a_a^c b_c^e \chi_e . \quad (36.86)$$

### 36.6.4 Column and row spinors are creation and annihilation operators

In REF, column and row spinors will emerge as describing respectively creation and annihilation operators in the quantum field theory of Dirac spinors.

The correct fermionic behaviour is built into the super spacetime algebra. The antisymmetry of the scalar product gives the correct anticommuting behaviour of spinors. The prescription that a column spinor cannot be multiplied by a column spinor prevents two spinors from being created in the same state. The prescription that a row spinor cannot be multiplied by a row spinor prevents two spinors from being annihilated out of the same state. A row spinor multiplying a column spinor corresponds to annihilation following creation, while a column spinor multiplying a row spinor corresponds to creation following annihilation, and these are allowed.

## 36.7 $C$ -conjugation

The super spacetime algebra possesses a discrete transformation called  $C$ -conjugation (or simply conjugation, when there is no ambiguity). The  $C$ -conjugate Dirac spinor  $\bar{\psi}$  is defined, equation (36.95), such that (a) its components are complex conjugates of those of the parent spinor  $\psi$ , (b) it has opposite chirality and spin to  $\psi$ , and (c) it Lorentz transforms according to the complex conjugate representation of the  $\gamma$ -matrices.

### 36.7.1 Conjugation operator $C$

Consider the conjugation operator  $C$  with the defining property that it converts any Lorentz rotor  $R$  in the chiral or Dirac representations to its complex conjugate (with respect to  $i$ ),

$$CRC^{-1} = R^* . \quad (36.87)$$

A Lorentz rotor  $R$  is a real (with respect to  $i$ ) linear combination of even Minkowski basis elements of the spacetime algebra, so a necessary and sufficient condition for (36.87) to hold is that  $C$  converts the basis Minkowski vectors  $\gamma_m$  to their complex conjugates, up to a possible sign,

$$C\gamma_m C^{-1} = \pm \gamma_m^* . \quad (36.88)$$

The  $\pm$  sign on the right hand side of equation (36.88) may be chosen positive without loss of generality, since it is actually the product of the signs on the right hand sides of equations (36.37) and (36.88) that distinguishes Dirac ( $-$ ) and Majorana ( $+$ ) spinors (Majorana has an extra chiral factor  $\gamma_5$  in one of  $\varepsilon$  or  $C$ , compared to Dirac). The conjugation operator  $C$  is therefore defined such that it converts the Dirac or chiral matrix representations of the Minkowski basis vectors  $\gamma_m$  to their complex conjugates (with respect to  $i$ ),

$$C\gamma_m C^{-1} = \gamma_m^* , \quad (36.89)$$

which holds for both Dirac and Majorana spinors. In either the Dirac representation (14.91) or the chiral representation (36.15), the condition (36.89) requires that  $C$  anticommute with  $\gamma_0$ ,  $\gamma_1$ , and  $\gamma_3$ , and commute with  $\gamma_2$ . The only basis element of the spacetime algebra with the required (anti)commutation properties is  $\gamma_2$ , so  $C$  must equal  $\gamma_2$  up to a possible scalar normalization,

$$C = \gamma_2 . \quad (36.90)$$

The chosen normalization is such that  $C$  is real and orthogonal. Its square is the unit matrix,

$$C^{-1} = C^\top , \quad C^2 = 1 . \quad (36.91)$$

In the chiral representation,

$$C = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} . \quad (36.92)$$

Notwithstanding the equality of  $C$  and  $\gamma_2$  in the Dirac or chiral representations, the conjugation operator  $C$  is defined such that it transforms as a Lorentz scalar, not as an element of the spacetime algebra. This is accomplished mathematically by demanding  $C$  to be a spinor tensor that Lorentz transforms as  $C \rightarrow RC\bar{R}^*$ . The spinor tensor that transforms in this way and remains invariant under Lorentz transformation is, up to a scalar or pseudoscalar factor, precisely the conjugation operator  $C$ .

The conjugation operator anticommutes with the Dirac spinor metric tensor, but commutes with the Majorana spinor metric tensor,

$$C\varepsilon = \mp \varepsilon C , \quad (36.93)$$

with upper (–) sign for Dirac, lower (+) sign for Majorana.

### 36.7.2 $C$ -conjugate spinor

The complex conjugate  $\psi^*$  of a Dirac or Majorana spinor  $\psi = \psi^a \gamma_a$  is defined to be the spinor with complex-conjugated (with respect to  $i$ ) coefficients in the Dirac or chiral matrix representation of the spinor,

$$\psi^* \equiv \psi^{a*} \gamma_a . \quad (36.94)$$

In effect, the spinor basis elements  $\gamma_a$  are taken to be real in the Dirac or chiral representations, in contrast to the basis vectors  $\gamma_m$ , whose Dirac or chiral matrix representations are complex. Complex conjugation leaves the chirality and spin of a spinor unchanged.

The  $C$ -conjugate Dirac or Majorana spinor  $\bar{\psi}$  is defined by

$$\boxed{\bar{\psi} \equiv C\psi^*} , \quad (36.95)$$

where  $C$  is the conjugation operator defined by equation (36.89). The  $C$ -conjugate spinor  $\bar{\psi}$  is sometimes called the antiparticle corresponding to the particle spinor  $\psi$ , but this is not quite correct: the physical (positive mass) antiparticle is the  $CPT$ -conjugate, §38.3. The  $C$ -conjugate spinor  $\bar{\psi}$  is also commonly called the charge conjugate spinor, because its (electric) charge is opposite to that of the spinor  $\psi$ , equation (38.28). However, the conjugate spinor also has opposite chirality and opposite spin, so the name charge conjugate understates the case.

If  $\psi = \psi^a \gamma_s$ , then the  $C$ -conjugate spinor  $\bar{\psi}$  is

$$\bar{\psi} = \psi^{a*} \bar{\gamma}_a , \quad (36.96)$$

where the  $C$ -conjugate basis elements  $\bar{\gamma}_a$  are, in either an orthonormal or Newman-Penrose basis, and in either the Dirac or chiral representations,

$$\bar{\gamma}_a = C\gamma_a . \quad (36.97)$$

In view of equation (36.87), the conjugate spinor  $\bar{\psi}$  Lorentz transforms according to the complex conjugate representation of the  $\gamma$ -matrices,

$$R : \bar{\psi} \equiv C\psi^* \rightarrow CR\psi^* = R^*C\psi^* = R^*\bar{\psi} . \quad (36.98)$$

Conjugating a spinor  $\psi$  twice recovers the original spinor,

$$\bar{\bar{\psi}} = \psi . \quad (36.99)$$

The components  $\bar{\psi}^a$  of the conjugate spinor are projected out by pre-multiplying by the row spinor  $\gamma^a \cdot$ , equation (36.59),

$$\bar{\psi}^a = \gamma^a \cdot \bar{\psi} . \quad (36.100)$$

In the chiral representation, the components of the spinor and its conjugate are related by

$$\psi^a = \begin{pmatrix} \psi^{V\uparrow} \\ \psi^{U\downarrow} \\ \psi^{U\uparrow} \\ \psi^{V\downarrow} \end{pmatrix}, \quad \bar{\psi}^a = \begin{pmatrix} \psi^{V\downarrow*} \\ -\psi^{U\uparrow*} \\ -\psi^{U\downarrow*} \\ \psi^{V\uparrow*} \end{pmatrix}. \quad (36.101)$$

Equations (36.101) show that conjugation flips chirality and spin; for example, the conjugate of a right-handed spin-up spinor is a left-handed spin-down spinor.

If a spinor  $\psi$  is represented by a complex (with respect to  $I$ ) quaternion

$$\psi = \left\{ \begin{array}{cccc} w_R & x_R & y_R & z_R \\ w_I & x_I & y_I & z_I \end{array} \right\}, \quad (36.102)$$

then from  $\bar{\psi} = \varepsilon(\bar{\psi}\cdot)^T$  and equation (14.110) (valid for Dirac spinors), the conjugate spinor  $\bar{\psi}$  is represented by the complex quaternion

$$\bar{\psi} = \left\{ \begin{array}{cccc} -x_I & w_I & -z_I & y_I \\ x_R & -w_R & z_R & -y_R \end{array} \right\}. \quad (36.103)$$

Equation (36.103) holds for both Dirac and Majorana spinors because the definition (36.95) of the conjugate spinor  $\bar{\psi}$  is the same for both Dirac and Majorana spinors, and the conjugation operator  $C$ , equation (36.90), is the same for both. The *CPT*-conjugate spinor  $I\bar{\psi}$  is represented by the complex quaternion

$$I\bar{\psi} = \left\{ \begin{array}{cccc} -x_R & w_R & -z_R & y_R \\ -x_I & w_I & -z_I & y_I \end{array} \right\}. \quad (36.104)$$

### 36.7.3 Row conjugate spinor

The Dirac or Majorana row conjugate spinor  $\bar{\psi}\cdot$  corresponding to the column conjugate spinor  $\bar{\psi}$  is

$$\bar{\psi}\cdot \equiv \bar{\psi}^\top \varepsilon = \psi^\dagger C^\top \varepsilon = \begin{cases} -i\psi^\dagger \gamma_0 & \text{Dirac,} \\ \psi^\dagger I\gamma_0 & \text{Majorana.} \end{cases} \quad (36.105)$$

The row conjugate spinor  $\bar{\psi}\cdot$  coincides with what is commonly called the adjoint spinor  $\bar{\psi}$  in Dirac or Majorana theory. The column conjugate spinor  $\bar{\psi}$  is not the same as the conventional adjoint spinor. However, the translation from the conventional to the present notation is trivial: wherever the conventional adjoint spinor  $\bar{\psi}$  appears, replace it by the row conjugate spinor  $\bar{\psi}\cdot$ .

Equation (36.105) implies that

$$C^\top \varepsilon = \begin{cases} -i\gamma_0 & \text{Dirac,} \\ I\gamma_0 & \text{Majorana.} \end{cases} \quad (36.106)$$

Equation (36.106) holds in the chiral and Dirac representations, but in fact it must hold in any satisfactory representation of Dirac or Majorana spinors governed by the Dirac or Majorana Lagrangians (38.26) or

(38.67), in order that the Dirac number current density  $n^0 \equiv i\bar{\psi} \cdot \boldsymbol{\gamma}^0 \psi$ , equation (38.16), or Majorana number current density  $n^0 \equiv I\bar{\psi} \cdot \boldsymbol{\gamma}^0 \psi$ , equation (38.76), equal a positive number  $\psi^\dagger \psi$ .

The spinor metric  $\varepsilon$  and conjugation operator  $C$  are defined by their actions (36.37) and (36.89) on the Minkowski basis vectors  $\boldsymbol{\gamma}_m$ , namely  $\boldsymbol{\gamma}_m^\top = \mp\varepsilon^{-1}\boldsymbol{\gamma}_m\varepsilon$  and  $\boldsymbol{\gamma}_m^* = C\boldsymbol{\gamma}_m C^{-1}$ . If equation (36.106) holds, as it must do, and if in addition  $C$  is real and unitary,  $C^{-1} = C^\top$ , as is true in the chiral or Dirac representations, then the Hermitian conjugates of the basis vectors  $\boldsymbol{\gamma}_m$  satisfy

$$\boldsymbol{\gamma}_m^\dagger = (\boldsymbol{\gamma}_m^*)^\top = \begin{cases} -\varepsilon^{-1}C\boldsymbol{\gamma}_m C^{-1}\varepsilon = -(C^\top\varepsilon)^{-1}\boldsymbol{\gamma}_m(C^\top\varepsilon) = -(\boldsymbol{\gamma}_0)^{-1}\boldsymbol{\gamma}_m(\boldsymbol{\gamma}_0) \\ \varepsilon^{-1}C\boldsymbol{\gamma}_m C^{-1}\varepsilon = (C^\top\varepsilon)^{-1}\boldsymbol{\gamma}_m(C^\top\varepsilon) = (I\boldsymbol{\gamma}_0)^{-1}\boldsymbol{\gamma}_m(I\boldsymbol{\gamma}_0) \end{cases} = \boldsymbol{\gamma}_0\boldsymbol{\gamma}_m\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}^m, \quad (36.107)$$

where the upper line is for Dirac, the lower for Majorana. Equation (36.107) shows that  $\boldsymbol{\gamma}_m$  are unitary, which is the condition (14.89) originally adopted for the Dirac  $\gamma$ -matrices.

### 36.7.4 $C$ -conjugate multivector

The complex conjugate (with respect to  $i$ )  $\mathbf{a}^*$  of a multivector  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  is defined to be, in either the Dirac or chiral representation (and in either an orthonormal or Newman-Penrose basis),

$$\mathbf{a}^* \equiv a^{A*} \boldsymbol{\gamma}_A^*. \quad (36.108)$$

So defined, complex conjugation is multiplicative over multivectors and spinors,

$$(\mathbf{a}\psi)^* = \mathbf{a}^*\psi^*, \quad (36.109)$$

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,

$$(\psi\chi\cdot)^* = \psi^*\chi^*\cdot. \quad (36.110)$$

Complex conjugation leaves the chirality and spin of a multivector unchanged.

The  $C$ -conjugate multivector  $\bar{\mathbf{a}}$  (not to be confused with the reverse multivector  $\overline{\mathbf{a}}$ ) of a multivector  $\mathbf{a}$  is defined to be

$$\boxed{\bar{\mathbf{a}} \equiv C\mathbf{a}^*C}. \quad (36.111)$$

$C$ -conjugation is multiplicative over multivectors and spinors,

$$\overline{\mathbf{a}\psi} = \bar{\mathbf{a}}\bar{\psi}. \quad (36.112)$$

The  $C$ -conjugate of a multivector that is an outer product of spinors is minus (Dirac) or plus (Majorana) the outer product of the conjugate spinors,

$$\overline{\psi\chi\cdot} = \begin{cases} -\bar{\psi}\bar{\chi}\cdot & \text{Dirac} \\ \bar{\psi}\bar{\chi}\cdot & \text{Majorana} \end{cases} \quad (36.113)$$

The sign comes from the (anti)commutation of the conjugation operator with the spinor metric tensor, equation (36.93).

If  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$ , then the  $C$ -conjugate multivector  $\bar{\mathbf{a}}$  is

$$\bar{\mathbf{a}} = a^{A*} \bar{\boldsymbol{\gamma}}_A , \quad (36.114)$$

where the  $C$ -conjugate basis elements  $\bar{\boldsymbol{\gamma}}_A$  are, in either an orthonormal or Newman-Penrose basis, and in either the Dirac or chiral representations,

$$\bar{\boldsymbol{\gamma}}_A = C \boldsymbol{\gamma}_A^* C . \quad (36.115)$$

The  $C$ -conjugates of orthonormal basis elements are equal to themselves,  $\bar{\boldsymbol{\gamma}}_A = \boldsymbol{\gamma}_A$ , but the  $C$ -conjugates of Newman-Penrose basis vectors are not equal to themselves.

Just as the  $C$ -conjugate spinor  $\bar{\psi}$  has opposite chirality and spin to  $\psi$ , so also the  $C$ -conjugate multivector  $\bar{\mathbf{a}}$  has opposite chirality and spin to  $\mathbf{a}$ . For Newman-Penrose,  $C$ -conjugation flips the spin indices  $+\leftrightarrow-$  of the Newman-Penrose basis vectors  $\boldsymbol{\gamma}_m$ ,

$$\bar{\boldsymbol{\gamma}}_v = \boldsymbol{\gamma}_v , \quad \bar{\boldsymbol{\gamma}}_u = \boldsymbol{\gamma}_u , \quad \bar{\boldsymbol{\gamma}}_+ = \boldsymbol{\gamma}_- , \quad \bar{\boldsymbol{\gamma}}_- = \boldsymbol{\gamma}_+ , \quad (36.116)$$

as can be verified by direct calculation from the matrices (36.18) and (36.92). This is true in general: the  $C$ -conjugate of any Newman-Penrose basis multivector  $\boldsymbol{\gamma}_A$  is obtained by flipping its spin indices  $+\leftrightarrow-$ . The chiral matrix  $\gamma_5$  expressed in the Newman-Penrose tetrad is

$$\gamma_5 \equiv -iI = -\frac{i}{4!} \varepsilon^{vu+-} \boldsymbol{\gamma}_v \boldsymbol{\gamma}_u \boldsymbol{\gamma}_+ \boldsymbol{\gamma}_- = -\boldsymbol{\gamma}_v \wedge \boldsymbol{\gamma}_u \wedge \boldsymbol{\gamma}_+ \wedge \boldsymbol{\gamma}_- , \quad (36.117)$$

where the imaginary factor  $i$  in the definition of  $\gamma_5$  cancels against the imaginary determinant of the transformation from Minkowski to Newman-Penrose tetrad, leaving a real factor in the rightmost expression of equations (36.117). Conjugation flips the sign of the chiral operator  $\gamma_5$ ,

$$\bar{\gamma}_5 = -\gamma_5 . \quad (36.118)$$

### 36.7.5 Real multivector

Conventionally, a multivector  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  is said to be real if its  $C$ -conjugate is itself (the overbar here denotes the  $C$ -conjugate, not the reverse),

$$\bar{\mathbf{a}} = \mathbf{a} . \quad (36.119)$$

In an orthonormal basis, the  $C$ -conjugates of the basis elements are themselves,  $\bar{\boldsymbol{\gamma}}_A = \boldsymbol{\gamma}_A$ , and a multivector  $\mathbf{a}$  is then real if and only if the coefficients  $a^A$  of its expansion  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  in the orthonormal basis are real.

Most classical multivectors are real. For example, derivatives are real, Lorentz rotors are real, the electromagnetic field is real.

## 36.8 (Anti)symmetry of the scalar product of spinors

The complex conjugate of the scalar product of Dirac or Majorana spinors satisfies

$$(\bar{\psi} \cdot \chi)^* = \mp \psi \cdot \bar{\chi} , \quad (36.120)$$

where the upper (–) sign is for Dirac spinors, the lower (+) sign for Majorana spinors. The sign comes from the anticommutation (Dirac) or commutation (Majorana) of the conjugation operator  $C$  with the spinor metric tensor  $\varepsilon$ , equation (36.93). In particular the complex conjugate of the scalar product of a spinor  $\psi$  with its conjugate  $\bar{\psi}$  is

$$(\bar{\psi} \cdot \psi)^* = \mp \psi \cdot \bar{\psi}. \quad (36.121)$$

The Dirac and Majorana Lagrangians (38.25) and (38.67) involve mass terms proportional to  $\bar{\psi} \cdot \psi$  (for Majorana the mass term involves an extra factor of the pseudoscalar  $I$ , but the factor does not affect complex conjugation of the scalar mass term, equation (38.69)). If it is insisted that the mass term be non-vanishing (or alternatively that a charge term be non-vanishing), then reality of the Lagrangian requires that the mass term be real, and consequently

$$\bar{\psi} \cdot \psi = \mp \psi \cdot \bar{\psi}. \quad (36.122)$$

with upper (–) sign for Dirac, lower (+) for Majorana. It follows that the spinor scalar product must be anticommuting for Dirac, and commuting for Majorana,

$$\begin{aligned} \psi \cdot \chi &= -\chi \cdot \psi && \text{Dirac}, \\ \psi \cdot \chi &= \chi \cdot \psi && \text{Majorana}. \end{aligned} \quad (36.123)$$

Since the scalar products of basis spinors  $\gamma_a$  are anticommuting for both Dirac and Majorana, the coefficients  $\psi^a$  and  $\chi^a$  must be commuting complex numbers for Dirac, but anticommuting complex numbers for Majorana,

$$\psi^a \chi^b = \begin{cases} \chi^b \psi^a & \text{Dirac}, \\ -\chi^b \psi^a & \text{Majorana}. \end{cases} \quad (36.124)$$

The notion of anticommuting complex numbers, called **Grassmann numbers**, is decidedly odd.

In chiral components, the scalar product  $\bar{\psi} \cdot \psi$  is

$$\bar{\psi} \cdot \psi = \psi^{V\downarrow *} \psi^{U\downarrow} + \psi^{U\uparrow *} \psi^{V\uparrow} \pm \psi^{U\downarrow *} \psi^{V\downarrow} \pm \psi^{V\uparrow *} \psi^{U\uparrow}, \quad (36.125)$$

where the upper sign is for Dirac, lower for Majorana.

### 36.9 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.

### 36.9.1 Parity inversion $P$

The parity inversion operation  $P$  reverses all the spatial axes, while keeping the time axis unchanged,

$$P : \boldsymbol{\gamma}_m \rightarrow P\boldsymbol{\gamma}_m P^{-1} = \begin{cases} \boldsymbol{\gamma}_m & m = 0, \\ -\boldsymbol{\gamma}_m & m = 1, 2, 3. \end{cases} \quad (36.126)$$

Parity reversal transforms a Dirac spinor  $\psi$  as

$$P : \psi \rightarrow P\psi. \quad (36.127)$$

In any representation, the transformation (36.126) 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:

$$P = i\boldsymbol{\gamma}_0. \quad (36.128)$$

The scalar factor of  $i$  is inserted, arbitrarily, so that, in the Dirac and chiral representations, the parity operator  $P$  (36.128) is real, Hermitian, and unitary, satisfying  $P^2 = 1$ . Parity reversal commutes with spatial rotations, but not with Lorentz boosts.

Some texts argue that, among other things, parity reversal transforms  $\psi(t, \mathbf{x}) \rightarrow \psi(t, -\mathbf{x})$ , but this is incorrect. In general relativity, Lorentz invariance is a local symmetry: Lorentz transformations rotate the tetrad frame at each point. They do not transform the coordinates. Similarly parity reversal, which is an improper Lorentz transformation, flips the spatial tetrad axes at each point. It does not change the coordinates.

### 36.9.2 Time reversal $T$

The time reversal operation  $T$  reverses the time axis, while keeping all the spatial axes unchanged,

$$T : \boldsymbol{\gamma}_m \rightarrow T\boldsymbol{\gamma}_m T^{-1} = \begin{cases} -\boldsymbol{\gamma}_m & m = 0, \\ \boldsymbol{\gamma}_m & m = 1, 2, 3. \end{cases} \quad (36.129)$$

Time reversal transforms a Dirac spinor  $\psi$  as

$$T : \psi \rightarrow T\psi. \quad (36.130)$$

In any representation, the transformation (36.129) 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  $\boldsymbol{\gamma}_0 I$ , so  $T$  must equal that pseudovector up to a possible scalar normalization:

$$T = i\boldsymbol{\gamma}_0 I. \quad (36.131)$$

The scalar normalization factor of  $i$  is chosen, arbitrarily, so that, in the Dirac and chiral representations,  $T$  is imaginary, Hermitian, and unitary, satisfying  $T^2 = 1$ .

Again, some texts argue that time reversal should transform  $\psi(t, \mathbf{x}) \rightarrow \psi(-t, \mathbf{x})$ , but again this confuses Lorentz transformations with coordinate transformations. Time reversal is an improper Lorentz transformation, not a coordinate transformation.

### 36.9.3 $PT$

The product  $PT$  of the parity and time inversion operators,

$$PT = I , \quad (36.132)$$

reverses all 4 spacetime axes  $\boldsymbol{\gamma}_m$ ,

$$PT : \boldsymbol{\gamma}_m \rightarrow I\boldsymbol{\gamma}_m I^{-1} = -\boldsymbol{\gamma}_m , \quad (36.133)$$

and transforms a Dirac spinor  $\psi$  as

$$PT : \psi \rightarrow I\psi . \quad (36.134)$$

The  $PT$  operation is Lorentz-invariant, since the pseudoscalar  $I$  is Lorentz-invariant. The pseudoscalar is related to the chiral matrix by  $I = i\gamma_5$ , so the spinor basis elements  $\gamma_a$  in the chiral representation are  $PT$ -eigenstates.

## 36.10 Self-conjugate Majorana spinor

Consider the condition that a spinor  $\psi$  be its own  $C$ -conjugate,

$$\bar{\psi} \equiv \psi . \quad (36.135)$$

The condition (36.135) is Lorentz-invariant. A Dirac spinor cannot be self-conjugate, because the Dirac Lagrangian (38.3) vanishes identically for a self-conjugate spinor. A Majorana spinor on the other hand can be self-conjugate, its dynamics being described by the Majorana Lagrangian (38.67).

In the chiral representation, the components  $\psi^a$  of a real self-conjugate Majorana spinor  $\psi$  satisfy, equation (36.101),

$$\psi^{V\uparrow} = \psi^{V\downarrow*} , \quad \psi^{U\uparrow} = -\psi^{U\downarrow*} . \quad (36.136)$$

The 4-component real self-conjugate spinor decomposes into two 2-component real self-conjugate spinors of boost weights  $+\frac{1}{2}$  ( $V$ ) and  $-\frac{1}{2}$  ( $U$ ),

$$\psi = \psi_V + \psi_U , \quad \bar{\psi}_V = \psi_V , \quad \bar{\psi}_U = \psi_U . \quad (36.137)$$

Under a Lorentz boost by rapidity  $\theta$  in the 3-direction, the self-conjugate Majorana spinor (36.137) transforms to

$$\psi = e^{\theta/2}\psi_V + e^{-\theta/2}\psi_U . \quad (36.138)$$

Neither  $\psi_V$  nor  $\psi_U$  has definite chirality, and Lorentz boosting the spinor does not change that fact. Thus a single real self-conjugate Majorana spinor cannot have definite chirality.

A self-conjugate Majorana spinor has the signature property of being its own antiparticle. Indeed, a self-conjugate Majorana spinor is not only its own  $C$ -conjugate, but also its own  $CPT$ -conjugate (spatially rotated), as shown immediately following. Because the property of a particle being its own antiparticle is so distinctive, physicists often use the term Majorana spinor to signify a self-conjugate spinor.

A self-conjugate Majorana spinor  $\psi$  can be represented by a complex quaternion of the form

$$\psi = \begin{Bmatrix} w & x & y & z \\ x & -w & z & -y \end{Bmatrix}, \quad (36.139)$$

as follows from equating the complex quaternionic representations (36.102) and (36.103) of the spinor  $\psi$  and its conjugate  $\bar{\psi}$ . The complex quaternion (36.139) is null, considered as a complex quaternion. However, the quaternionic scalar product is the Dirac scalar product, which is not the same as the Majorana scalar product, so a self-conjugate Majorana spinor can be massive.

The  $CPT$ -conjugate Majorana spinor  $I\psi$  corresponding to the spinor represented by the complex quaternion (36.139) is

$$I\psi = \begin{Bmatrix} -x & w & -z & y \\ w & x & y & z \end{Bmatrix}. \quad (36.140)$$

Rotating the complex quaternion (36.139) spatially by angle  $\pi$  about the 1-direction ( $x$ -axis) multiplies it by  $I$ . Thus the  $PT$ -conjugate  $I\psi$  of a self-conjugate Majorana spinor is itself (spatially rotated). Not only is a self-conjugate Majorana spinor its own  $C$ -conjugate, it is its own  $CPT$ -conjugate.

### 36.10.1 Complex self-conjugate Majorana spinor

Whereas a Dirac spinor has 8 degrees of freedom, a self-conjugate Majorana spinor has 4 degrees of freedom. A Majorana spinor can also be complex. A complex 8-component Majorana spinor can be formed from a complex linear combination of two real 4-component Majorana spinors satisfying the self-conjugate conditions  $\overline{\psi_R} = \psi_R$  (Real part) and  $\overline{\psi_I} = \psi_I$  (Imaginary part),

$$\psi = \psi_R + i\psi_I, \quad (36.141)$$

whose  $C$ -conjugate is

$$\bar{\psi} = \psi_R - i\psi_I. \quad (36.142)$$

Such a complex 8-component Majorana spinor  $\psi$  can be decomposed into a sum of two self-conjugate 4-component Majorana spinors,

$$\psi_R \equiv \frac{1}{2}(\psi + \bar{\psi}), \quad \psi_I \equiv \frac{1}{2i}(\psi - \bar{\psi}). \quad (36.143)$$

The decomposition (36.143) of a complex Majorana spinor is not unique: the spinor and its conjugate can be rotated by a phase,  $\psi \rightarrow e^{-i\theta}\psi$  and  $\bar{\psi} \rightarrow e^{i\theta}\bar{\psi}$ , before decomposing as (36.143). A complex 8-component

Majorana spinor differs from two real 4-component Majorana spinors by having an additional symmetry under rotation by a phase. Despite having the same number of components, a complex Majorana spinor is not the same as a Dirac spinor, because it satisfies a different scalar product, and has a different Lagrangian, §38.6.

In contrast to a single-component self-conjugate Majorana spinor, a complex Majorana spinor that is a complex linear combination of self-conjugate Majorana spinors can have definite chirality.

### 36.11 Real, or Majorana, representation of $\gamma$ -matrices

A real, or Majorana, representation is constructed so that the Dirac  $\gamma$ -matrices are real, and a Majorana spinor can be represented with 4 real components. The choice of representation makes no difference whatsoever to the physics. A Majorana spinor can be represented in a Dirac or chiral representation, and a Dirac spinor can be represented in a Majorana representation. The difference between a Dirac and Majorana spinor is not in their representation, but rather in the scalar product that they satisfy, Dirac or Majorana, §36.5.1.

Whereas in the Dirac representation the time matrix  $\gamma_0$  is diagonal, and in the chiral representation the chiral matrix  $\gamma_5$  is diagonal, in the Majorana representation the conjugation matrix  $C$  is diagonal.

A suitable real Majorana representation may be obtained from the chiral representation (36.15) by the transformation

$$X : \boldsymbol{\gamma}_m \rightarrow X\boldsymbol{\gamma}_m X^{-1} , \quad (36.144)$$

where  $X$  is the unitary ( $X^{-1} = X^\dagger$ ) matrix

$$X \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ -i & 0 & 0 & i \\ 0 & 1 & -1 & 0 \\ 0 & -i & -i & 0 \end{pmatrix} . \quad (36.145)$$

The matrix  $X$  (36.145) is chosen so that the components of a self-conjugate Majorana spinor  $\psi$  comprise the real and imaginary parts of its right-handed components in the chiral representation,

$$\psi^a = \begin{pmatrix} \text{Re } \psi^{V\uparrow} \\ \text{Im } \psi^{V\uparrow} \\ \text{Re } \psi^{U\downarrow} \\ \text{Im } \psi^{U\downarrow} \end{pmatrix} = \begin{pmatrix} \text{Re } \psi^{V\downarrow} \\ -\text{Im } \psi^{V\downarrow} \\ -\text{Re } \psi^{U\uparrow} \\ \text{Im } \psi^{U\uparrow} \end{pmatrix} . \quad (36.146)$$

However, the mere use of a Majorana representation does not force the spinor  $\psi$  to be real. A complex Majorana spinor (36.142) is complex. A Dirac spinor in a Majorana representation is necessarily complex.

The mapping  $\gamma_a \rightarrow X\gamma_a$  transforms the spinor basis elements in the chiral representation to those in the

Majorana representation, yielding

$$\gamma_{VR} = \frac{1}{\sqrt{2}}(\gamma_{V\uparrow} + \gamma_{V\downarrow}) , \quad (36.147a)$$

$$\gamma_{VI} = \frac{1}{\sqrt{2}i}(\gamma_{V\uparrow} - \gamma_{V\downarrow}) , \quad (36.147b)$$

$$\gamma_{UR} = \frac{1}{\sqrt{2}}(\gamma_{U\uparrow} - \gamma_{U\downarrow}) , \quad (36.147c)$$

$$\gamma_{UI} = \frac{1}{\sqrt{2}i}(\gamma_{U\uparrow} + \gamma_{U\downarrow}) . \quad (36.147d)$$

The indices  $R$  and  $I$  (“Real” and “Imaginary”) are motivated by equation (36.146). The inverse of equations (36.147) is

$$\gamma_{V\uparrow} = \frac{1}{\sqrt{2}}(\gamma_{VR} + i\gamma_{VI}) , \quad (36.148a)$$

$$\gamma_{U\downarrow} = \frac{1}{\sqrt{2}}(\gamma_{UI} + i\gamma_{UR}) , \quad (36.148b)$$

$$\gamma_{U\uparrow} = -\frac{1}{\sqrt{2}}(\gamma_{UI} - i\gamma_{UR}) , \quad (36.148c)$$

$$\gamma_{V\downarrow} = \frac{1}{\sqrt{2}}(\gamma_{VR} - i\gamma_{VI}) . \quad (36.148d)$$

The  $\gamma$ -matrices in the Majorana representation are the real orthogonal matrices

$$\boldsymbol{\gamma}_0 = \begin{pmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{pmatrix} , \quad \boldsymbol{\gamma}_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & -\sigma_1 \end{pmatrix} , \quad \boldsymbol{\gamma}_2 = \begin{pmatrix} \sigma_3 & 0 \\ 0 & \sigma_3 \end{pmatrix} , \quad \boldsymbol{\gamma}_3 = \begin{pmatrix} 0 & -\sigma_1 \\ -\sigma_1 & 0 \end{pmatrix} . \quad (36.149)$$

The bivectors  $\boldsymbol{\gamma}_0 \boldsymbol{\gamma}_a \equiv \boldsymbol{\sigma}_a$  and  $\varepsilon_{abc} \boldsymbol{\gamma}_b \boldsymbol{\gamma}_c \equiv I \boldsymbol{\sigma}_a$  and the pseudoscalar  $I$  are the real orthogonal matrices

$$\begin{aligned} \boldsymbol{\sigma}_1 &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} , & \boldsymbol{\sigma}_2 &= \begin{pmatrix} 0 & i\sigma_2 \\ -i\sigma_2 & 0 \end{pmatrix} , & \boldsymbol{\sigma}_3 &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} , \\ I\boldsymbol{\sigma}_1 &= \begin{pmatrix} 0 & -i\sigma_2 \\ -i\sigma_2 & 0 \end{pmatrix} , & I\boldsymbol{\sigma}_2 &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} , & I\boldsymbol{\sigma}_3 &= \begin{pmatrix} -i\sigma_2 & 0 \\ 0 & i\sigma_2 \end{pmatrix} , \\ I &= \begin{pmatrix} -i\sigma_2 & 0 \\ 0 & -i\sigma_2 \end{pmatrix} . \end{aligned} \quad (36.150)$$

In the Majorana representation, the basis spinors  $\gamma_a$  have definite boost weights, but not definite spin weights, as is apparent from equations (36.146), and evident from the fact that  $\boldsymbol{\sigma}_3$  is diagonal but  $I\boldsymbol{\sigma}_3$  is not. The chiral matrix  $\gamma_5$  is the pure imaginary Hermitian matrix

$$\gamma_5 \equiv -iI = \begin{pmatrix} -\sigma_2 & 0 \\ 0 & -\sigma_2 \end{pmatrix} . \quad (36.151)$$

The chiral matrix is not diagonal, so the basis spinors in the Majorana representation do not have definite chirality.

Although the basis spinors in the Majorana representation do not have definite chirality, complex linear combinations of the basis spinors can have definite chirality, as illustrated by equations (36.148).

### 36.11.1 Spinor metric and conjugation operator in the Majorana representation

A key property of the conjugation operator  $C$  is that it flips chirality. The conjugation operator is not redefined in the Majorana representation to change the  $\gamma$ -matrices to their complex conjugates (which would mean that  $C$  would be the unit matrix, and do nothing). Rather, the properties of the conjugation operator are carried over unchanged to the Majorana representation. Thus the conjugation matrix in the Majorana representation is still  $C = \boldsymbol{\gamma}_2$ , as it is in the Dirac and chiral representations, equation (36.90),

$$C = \boldsymbol{\gamma}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (36.152)$$

The conjugation matrix  $C$  leaves the “real” basis spinors  $\gamma_{VR}$  and  $\gamma_{UR}$  unchanged, while flipping the sign of the “imaginary” basis spinors  $\gamma_{VI}$  and  $\gamma_{UI}$ . This has the effect of flipping chirality, the chiral basis spinors being related to the Majorana basis spinors by equations (36.147).

Likewise, the properties of the spinor metric tensor  $\varepsilon$  are carried over from the Dirac and chiral representations. The spinor metric tensor  $\varepsilon$  in the Majorana representation equals  $i\sigma_2$  for Dirac spinors, and  $I\sigma_2$  for Majorana spinors, as in the Dirac and chiral representations, equation (36.38),

$$\varepsilon = \begin{cases} i\sigma_2 = \begin{pmatrix} 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix} & \text{Dirac ,} \\ I\sigma_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} & \text{Majorana .} \end{cases} \quad (36.153)$$

### 36.11.2 Chiral decomposition of a Majorana spinor

Like a Dirac spinor, a Majorana spinor  $\psi$  decomposes into right- and left-handed chiral components that are the chiral projections of  $\psi$ , equation (36.34),

$$\psi = \psi_R + \psi_L. \quad (36.154)$$

Merely decomposing a Majorana spinor into right- and left-handed components does not mean that either chiral component is an eigenstate. A single real self-conjugate spinor cannot be a chiral eigenstate. However, a complex superposition of self-conjugate Majorana spinors can be a chiral eigenstate.

Since  $C$ -conjugation flips chirality regardless of representation, the right-handed component of a self-conjugate Majorana spinor must be the conjugate of the left-handed component (and vice-versa),

$$\psi_R = \overline{\psi_L}. \quad (36.155)$$

The right- and left-handed chiral components of a self-conjugate Majorana spinor can be represented as complex quaternions

$$\psi_R = \begin{Bmatrix} a & b & c & d \\ -d & -c & b & a \end{Bmatrix}, \quad \psi_L = \begin{Bmatrix} c & -d & -a & b \\ b & -a & d & -c \end{Bmatrix}, \quad (36.156)$$

whose coefficients  $a, b, c, d$  are related to those  $w, x, y, z$  of the Majorana self-conjugate complex quaternion (36.139) by

$$a = \frac{1}{2}(w - y), \quad b = \frac{1}{2}(x + z), \quad c = \frac{1}{2}(y + w), \quad d = \frac{1}{2}(z - x). \quad (36.157)$$

# 37

---

## Geometric Differentiation and Integration of Spinors

### 37.1 Covariant derivative of a Dirac spinor

In the case of a Dirac spinor  $\psi$ , a Lorentz transformation by rotor  $R$  transforms the spinor by  $\psi \rightarrow R\psi$ . An infinitesimal Lorentz transformation  $R = 1 + \epsilon\boldsymbol{\Gamma}/2$  generated by a bivector  $\boldsymbol{\Gamma}$  transforms  $\psi \rightarrow \psi + \frac{1}{2}\epsilon\boldsymbol{\Gamma}\psi$ . Consequently the action of the connection operator  $\hat{\Gamma}_n$  on a Dirac spinor  $\psi$  is

$$\hat{\Gamma}_n\psi = \frac{1}{2}\boldsymbol{\Gamma}_n\psi , \quad (37.1)$$

where  $\boldsymbol{\Gamma}_n$  is the  $N$ -tuple of bivectors (15.9). The covariant derivative of a Dirac spinor  $\psi$  is thus

$$D_n\psi = \partial_n\psi + \frac{1}{2}\boldsymbol{\Gamma}_n\psi , \quad (37.2)$$

In equation (37.3), as previously in equations (15.6) and (15.15), for a spinor  $\psi = \gamma_a\psi^a$ , the directed derivative  $\partial_n$  is to be interpreted as acting only on the components  $\psi^a$  of the spinor,  $\partial_n\psi = \gamma_a\partial_n\psi^a$ . In the convention (36.79) that multivectors acting to the right of a column spinor yield zero, the connection term in equation (37.3) can be written as a commutator, in the same form as (15.15),

$$D_n\psi = \partial_n\psi + \frac{1}{2}[\boldsymbol{\Gamma}_n, \psi] . \quad (37.3)$$

### 37.2 Covariant derivative in a spinor basis

The covariant derivative  $D_n$  can also be expressed in a spinor basis.

The spinor tetrad connections  $\Gamma_{an}^b$  are defined, analogously to the definition (11.37) of the tetrad connections  $\Gamma_{mn}^k$ , to be the coefficients of the change of the spinor axes  $\gamma_a$  parallel-transported along the direction  $\boldsymbol{\gamma}_n$ ,

$$\Gamma_{an}^b \gamma_b \equiv \partial_n \gamma_a . \quad (37.4)$$

The same equation (37.4) with a trailing dot appended on both sides holds for row spinors. The constancy

and antisymmetry of the spinor metric,

$$0 = \partial_n \varepsilon_{ab} = \partial_n (\gamma_a \cdot \gamma_b) = \Gamma_{bn}^c \gamma_a \cdot \gamma_c + \Gamma_{an}^c \gamma_c \cdot \gamma_b = \Gamma_{bn}^c \varepsilon_{ac} + \Gamma_{an}^c \varepsilon_{cb} = \Gamma_{abn} - \Gamma_{ban} , \quad (37.5)$$

implies that the spinor tetrad connection  $\Gamma_{abn}$  is symmetric in its first two indices,

$$\Gamma_{abn} = \Gamma_{ban} . \quad (37.6)$$

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  $\Gamma_{abn}$  are non-vanishing only when  $a$  and  $b$  have the same chirality,

$$\Gamma_{abn} = 0 \quad \text{for } a, b \text{ of opposite chirality} . \quad (37.7)$$

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  $\Gamma_{abn}$  are related to the 24 tetrad connection coefficients  $\Gamma_{kmn}$  by

$$\Gamma_{abn} = \gamma_{ab}^{km} \Gamma_{kmn} , \quad (37.8)$$

where  $\gamma_{ab}^{km}$  is the matrix defined by equation (36.73) 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 (36.64) and (36.65).

The connection  $\Gamma_n$  defined by equation (15.9) is terms of the spinor basis

$$\Gamma_n = \Gamma_{abn} \gamma^a \gamma^b , \quad (37.9)$$

implicitly summed over distinct symmetric self-chiral pairs  $ab$  of spinor indices. Expressions (15.15) and (37.3) for the covariant derivatives of multivectors and spinors remain valid with the connection  $\Gamma_n$  given by equation (37.9).

### 37.3 Ashtekhar variables

WHAT TO DO WITH THIS? Left-handed  $\Gamma_{ab\mu}$  is an Ashtekhar variable.

$$\mathbf{R}_{\kappa\lambda}^R = \frac{\partial \mathbf{\Gamma}_{\lambda}^R}{\partial x^{\kappa}} - \frac{\partial \mathbf{\Gamma}_{\kappa}^R}{\partial x^{\lambda}} + \frac{1}{2} [\mathbf{\Gamma}_{\kappa}^R, \mathbf{\Gamma}_{\lambda}^R] , \quad (37.10)$$

### 37.4 Covariant spacetime derivative of a spinor

Acting on a Dirac (respectively Majorana) column spinor  $\psi$ , the covariant spacetime derivative yields another Dirac (respectively Majorana) spinor

$$\mathbf{D}\psi \quad \text{column spinor} . \quad (37.11)$$

This derivative is a fundamental ingredient in the Lagrangian for a Dirac or Majorana field, and in the resulting Dirac or Majorana 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  $\psi$ , 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 (Dirac) or plus (Majorana) the spacetime derivative acting on the row spinor to the left:

$$\mathbf{D}\psi \cdot \equiv (\mathbf{D}\psi) \cdot = (\mathbf{D}\psi)^T \varepsilon = \psi^T \overset{\leftarrow}{\mathbf{D}}^T \varepsilon = \begin{cases} -\psi^T \varepsilon \overset{\leftarrow}{\mathbf{D}} = -\psi \cdot \overset{\leftarrow}{\mathbf{D}} & \text{Dirac} \\ \psi^T \varepsilon \overset{\leftarrow}{\mathbf{D}} = \psi \cdot \overset{\leftarrow}{\mathbf{D}} & \text{Majorana} \end{cases} \quad \text{row spinor .} \quad (37.12)$$

The penultimate step is true because  $\boldsymbol{\gamma}^{n\top} \varepsilon = \mp \varepsilon \boldsymbol{\gamma}^n$  for Dirac and Majorana spinors respectively, equation (36.37).

### 37.5 Covariant spacetime derivative of a true scalar

The super spacetime algebra contains not only scalars that are multivectors, proportional to the unit matrix in any representation, but also true scalars that are just complex numbers, not matrices. These two kinds of scalar must be distinguished. For example, in least action, the action is a true scalar. By contrast, a so-called scalar field is a multivector scalar. Whereas the covariant spacetime derivative of a multivector scalar yields a vector, equation (15.39a), the covariant spacetime derivative of a true scalar is defined so that it yields another true scalar.

The covariant spacetime derivative of the scalar product  $\psi \cdot \chi$  of a row spinor  $\psi \cdot$  with a column spinor  $\chi$ , which is a true scalar, is defined to satisfy the Leibniz rule:

$$\mathbf{D}(\psi \cdot \chi) = \psi \cdot (\mathbf{D}\chi) + (\mathbf{D}\psi) \cdot \chi = \begin{cases} \psi \cdot (\mathbf{D}\chi) - \chi \cdot (\mathbf{D}\psi) & \text{Dirac} \\ \psi \cdot (\mathbf{D}\chi) + \chi \cdot (\mathbf{D}\psi) & \text{Majorana} \end{cases} \quad \text{true scalar ,} \quad (37.13)$$

the right hand side of which is again a true scalar, a sum of scalar products of row spinors with column spinors. Equation (37.13) respects the antisymmetry (Dirac) or symmetry (Majorana) of the spinor scalar product, as it should.

Beware that a scalar product of multivectors is a multivector scalar, not a true scalar: equation (37.13) applies to spinor scalar products, not to multivector scalar products.

### 37.6 Stokes' and Gauss' theorems for spinors

Let  $\psi$  be a Dirac or Majorana column spinor. The analogue of Stokes' theorem (15.88) is

$$\int_V (p+1) \boldsymbol{\gamma}_{[m_1} \dots \boldsymbol{\gamma}_{m_p]} D_{n]} \psi \, d^{p+1} x^{m_1 \dots m_p n} = \oint_{\partial V} \boldsymbol{\gamma}_{[m_1} \dots \boldsymbol{\gamma}_{m_p]} \psi \, d^p x^{m_1 \dots m_p} , \quad (37.14)$$

with implicit summation over distinct antisymmetric sequences of respectively  $p+1$  and  $p$  indices on the left and right. Since the basis vectors  $\gamma_m$  are covariantly constant, they could equally well be moved in front of the derivative in the integrand on the left hand side (but still before  $\psi$ , since multivectors can only pre-multiply a column spinor). The simplest case,  $p = 0$ , is the fundamental theorem of calculus,

$$\int_{x_0}^{x_1} D_n \psi \, dx^n = \psi(x_1) - \psi(x_0) . \quad (37.15)$$

The analogue of Gauss' theorem (15.96) is obtained by pre-multiplying the integrand on the left hand side of Stokes' theorem (37.14) by the pseudoscalar  $I_N$ ,

$$\int_V \gamma_{[m_1} \dots \gamma_{m_{p-1}} \gamma_{n]} D^n \psi \, {}^* d^{q+1} x^{m_1 \dots m_{p-1}} = \oint_{\partial V} \gamma_{[m_1} \dots \gamma_{m_p]} \psi \, {}^* d^q x^{m_1 \dots m_p} , \quad (37.16)$$

with  $q \equiv N - p$ , and implicit summation over distinct antisymmetric sequences of respectively  $p-1$  and  $p$  indices on left and right. The simplest and most common application of Gauss' theorem (37.16) for spinors  $\psi$  is  $p = 1$ ,

$$\int \mathbf{D} \psi \, d^N x = \oint \gamma_n \psi \, d^{N-1} x^n , \quad (37.17)$$

where  $d^N x$  and  $d^{N-1} x^n$  denote respectively the dual scalar  $N$ -volume and the dual vector  $N-1$ -volume.

The proof of Stokes' theorem (37.14) for spinors is similar to that for multivectors. The key point is that the integrands on both sides are (coordinate and tetrad) gauge-invariant scalars, so one is free to choose whatever coordinate system is most convenient. Break the volume  $V$  of integration into patches, contrive rectangular coordinates on each patch, do the easy Euclidean integral on each patch, and add the patches.

### 37.7 Gauss' theorem for true scalars

In practical application to Lagrangians, Gauss' theorem for true scalars occurs in the form

$$\int (\psi \cdot \mathbf{D} \chi + \chi \cdot \mathbf{D} \psi) \, d^N x = \int D^n (\psi \cdot \gamma_n \chi) \, d^N x = \oint \psi \cdot \gamma_n \chi \, d^{N-1} x^n , \quad (37.18)$$

where  $\psi$  and  $\chi$  are spinors. Equation (37.18) is true for both Dirac and Majorana spinors. For Dirac spinors, a sign flip from the antisymmetry of the scalar product is cancelled by a sign flip from anticommutation of the basis vectors through the spinor scalar product,  $\chi \cdot (\mathbf{D} \psi) = -(\mathbf{D} \psi) \cdot \chi = (D^n \psi) \cdot \gamma^n \chi$ . The integrand on the left hand side of equation (37.18) is the same as the derivative  $\mathbf{D}(\psi \cdot \chi)$  defined by equation (37.13) for Majorana, but not for Dirac,

$$\psi \cdot \mathbf{D} \chi + \chi \cdot \mathbf{D} \psi = \begin{cases} \psi \cdot (\mathbf{D} \chi) - (\mathbf{D} \psi) \cdot \chi \neq \mathbf{D}(\psi \cdot \chi) & \text{Dirac ,} \\ \psi \cdot (\mathbf{D} \chi) + (\mathbf{D} \psi) \cdot \chi = \mathbf{D}(\psi \cdot \chi) & \text{Majorana .} \end{cases} \quad (37.19)$$

Whereas for Majorana the total derivative  $\mathbf{D}(\psi \cdot \chi)$  of a scalar product of spinors integrates to a surface term, for Dirac it does not.

# 38

---

## Action principle for fields

THIS CHAPTER NEEDS WORK, ESPECIALLY LATTER PARTS.

Fields are at the heart of modern physics. Quantum field theory starts with solutions of classical fields. The properties and dynamics of fields are most transparent when derived from a Lagrangian or Hamiltonian.

Lagrangians are known for fields of spin 0,  $\frac{1}{2}$ , 1,  $\frac{3}{2}$ , and 2. It is thought that there are no consistent quantum field theories for fundamental fields of spin other than these (BUT WEINBERG SAYS OTHERWISE).

In this chapter,  $D_n$  and  $\mathbf{D}$  denote the torsion-free covariant derivative (equivalent to the exterior derivative  $d$  in a differential forms approach). The torsion-free covariant derivative is prescribed because this is the only derivative to which Stokes' theorem applies, and which therefore allows integration by parts. Torsion need not vanish, but it cannot contribute to the Lagrangians considered. REALLY?

As expounded in Chapter 15,  $d^4x$  denotes the invariant scalar 4-volume, equation (15.97), not the pseudoscalar 4-volume. The units are  $c = \hbar = 1$ .

### 38.1 Dirac spinor field

#### 38.1.1 Dirac Lagrangian

The general relativistic scalar Lagrangian  $L$  of a free Dirac spinor field  $\psi$  of mass  $m$  is

$$L = \bar{\psi} \cdot (\mathbf{D} + m) \psi . \quad (38.1)$$

Here  $\mathbf{D} \equiv \gamma^n D_n$  is the torsion-free covariant spacetime derivative, equation (15.31), and  $\bar{\psi}$  is the  $C$ -conjugate field defined by equation (36.95). In flat (Minkowski) space the justification for the Dirac Lagrangian (38.1) is that it leads to equations that reproduce ample experiment. Equation (38.1) 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  $\psi$  has units of length $^{-3/2}$ .

To apply least action, the Lagrangian must be expressed as a function of fields and their derivatives. The Dirac spinor field  $\psi$  contains 4 complex (with respect to  $i$ , in a complex representation) components, 8 components altogether. In varying the Lagrangian, it is convenient to treat the field  $\psi$  and its conjugate

$\bar{\psi}$  as two independent fields containing 4 components each. The derivative  $\mathbf{D}\psi$  of the spinor field is itself a spinor, and invites identification as the velocity of the field.

As it stands, the Lagrangian (38.1) 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 (38.1) is complex, not real. Despite these flaws, the complex Lagrangian (38.1) does yield the correct Dirac equations, because, as shown below, equation (38.4), the imaginary part of the Lagrangian integrates to a surface term, and therefore has no effect on the equations of motion. However, a symmetric treatment is followed here, making manifest the symmetry between the field  $\psi$  and its conjugate  $\bar{\psi}$ . Symmetry and reality can be restored by symmetrizing the Lagrangian (38.1) 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  $\boldsymbol{\gamma}_n$ . The complex conjugate of any multivector  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  with real coefficients  $a^A$  is  $\mathbf{a}^* = a^A \boldsymbol{\gamma}_A^* = a^A C \boldsymbol{\gamma}_A C = C \mathbf{a} C$ . Consequently for any multivector  $\mathbf{a}$  whose coefficients are real in an orthonormal basis,

$$(\bar{\psi} \cdot \mathbf{a} \psi)^* = \psi^\top C \varepsilon \mathbf{a}^* C \bar{\psi} = -\psi^\top \varepsilon C \mathbf{a}^* C \bar{\psi} = -\psi \cdot \mathbf{a} \bar{\psi}, \quad (38.2)$$

where the sign flip in the third expression occurs because the conjugation operator  $C$  anticommutes with the Dirac spinor metric tensor  $\varepsilon$ , equation (36.93). The symmetrized, real Lagrangian is thus

$$L = \frac{1}{2} \bar{\psi} \cdot (\mathbf{D} + m) \psi - \frac{1}{2} \psi \cdot (\mathbf{D} + m) \bar{\psi}. \quad (38.3)$$

The antisymmetry of the Dirac scalar product, equation (36.124), allows a non-vanishing mass  $m$ .

Gauss' theorem (37.18) for true scalars shows that the imaginary part of  $\bar{\psi} \cdot \mathbf{D}\psi$  integrates to a surface term,

$$i \int \text{Im}(\bar{\psi} \cdot \mathbf{D}\psi) d^4x = \frac{1}{2} \int (\bar{\psi} \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}\bar{\psi}) d^4x = \frac{1}{2} \oint \bar{\psi} \cdot \boldsymbol{\gamma}_n \psi d^3x^n, \quad (38.4)$$

and therefore has no effect on the equations of motion.

Canonically conjugate momenta are obtained by differentiating the Lagrangian  $L(\psi, \bar{\psi}, \mathbf{D}\psi, \mathbf{D}\bar{\psi})$  with respect to the velocity spinors  $\mathbf{D}\psi$  and  $\mathbf{D}\bar{\psi}$ . The derivative of a scalar product of spinors with respect to a column or row spinor is defined in the natural way, satisfying the Leibniz rule, by

$$\frac{\partial \chi \cdot \psi}{\partial \psi} \equiv \frac{\partial \chi \cdot \psi}{\partial \psi_a} \boldsymbol{\gamma}_a \cdot = \chi \cdot, \quad \frac{\partial \chi \cdot \psi}{\partial \chi} = -\frac{\partial \psi \cdot \chi}{\partial \chi} = -\psi \cdot, \quad \frac{\partial \chi \cdot \psi}{\partial \psi \cdot} = -\frac{\partial \psi \cdot \chi}{\partial \psi \cdot} = -\chi, \quad \frac{\partial \chi \cdot \psi}{\partial \chi \cdot} = \psi \cdot. \quad (38.5)$$

The trailing dot on a spinor identifies it as a row spinor, which, as defined by equation (36.105), equals the transpose of the column spinor, post-multiplied by the spinor metric tensor. The trailing dot is symbolic of the spinor metric. The antisymmetry of the Dirac spinor scalar product introduces a minus sign when a row spinor is differentiated with respect to a column spinor, or a column spinor is differentiated with respect to a row spinor.

The momenta  $\frac{1}{2}\pi$  and  $-\frac{1}{2}\bar{\pi}$  canonically conjugate to the field  $\psi$  and its conjugate  $\bar{\psi}$  are given by

$$\frac{1}{2}\pi \cdot = \frac{\partial L}{\partial(\mathbf{D}\psi)} = \frac{1}{2}\bar{\psi} \cdot, \quad -\frac{1}{2}\bar{\pi} \cdot = \frac{\partial L}{\partial(\mathbf{D}\bar{\psi})} = -\frac{1}{2}\psi \cdot, \quad (38.6)$$

implying

$$\pi = \bar{\psi} , \quad -\bar{\pi} = -\psi . \quad (38.7)$$

The factors of  $\frac{1}{2}$  on  $\pi$  and  $\bar{\pi}$  in equation (38.72) are introduced to ensure that  $\bar{\pi}$  is the  $C$ -conjugate of  $\pi$ , as defined by equation (36.95). The momentum conjugate to  $\bar{\psi}$  is  $-\frac{1}{2}\bar{\pi}$  with a minus sign. The momenta canonically conjugate to the fields are, modulo the factors of  $\frac{1}{2}$  and the minus on  $-\frac{1}{2}\bar{\pi}$ , their  $C$ -conjugates.

Linearity of the derivative

$$\mathbf{D}(\psi + \delta\psi) = \mathbf{D}\psi + \mathbf{D}(\delta\psi) , \quad (38.8)$$

shows that the variation of the velocity equals the velocity of the variation,  $\delta(\mathbf{D}\psi) = \mathbf{D}(\delta\psi)$ . The variation  $\delta L_{\mathbf{D}\psi}$  of the Lagrangian with respect to the velocity  $\mathbf{D}\psi$  (and similarly  $\mathbf{D}\bar{\psi}$ ) leads to terms that warrant integration by parts,

$$\delta L_{\mathbf{D}\psi} = \frac{1}{2}\pi \cdot \mathbf{D}(\delta\psi) = \frac{1}{2}D^n(\pi \cdot \boldsymbol{\gamma}_n \delta\psi) - \frac{1}{2}\pi \cdot \overleftarrow{\mathbf{D}} \delta\psi = \frac{1}{2}D^n(\pi \cdot \boldsymbol{\gamma}_n \delta\psi) + \frac{1}{2}\mathbf{D}\pi \cdot \delta\psi . \quad (38.9)$$

The sign flip in the last expression is from equation (37.12). Varying the action with respect to the fields and their velocities thus gives

$$\delta S = \oint \frac{1}{2} (\pi \cdot \boldsymbol{\gamma}_n \delta\psi - \bar{\pi} \cdot \boldsymbol{\gamma}_n \delta\bar{\psi}) d^3x^n + \int \left[ \left( \frac{\partial L}{\partial \psi} + \frac{1}{2}\mathbf{D}\pi \cdot \right) \delta\psi + \left( \frac{\partial L}{\partial \bar{\psi}} - \frac{1}{2}\mathbf{D}\bar{\pi} \cdot \right) \delta\bar{\psi} \right] d^4x = 0 . \quad (38.10)$$

The resulting Euler-Lagrange equations of motion are

$$\frac{1}{2}\mathbf{D}\pi \cdot = -\frac{\partial L}{\partial \psi} = -\left(\frac{1}{2}\mathbf{D} + m\right)\bar{\psi} \cdot , \quad \frac{1}{2}\mathbf{D}\bar{\pi} \cdot = \frac{\partial L}{\partial \bar{\psi}} = -\left(\frac{1}{2}\mathbf{D} + m\right)\psi \cdot . \quad (38.11)$$

Substituting the momenta (38.7) into the Euler-Lagrange equations (38.11) and switching from row to column spinors (by dropping the trailing dots) yields the Dirac equations

$$(\mathbf{D} + m)\psi = 0 , \quad (38.12a)$$

$$(\mathbf{D} + m)\bar{\psi} = 0 . \quad (38.12b)$$

The free Dirac field  $\psi$  and its conjugate  $\bar{\psi}$  satisfy the same equation of motion.

### 38.1.2 Conserved Dirac number current

The Dirac Lagrangian (38.3) is unchanged if the field and its conjugate are changed by opposing complex phases,  $\psi \rightarrow e^{-i\epsilon}\psi$  and  $\bar{\psi} \rightarrow e^{i\epsilon}\bar{\psi}$ . In infinitesimal form, this transformation is

$$\psi \rightarrow \psi - i\epsilon\psi , \quad \bar{\psi} \rightarrow \bar{\psi} + i\epsilon\bar{\psi} . \quad (38.13)$$

The corresponding conserved Noether current, equation (16.17), is

$$n^m = \frac{1}{2}i(\pi \cdot \boldsymbol{\gamma}^m \psi + \bar{\pi} \cdot \boldsymbol{\gamma}^m \bar{\psi}) = \frac{1}{2}i(\bar{\psi} \cdot \boldsymbol{\gamma}^m \psi + \psi \cdot \boldsymbol{\gamma}^m \bar{\psi}) . \quad (38.14)$$

The relative sign between the two terms in the middle expression of equations (38.14) is positive because the fields vary with opposite sign under the transformation (38.13),  $\delta\bar{\psi} = -\delta\psi$ , and in addition the momenta

canonically conjugate to  $\psi$  and  $\bar{\psi}$  are  $\frac{1}{2}\pi$  and  $-\frac{1}{2}\bar{\pi}$ , equations (38.72). The last expression of equations (38.14) follows from substituting the expressions (38.7) for  $\pi$  and  $\bar{\pi}$ . Since the two terms in the last expression are the same,

$$\psi \cdot \boldsymbol{\gamma}^m \bar{\psi} = -(\boldsymbol{\gamma}^m \bar{\psi}) \cdot \psi \equiv -(\boldsymbol{\gamma}^m \bar{\psi})^\top \boldsymbol{\varepsilon} \psi = -\bar{\psi}^\top \boldsymbol{\gamma}^{m\top} \boldsymbol{\varepsilon} \psi = \bar{\psi}^\top \boldsymbol{\varepsilon} \boldsymbol{\gamma}^m \psi = \bar{\psi} \cdot \boldsymbol{\gamma}^m \psi , \quad (38.15)$$

equation (38.14) simplifies to

$$n^m = i \bar{\psi} \cdot \boldsymbol{\gamma}^m \psi . \quad (38.16)$$

The Dirac current (38.16) is covariantly conserved in accordance with Noether's theorem, equation (16.18),

$$D_m n^m = 0 . \quad (38.17)$$

The factor  $i$  in the Dirac current (38.16) is introduced so that the time component  $n^0$  is a positive number,

$$n^0 = \psi^\dagger \psi , \quad (38.18)$$

where, in accordance with equation (36.105),  $\psi^\dagger \equiv \psi^{*\top}$  is the Hermitian conjugate of  $\psi$ . The Dirac current  $n^m$  is interpreted as a conserved probability number current with a positive density  $n^0$ .

If the current (38.16) is written

$$\mathbf{n} = \boldsymbol{\gamma}_m n^m = i \boldsymbol{\gamma}_m \bar{\psi} \cdot \boldsymbol{\gamma}^m \psi , \quad (38.19)$$

then the probability conservation equation (38.17) is

$$\mathbf{D} \cdot \mathbf{n} = 0 . \quad (38.20)$$

If the Dirac spinor is null,  $\bar{\psi} \cdot \psi = 0$ , 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

$$n_5^m \equiv i \bar{\psi} \cdot \gamma_5 \boldsymbol{\gamma}^m \psi \quad (38.21)$$

is also conserved.

## 38.2 Dirac field with electromagnetism

Electromagnetism emerges from elevating the invariance of the Lagrangian under rotation of the Dirac fields  $\psi$  and  $\bar{\psi}$  by a complex phase from a global to a local symmetry. This kind of transformation is called a **gauge transformation**. The three forces of the standard model, the electromagnetic, weak, and strong forces, all emerge from gauge transformations. Electromagnetism is the simplest gauge field, based on the 1-dimensional unitary group  $U(1)$  of rotations about a circle.

Under an electromagnetic gauge transformation, a Dirac field  $\psi$  of charge  $e$ , and its conjugate field  $\bar{\psi}$ , which is proportional to the complex conjugate of the field, equation (36.95), transform as

$$\psi \rightarrow e^{-ie\theta} \psi , \quad \bar{\psi} \rightarrow e^{ie\theta} \bar{\psi} , \quad (38.22)$$

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 (38.3) remains invariant under the gauge transformation, the derivative  $\mathbf{D}$  must be replaced by a gauge-covariant derivative  $\mathbf{D} \pm ie\mathbf{A}$  which, when acting on the field and its conjugate, transforms under the gauge transformation (38.22) as

$$(\mathbf{D} + ie\mathbf{A})\psi \rightarrow e^{-ie\theta}(\mathbf{D} + ie\mathbf{A})\psi, \quad (\mathbf{D} - ie\mathbf{A})\bar{\psi} \rightarrow e^{ie\theta}(\mathbf{D} - ie\mathbf{A})\bar{\psi}. \quad (38.23)$$

The gauge-covariant derivative transforms correctly provided that the gauge field  $\mathbf{A}$  transforms under the electromagnetic gauge transformation (38.22) as

$$\mathbf{A} \rightarrow \mathbf{A} + \mathbf{D}\theta. \quad (38.24)$$

The gauge field  $\mathbf{A}$  is the electromagnetic potential.

The general relativistic scalar Lagrangian  $L$  of a Dirac spinor field  $\psi$  of mass  $m$  and charge  $e$  is obtained from the uncharged Dirac Lagrangian (38.1) by changing the covariant derivative  $\mathbf{D}$  to the gauge covariant derivative  $\mathbf{D} + ie\mathbf{A}$ ,

$$L = \bar{\psi} \cdot (\mathbf{D} + ie\mathbf{A} + m)\psi. \quad (38.25)$$

Symmetrized with its complex conjugate, the charged Dirac Lagrangian (38.25) is

$$L = \frac{1}{2}\bar{\psi} \cdot (\mathbf{D} + ie\mathbf{A} + m)\psi - \frac{1}{2}\psi \cdot (\mathbf{D} - ie\mathbf{A} + m)\bar{\psi}. \quad (38.26)$$

Going through the same steps as in §38.1.1 yields Euler-Lagrange equations of motion

$$\frac{1}{2}\mathbf{D}\pi \cdot = -\frac{\partial L}{\partial \psi} = -\left(\frac{1}{2}\mathbf{D} - ie\mathbf{A} + m\right)\bar{\psi} \cdot, \quad \frac{1}{2}\mathbf{D}\bar{\pi} \cdot = \frac{\partial L}{\partial \bar{\psi}} = -\left(\frac{1}{2}\mathbf{D} + ie\mathbf{A} + m\right)\psi \cdot, \quad (38.27)$$

which simplify to the charged Dirac equations

$$\boxed{(\mathbf{D} + ie\mathbf{A} + m)\psi = 0}, \quad (38.28a)$$

$$\boxed{(\mathbf{D} - ie\mathbf{A} + m)\bar{\psi} = 0}. \quad (38.28b)$$

The Dirac equation for the conjugate field  $\bar{\psi}$  looks like that for the field  $\psi$  but with opposite charge  $e$ , which explains why  $\bar{\psi}$  is commonly called the charge conjugate field.

The charged Dirac field has an electric current  $\mathbf{j}$  given by the product of the charge  $e$  and the conserved number current  $\mathbf{n} \equiv \gamma_m n^m$ , equation (38.16),

$$\mathbf{j} \equiv e\mathbf{n}. \quad (38.29)$$

Like the number current, equation (38.17), the electric current  $\mathbf{j}$  is covariantly conserved,

$$\mathbf{D} \cdot \mathbf{j} = 0. \quad (38.30)$$

Current conservation is a consequence of the invariance of the Lagrangian under a global electromagnetic gauge transformation (38.22).

The electromagnetic contribution to the Dirac Lagrangian (38.26) can be interpreted as describing the interaction between the electromagnetic field  $\mathbf{A}$  and the Dirac electric current  $\mathbf{j}$ ,

$$L_{\text{int}} = ie \bar{\psi} \cdot \mathbf{A} \psi = \mathbf{A} \cdot \mathbf{j} . \quad (38.31)$$

### 38.2.1 Dirac Super-Hamiltonian

A super-Hamiltonian approach to the Dirac spinor field brings additional insights, and provides a blueprint to quantization.

The Dirac super-Hamiltonian  $H$  is defined in terms of the Lagrangian  $L$  by

$$H \equiv \frac{1}{2}\pi \cdot \mathbf{D}\psi - \frac{1}{2}\bar{\pi} \cdot \mathbf{D}\bar{\psi} - L , \quad (38.32)$$

the momenta canonically conjugate to  $\psi$  and  $\bar{\psi}$  being  $\frac{1}{2}\pi$  and  $-\frac{1}{2}\bar{\pi}$ , equation (38.7). With the Lagrangian and momenta from equation (38.3), the uncharged Dirac super-Hamiltonian  $H$  is

$$H = -m \bar{\psi} \cdot \psi . \quad (38.33)$$

A negative super-Hamiltonian signifies that the field is timelike. Do not confuse the super-Hamiltonian, which is a scalar associated with rest mass  $m$ , with the conventional Hamiltonian, which is a time-component of a 4-vector associated with energy.

The Hamiltonian is required to be written as a function  $H(\psi, \bar{\psi}, \pi, \bar{\pi})$  of the fields  $\psi, \bar{\psi}$  and their canonical momenta  $\frac{1}{2}\pi$  and  $-\frac{1}{2}\bar{\pi}$ . But there is ambiguity in how to write the Hamiltonian because the Lagrangian approach has shown that the conjugate momenta are proportional to the conjugate fields,  $\pi = \bar{\psi}$  and  $-\bar{\pi} = -\psi$ , equations (38.7). One way to proceed might be to discard the conjugate field  $\bar{\psi}$  as superfluous, and to write the Hamiltonian as

$$H = -m \pi \cdot \psi . \quad (38.34)$$

The resulting Hamilton's equations recover the Dirac equations for each of  $\psi$  and  $\pi$ , but the equations are decoupled, with no relation between  $\psi$  and  $\pi$ . One way to proceed would be to impose the condition  $\pi = \bar{\psi}$  ad hoc after the equations of motion are derived. In fact this “effective” approach is legitimate, but only because a more rigorous approach works. One might also try to add a constraint term to the Hamiltonian to enforce  $\pi = \bar{\psi}$ . However, the natural constraint term, a scalar quadratic term  $\mu(\pi - \bar{\psi}) \cdot (\bar{\pi} - \psi)$  with a Lagrange multiplier  $\mu$ , would merely force  $\pi - \bar{\psi}$  to be null, not necessarily zero.

An alternative approach is to take the ambiguity seriously, and to write the Hamiltonian in a symmetrical form that looks like a simple harmonic oscillator:

$$\boxed{H = -\frac{1}{2}m(-\bar{\pi} \cdot \pi + \bar{\psi} \cdot \psi)} . \quad (38.35)$$

Equation (38.35) is a valid option because  $-\bar{\pi} \cdot \pi = -\psi \cdot \bar{\psi} = \bar{\psi} \cdot \psi$ .

Hamilton's equations are obtained by varying the action

$$S = \int \left( \frac{1}{2}\pi \cdot \mathbf{D}\psi - \frac{1}{2}\bar{\pi} \cdot \mathbf{D}\bar{\psi} - H \right) d^4x \quad (38.36)$$

with respect to the fields  $\psi$  and  $\bar{\psi}$  and their conjugate momenta  $\frac{1}{2}\pi$  and  $-\frac{1}{2}\bar{\pi}$ . The variation of the first term in the integrand of equation (38.36) is

$$\delta(\pi \cdot \mathbf{D}\psi) = \pi \cdot \mathbf{D}(\delta\psi) - \mathbf{D}\psi \cdot \delta\pi = D_n(\pi \cdot \boldsymbol{\gamma}^n \delta\psi) + \mathbf{D}\pi \cdot \delta\psi - \mathbf{D}\psi \cdot \delta\pi , \quad (38.37)$$

the first term on the right hand side of which integrates to a surface term in accordance with Gauss' theorem (37.18) for true scalars. Varying the action (38.36) with respect to the fields and conjugate momenta thus gives

$$\begin{aligned} \delta S = & \oint \frac{1}{2} (\pi \cdot \boldsymbol{\gamma}_n \delta\psi - \bar{\pi} \cdot \boldsymbol{\gamma}_n \delta\bar{\psi}) d^3x^n \\ & + \int \left[ \left( \frac{1}{2} \mathbf{D}\pi \cdot - \frac{\partial H}{\partial \psi} \right) \delta\psi - \left( \frac{1}{2} \mathbf{D}\bar{\pi} \cdot + \frac{\partial H}{\partial \bar{\psi}} \right) \delta\bar{\psi} - \left( \frac{1}{2} \mathbf{D}\psi \cdot + \frac{\partial H}{\partial \pi} \right) \delta\pi + \left( \frac{1}{2} \mathbf{D}\bar{\psi} \cdot - \frac{\partial H}{\partial \bar{\pi}} \right) \delta\bar{\pi} \right] d^4x . \end{aligned} \quad (38.38)$$

Dropping the surface term and setting the variation to zero, as required by least action, yields Hamilton's equations for the fields and momenta

$$\frac{1}{2} \mathbf{D}\pi \cdot = \frac{\partial H}{\partial \psi} = -\frac{1}{2} m \bar{\psi} \cdot , \quad \frac{1}{2} \mathbf{D}\bar{\pi} \cdot = -\frac{\partial H}{\partial \bar{\psi}} = -\frac{1}{2} m \psi \cdot , \quad (38.39a)$$

$$\frac{1}{2} \mathbf{D}\psi \cdot = -\frac{\partial H}{\partial \pi} = -\frac{1}{2} m \bar{\pi} \cdot , \quad \frac{1}{2} \mathbf{D}\bar{\psi} \cdot = \frac{\partial H}{\partial \bar{\pi}} = -\frac{1}{2} m \pi \cdot . \quad (38.39b)$$

Dropping the trailing dots converts these row spinor equations to column spinor equations. Hamilton's equations (38.39) rearrange to

$$(\mathbf{D} + m)(\pi + \bar{\psi}) = 0 , \quad (\mathbf{D} + m)(\bar{\pi} + \psi) = 0 , \quad (38.40a)$$

$$(\mathbf{D} - m)(\pi - \bar{\psi}) = 0 , \quad (\mathbf{D} - m)(\bar{\pi} - \psi) = 0 . \quad (38.40b)$$

Hamilton's equations (38.40) appear to describe not only solutions of positive mass  $m$ , but also solutions of negative mass  $-m$ . If  $\pi = \bar{\psi}$  initially, then the negative mass Dirac equations (38.40b) ensure that  $\pi = \bar{\psi}$  thereafter, and then the positive mass Dirac equations (38.40a) reproduce the usual Dirac equations (38.12). But there are other solutions, satisfying  $\pi = -\bar{\psi}$ , for which the positive mass equations (38.40a) vanish, and the negative mass equations (38.40b) hold. This is Dirac's celebrated problem of negative mass states.

### 38.3 CPT-conjugates: Antiparticles

To understand the problem of negative mass states, suppose that either the positive or negative mass condition  $\pi = \pm \bar{\psi}$  is imposed, so that the Dirac Hamiltonian is

$$H = -m \bar{\psi} \cdot \psi . \quad (38.41)$$

Whereas the number density  $n^0 \equiv i\bar{\psi} \cdot \boldsymbol{\gamma}^0 \psi$  of the Dirac field is always positive, equation (38.18), the scalar product  $\bar{\psi} \cdot \psi$  can be either positive or negative. If  $\bar{\psi} \cdot \psi$  is positive, then the Hamiltonian (38.41) describes a timelike field. If on the other hand  $\bar{\psi} \cdot \psi$  is negative, then the field would appear to be spacelike. Only

massive fields have this problem: massless fields have  $\bar{\psi} \cdot \psi = 0$ , for which the Hamiltonian is lightlike. Again, do not confuse the scalar super-Hamiltonian (38.41) with the conventional Hamiltonian, which is the time component of a 4-vector.

Since the Hamiltonian is conserved in the absence of interactions, equation (38.66), a timelike Hamiltonian will remain timelike as the field evolves freely, so there is no immediate disaster. But the problem of a possibly spacelike Hamiltonian warrants closer examination.

A massive Dirac spinor inevitably contains contributions from both right- and left-handed chiral components, since pure right- or left-handed chiral Dirac fields are null. Thus a massive Dirac spinor inevitably contains contributions from both the field  $\psi$  and its conjugate  $\bar{\psi}$ . The positive mass Dirac equations (38.40a) imply that in flat (Minkowski) space, and in its own rest frame (where  $\psi \propto \gamma_{\uparrow\uparrow}$ ), a massive Dirac spinor  $\psi$  and its conjugate  $\bar{\psi}$  defined by equation (36.95) satisfy

$$\psi \propto e^{-imt} \gamma_{\uparrow\uparrow}, \quad \bar{\psi} \propto -e^{imt} \gamma_{\downarrow\downarrow}. \quad (38.42)$$

Lorentz transformed by rotor  $R$  into an arbitrary inertial frame, the Dirac spinor  $\psi$  and its conjugate  $\bar{\psi}$  satisfy

$$\psi \propto e^{-i(\omega t - k_a x^a)} R \gamma_{\uparrow\uparrow}, \quad \bar{\psi} \propto -e^{i(\omega t - k_a x^a)} R^* \gamma_{\downarrow\downarrow}. \quad (38.43)$$

If the field  $\psi$  has positive frequency  $\omega$ , then the conjugate field  $\bar{\psi}$  must have negative frequency. Thus it appears that the conjugate field  $\bar{\psi}$  has negative mass. The conclusion is confirmed by reordering the spinors in the Hamiltonian (38.41), which changes the apparent sign because of the antisymmetry of the Dirac spinor scalar product,

$$H = m \psi \cdot \bar{\psi}. \quad (38.44)$$

If  $\bar{\psi} \cdot \psi$  is positive, then the field  $\psi$  dominates, and the Hamiltonian (38.41) is timelike as it should be. If on the other hand  $\psi \cdot \bar{\psi}$  is positive, then the conjugate field  $\bar{\psi}$  dominates, and either the Hamiltonian (38.44) is spacelike, which is unphysical, or else the mass  $m$  is negative.

Sensible behaviour when the conjugate field  $\bar{\psi}$  dominates is restored by taking  $PT$  conjugates of the fields. Let primed fields denote the  $PT$ -conjugate fields obtained by pre-multiplying by the pseudoscalar, in accordance with equation (36.134),

$$\psi' \equiv I\psi, \quad \bar{\psi}' \equiv I\bar{\psi}. \quad (38.45)$$

The  $PT$ -conjugate fields  $\psi'$  and  $\bar{\psi}'$  are conjugates to each other, equation (36.95), since  $CI^* = IC$ . Since the pseudoscalar satisfies  $I^2 = -1$ , the Hamiltonian of the  $CPT$ -conjugate field  $\bar{\psi}'$  is

$$H = -m \psi' \cdot \bar{\psi}', \quad (38.46)$$

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  $\psi$  and  $\bar{\psi}$  cancel each other.

The conclusion is that the physical antiparticle field associated with the particle field  $\psi$  is not the  $C$ -conjugate field  $\bar{\psi}$  itself, but rather the  $CPT$ -conjugate field  $\bar{\psi}'$ . Since  $C$  flips the mass, while  $PT$  flips

the spacetime directions, one arrives at Feynman's famous aphorism, **an antiparticle is a negative mass particle moving backwards in time**. Of course, negative mass going backwards in time is indistinguishable from positive mass going forwards in time.

### 38.4 Negative mass particles?

The identification of antiparticles as *CPT* conjugates solves the problem of negative mass Dirac particles. The charged Dirac Lagrangian (38.26) rewritten in terms of the *CPT*-conjugate fields is

$$L = \frac{1}{2} \bar{\psi}' \cdot (\mathbf{D} + ie\mathbf{A} - m) \psi' - \frac{1}{2} \psi' \cdot (\mathbf{D} - ie\mathbf{A} - m) \bar{\psi}' . \quad (38.47)$$

The mass  $m$  terms acquire a minus sign because  $I^2 = -1$ , while the covariant derivative terms  $\mathbf{D} \pm ie\mathbf{A}$  are unchanged in sign because in addition  $I$  anticommutes with the basis vectors  $\boldsymbol{\gamma}^n$ . The resulting Euler-Lagrange equations of motion for the *CPT*-conjugate fields are

$$(\mathbf{D} + ie\mathbf{A} - m)\psi' = 0 , \quad (38.48a)$$

$$(\mathbf{D} - ie\mathbf{A} - m)\bar{\psi}' = 0 . \quad (38.48b)$$

Thus the *CPT*-conjugate fields satisfy a Dirac equation of apparently opposite sign of the mass  $m$ , solving the problem of where the negative mass particles are.

As a sanity check that the signs are right, consider flat space, and the non-relativistic limit where the time derivative  $\partial_0$  dominates over the spatial derivatives, and the electromagnetic potential is a weak, pure electric potential  $A_0$ . If the particle field  $\psi$  dominates, then the  $\psi^\dagger$  components of the Dirac spinor field dominate, and  $\boldsymbol{\gamma}_0 = i$  acting on these components, §14.8. Conversely if the antiparticle field  $\bar{\psi}'$  dominates, then the  $\psi^\dagger$  components of the spinor field dominate, and  $\boldsymbol{\gamma}_0 = -i$  acting on these components. In this limit, the Dirac equation for the field  $\psi$  and the *CPT*-conjugate field  $\bar{\psi}'$  are

$$(i\partial_0 - eA_0 - m)\psi = 0 , \quad (38.49a)$$

$$(i\partial_0 + eA_0 - m)\bar{\psi}' = 0 . \quad (38.49b)$$

The particle and antiparticle fields have opposite electric charges, and both fields have positive mass, positive frequency,  $\psi \sim e^{-imt}$  and  $\bar{\psi}' \sim e^{-imt}$ .

The same Lagrangian or super-Hamiltonian serves to describe both particles and antiparticles. The condition that the super-Hamiltonian be timelike determines whether a particle or antiparticle is present.

The electric current  $j^k$  has the same sign regardless of which field is considered dominant:

$$j^k = ie \bar{\psi} \cdot \boldsymbol{\gamma}^k \psi = ie \psi \cdot \boldsymbol{\gamma}^k \bar{\psi} = ie \psi' \cdot \boldsymbol{\gamma}^k \bar{\psi}' . \quad (38.50)$$

The equality of the middle two expressions is from equations (38.15), and the last follows because  $I^2 = -1$  and  $I$  anticommutes with  $\boldsymbol{\gamma}^k$ , equation (14.10).

It is sometimes stated that the problem of negative mass states is solved only by quantum field theory. But the problem and its solution are evident already at the classical level. The negative mass field is the

conjugate field  $\bar{\psi}$ , but the Hamiltonian is spacelike, therefore unphysical, if the conjugate field dominates, that is, if  $\psi \cdot \bar{\psi}$  is positive. The *CPT*-conjugate field  $\bar{\psi}'$  has negative mass and goes backwards in spacetime, but has a physical timelike Hamiltonian when the *CPT*-conjugate field dominates, that is, when  $\psi' \cdot \bar{\psi}'$  is positive. The *CPT*-conjugate field corresponds to an antiparticle that has positive mass going forwards in time.

### 38.4.1 Chiral decomposition of the Dirac Lagrangian

If a Dirac spinor is decomposed into its right- and left-handed chiral (Weyl) components,  $\psi = \psi_R + \psi_L$ , the Dirac Lagrangian (38.3) becomes

$$L = \overline{\psi_R} \cdot \mathbf{D}\psi_R + \overline{\psi_L} \cdot \mathbf{D}\psi_L + m(\overline{\psi_L} \cdot \psi_R + \overline{\psi_R} \cdot \psi_L). \quad (38.51)$$

Note that conjugation flips chirality, so that, for example, the conjugate  $\overline{\psi_R}$  of the right-handed component is left-handed. Equation (38.51) holds because the derivative  $\mathbf{D} \equiv \boldsymbol{\gamma}^n D_n$  connects components of opposite chirality, while the mass term  $m$  connects components of the same chirality.

A purely chiral Dirac spinor must be massless. However, a massive Dirac spinor can appear to be nearly chiral if it is moving relativistically. For example, the neutrino  $\nu_L$  is left-handed and almost massless. Let  $\psi$  be a Dirac spinor whose spin points in the negative 3-direction  $\downarrow$ . The spinor then takes the form

$$\psi = \psi^{V\downarrow} \gamma_{V\downarrow} + \psi^{U\downarrow} \gamma_{U\downarrow}. \quad (38.52)$$

Under a Lorentz boost by rapidity  $\theta$  in the positive 3-direction, the spinor becomes

$$\psi = e^{\theta/2} \psi^{V\downarrow} \gamma_{V\downarrow} + e^{-\theta/2} \psi^{U\downarrow} \gamma_{U\downarrow}. \quad (38.53)$$

If the boost  $\theta$  is large and positive, then the spinor will appear to be almost entirely left-handed. Conversely if the boost  $\theta$  is large and negative, then the spinor will appear to be almost entirely right-handed. In the middle, there is a rest-frame where the spinor is equal parts left and right,

$$|\psi^{V\downarrow}| = |\psi^{U\downarrow}|. \quad (38.54)$$

To the extent that a neutrino  $\nu$  is a massless left-handed Dirac spinor, its Lagrangian approximates

$$L \approx \overline{\nu_L} \cdot \mathbf{D}\nu_L. \quad (38.55)$$

Symmetrized with its complex conjugate, the neutrino Lagrangian (38.55) is

$$L \approx \frac{1}{2} \overline{\nu_L} \cdot \mathbf{D}\nu_L - \frac{1}{2} \nu_L \cdot \mathbf{D}\overline{\nu_L}. \quad (38.56)$$

The antineutrino  $\bar{\nu}_R$  (the bar to denote the antineutrino is conventional) is the *CPT*-conjugate of the left-handed neutrino, and is thus right-handed,

$$\bar{\nu}_R \equiv I\overline{\nu_L}. \quad (38.57)$$

The Lagrangian (38.55) can be rewritten in terms of the *CPT*-conjugate antineutrino

$$L \approx -\overline{\bar{\nu}_R} \cdot \mathbf{D}\bar{\nu}_R. \quad (38.58)$$

### 38.5 Super-Hamiltonian as generator of evolution

Let  $f(\psi, \bar{\psi})$  be some function of the field  $\psi$  and its conjugate  $\bar{\psi}$  that is either a true scalar, or a column spinor. The evolution of such a true scalar or spinor function  $f$  can be expressed as

$$\mathbf{D}f = \frac{\partial f}{\partial \psi} \mathbf{D}\psi + \frac{\partial f}{\partial \bar{\psi}} \mathbf{D}\bar{\psi} . \quad (38.59)$$

The covariant spacetime derivative  $\mathbf{D}f$  is a true scalar if  $f$  is a true scalar, and a spinor if  $f$  is a spinor. Substituting Hamilton's equations (38.39), and taking into account factors of 2 from  $\pi = \bar{\psi}$ , and a sign flip from equation (38.5), gives

$$\mathbf{D}f = \frac{\partial f}{\partial \psi} \frac{\partial H}{\partial \bar{\psi}} - \frac{\partial f}{\partial \bar{\psi}} \frac{\partial H}{\partial \psi} . \quad (38.60)$$

Define the Poisson bracket  $\{f, h\}$ , not to be confused with the anticommutator, by

$$\{f, h\} \equiv \frac{\partial f}{\partial \psi} \frac{\partial h}{\partial \bar{\psi}} - \frac{\partial f}{\partial \bar{\psi}} \frac{\partial h}{\partial \psi} . \quad (38.61)$$

Then the evolution equation (38.60) for the function  $f$  can be written as its Poisson bracket with the super-Hamiltonian,

$$\mathbf{D}f = \{f, H\} . \quad (38.62)$$

Equation (38.62) encapsulates mathematically the physical idea that the super-Hamiltonian is a generator of evolution.

If  $f$  and  $h$  are both true scalars, then their Poisson bracket is symmetric

$$\{f, h\} = \{h, f\} \quad \text{if } f \text{ and } h \text{ are both true scalars} . \quad (38.63)$$

To derive equation (38.63), note that each factor, such as  $(\partial f / \partial \psi)(\partial h / \partial \bar{\psi})$ , on the right hand side of equation (38.61) is a scalar product of a row spinor with a column spinor, in accordance with equation (38.5). Sign flips come from exchange of the two terms on the right hand side of equation (38.61), from the antisymmetry of the spinor scalar product, and from a pair of sign flips from swapping derivatives with respect to row and column spinors, per equation (38.5), a total of four sign flips altogether.

The Poisson bracket is not symmetric if either  $f$  or  $h$  is a spinor. The Poisson brackets of the fields  $\psi$  and  $\bar{\psi}$  with themselves and with each other are

$$\{\psi, \psi\} = \{\bar{\psi}, \bar{\psi}\} = 0 , \quad (38.64a)$$

$$\{\psi, \bar{\psi}\} = -\{\bar{\psi}, \psi\} = 1 . \quad (38.64b)$$

The Poisson brackets (38.64) of the fields are true scalars.

A quantity that is conserved satisfies

$$\mathbf{D}f = \{f, H\} = 0 . \quad (38.65)$$

Explicit calculation from the Hamiltonian (38.41) shows that the super-Hamiltonian itself is conserved,

$$\mathbf{D}H = \{H, H\} = 0 . \quad (38.66)$$

## 38.6 Majorana spinor field

A Majorana spinor is a spinor that satisfies the Majorana scalar product condition, equation (36.37). The Dirac Lagrangian (38.3) vanishes if the scalar product is Majorana instead of Dirac. An alternative Lagrangian that does not vanish is the Majorana Lagrangian (38.67).

### 38.6.1 Majorana Lagrangian

The Lagrangian for a complex, massive, charged Majorana spinor  $\psi$  is

$$L = I\bar{\psi} \cdot (\mathbf{D} + ie\mathbf{A} + m)\psi . \quad (38.67)$$

The Majorana Lagrangian (38.67) differs from the Dirac Lagrangian (38.25) in that it involves the spinor field  $\psi$  and its *CPT*-conjugate  $I\bar{\psi}$  in place of the *C*-conjugate  $\bar{\psi}$ , and the scalar product in equation (38.67) is the Majorana scalar product, not the Dirac scalar product, equation (36.37). The Majorana Lagrangian (38.67) is valid in any representation of the  $\gamma$ -matrices, Dirac, chiral, or Majorana. The Majorana Lagrangian (38.67) allows for the possibility of a complex Majorana field  $\psi$ .

As previously with the Dirac Lagrangian (38.1), the Majorana Lagrangian (38.70) as it stands is neither real, nor symmetric with respect to the fields  $\psi$  and  $I\bar{\psi}$ . The covariant spacetime derivative  $\mathbf{D} = \boldsymbol{\gamma}_n D^n$ , the gauge field  $\mathbf{A} = A^n \boldsymbol{\gamma}_n$ , and the pseudoscalar  $I$  are all multivectors with real coefficients in an orthonormal basis. For the Majorana scalar product, the complex conjugate of any multivector  $\mathbf{a} = a^A \boldsymbol{\gamma}_A$  with real coefficients  $a^A$  in an orthonormal basis satisfies

$$(\bar{\psi} \cdot \mathbf{a}\psi)^* = \psi \cdot \mathbf{a}\bar{\psi} , \quad (38.68)$$

with a  $+$  sign on the right hand side, in contrast to the  $-$  for the Dirac scalar product, equation (38.2), because the conjugation operator  $C$  commutes with the Majorana spinor metric tensor, but anticommutes with the Dirac spinor metric tensor, equation (36.93). The Majorana Lagrangian (38.70) involves the *CPT*-conjugate  $I\bar{\psi}$  rather than the *C*-conjugate  $\bar{\psi}$ . With the additional factor of the pseudoscalar  $I$ , the complex conjugate of any grade- $p$  multivector  $\mathbf{a}$  with real coefficients satisfies

$$(I\bar{\psi} \cdot \mathbf{a}\psi)^* = (-)^p \psi \cdot \mathbf{a}I\bar{\psi} , \quad (38.69)$$

where the factor of  $(-)^p$  arises because  $I$  anticommutes with odd multivectors, but commutes with even multivectors. Symmetrizing the Lagrangian (38.67) with its complex conjugate gives the real Majorana Lagrangian for a complex Majorana field  $\psi$ ,

$$L = \frac{1}{2}I\bar{\psi} \cdot (\mathbf{D} + ie\mathbf{A} + m)\psi - \frac{1}{2}\psi \cdot (\mathbf{D} - ie\mathbf{A} - m)I\bar{\psi} . \quad (38.70)$$

The symmetry of the Majorana scalar product, equation (36.124), allows a non-vanishing mass  $m$  and charge  $e$ .

As with the Dirac Lagrangian, equation (38.4), the imaginary part of the derivative term of the Majorana Lagrangian integrates to a surface term, per Gauss' theorem (37.18) for true scalars,

$$i \int \text{Im}(I\bar{\psi} \cdot \mathbf{D}\psi) d^4x = \frac{1}{2} \int (I\bar{\psi} \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}I\bar{\psi}) d^4x = \frac{1}{2} \oint I\bar{\psi} \cdot \boldsymbol{\gamma}_n \psi d^3x^n , \quad (38.71)$$

and therefore has no effect on the equations of motion. Equation (38.71) reveals why the pseudoscalar factor  $I$  must be included in the Majorana Lagrangian. Without this factor it would be the real part, not the imaginary part, of the derivative term in the Majorana Lagrangian that would integrate to a surface term, and there would be no Majorana theory.

### 38.6.2 Majorana equation

In varying the Lagrangian (38.70) for a complex 8-component Majorana field, it is convenient to regard  $\psi$  and  $I\bar{\psi}$  as two independent fields of 4 components each. The momenta canonically conjugate to  $\psi$  and  $I\bar{\psi}$  are  $\frac{1}{2}\pi$  and  $\frac{1}{2}I\bar{\pi}$  given by

$$\frac{1}{2}\pi \cdot = \frac{\partial L}{\partial(\mathbf{D}\psi)} = \frac{1}{2}I\bar{\psi} \cdot , \quad \frac{1}{2}I\bar{\pi} \cdot = \frac{\partial L}{\partial(I\bar{\psi})} = -\frac{1}{2}\psi \cdot , \quad (38.72)$$

implying

$$\pi = I\bar{\psi} . \quad (38.73)$$

The Euler-Lagrange equations for the complex Majorana field are

$$(\mathbf{D} + ie\mathbf{A} + m)\psi = 0 , \quad (38.74a)$$

$$(\mathbf{D} - ie\mathbf{A} - m)I\bar{\psi} = 0 . \quad (38.74b)$$

The Majorana equations (38.74) are the same as those for the Dirac field  $\psi$ , equation (38.28a), and its *CPT*-conjugate  $\bar{\psi}' \equiv I\bar{\psi}$ , equation (38.48b).

### 38.6.3 Conserved Majorana current

The complex Majorana Lagrangian (38.70) is invariant under rotations of the Majorana fields by opposing complex phases,  $\psi \rightarrow e^{-i\epsilon}\psi$  and  $I\bar{\psi} \rightarrow e^{i\epsilon}I\bar{\psi}$ , similar to the symmetry of the Dirac Lagrangian. Noether's theorem takes the form

$$0 = \pi \cdot \mathbf{D}\psi - I\bar{\pi} \cdot \mathbf{D}I\bar{\psi} = I\bar{\psi} \cdot \mathbf{D}\psi + \psi \cdot \mathbf{D}I\bar{\psi} = D_m(I\bar{\psi} \cdot \boldsymbol{\gamma}^m \psi) , \quad (38.75)$$

where the last step is an application of Gauss' theorem (37.18) for true scalars. The conserved Majorana current is

$$n^m \equiv I\bar{\psi} \cdot \boldsymbol{\gamma}^m \psi = i\bar{\psi} \cdot \gamma_5 \boldsymbol{\gamma}^m \psi = i\psi^\dagger C\varepsilon\gamma_5 \boldsymbol{\gamma}^m \psi = \psi^\dagger \gamma_0 \boldsymbol{\gamma}^m \psi . \quad (38.76)$$

The Majorana current is covariantly conserved,

$$D_m n^m = 0 , \quad (38.77)$$

and its time component  $n^0$  is the positive number

$$n^0 = \psi^\dagger \psi . \quad (38.78)$$

The current (38.76) may be interpreted as a conserved probability current for the Majorana field.

### 38.6.4 Chiral decomposition of the Majorana field

If the complex Majorana field  $\psi$  is decomposed into its right- and left-handed chiral components,  $\psi = \psi_R + \psi_L$ , then the Majorana Lagrangian (38.67) becomes, similarly to the Dirac Lagrangian (38.51),

$$L = I\bar{\psi}_R \cdot \mathbf{D}\psi_R + I\bar{\psi}_L \cdot \mathbf{D}\psi_L + Im(\bar{\psi}_L \cdot \psi_R + \bar{\psi}_R \cdot \psi_L) . \quad (38.79)$$

If the Majorana field  $\psi$  is a 4-component self-conjugate field, rather than an 8-component complex field, then

$$L = I\bar{\psi}_L \cdot \mathbf{D}\psi_L + Im\psi_L \cdot \psi_L . \quad (38.80)$$

The self-conjugate Lagrangian (38.80) is deceiving because it looks like only a left-handed chiral spinor  $\psi_L$  is present. As discussed in §36.10, a single real self-conjugate Majorana spinor cannot have definite chirality. However, a complex linear combination of self-conjugate Majorana spinors can have definite chirality. The Lagrangian (38.80) appears to describe a single left-handed self-conjugate Majorana spinor because the Lagrangian is actually complex. When the Lagrangian (38.80) is symmetrized with its complex conjugate, then it becomes evident that a combination of right- and left-handed spinors is present.

## 38.7 Electromagnetic field

The next field to consider is the electromagnetic field, which is a massless spin-1 field. As with the Dirac spin- $\frac{1}{2}$  field, the electromagnetic field is well understood theoretically, thanks to ample experimental evidence.

### 38.7.1 Electromagnetic Lagrangian

As seen in §38.2, electromagnetism emerges when the global invariance of the Dirac Lagrangian under rotations of the spinor field  $\psi$  by a complex phase is elevated to a local invariance. The electromagnetic potential  $\mathbf{A}$  is the connection in the gauge-covariant derivative, analogous to the connection  $\mathbf{\Gamma}$  of the general relativistic covariant derivative.

The electromagnetic Lagrangian must be constructed from the electromagnetic potential  $\mathbf{A}$  and its derivative. WHY TORSION-FREE? TO PRESERVE GAUGE INVARIANCE. The torsion-free covariant derivative of the field is

$$\mathbf{D}\mathbf{A} = \mathbf{D} \cdot \mathbf{A} + \mathbf{D} \wedge \mathbf{A} . \quad (38.81)$$

In index notation, the derivative  $\mathbf{D}\mathbf{A}$  is

$$\mathbf{D}\mathbf{A} = \mathbf{D} \cdot \mathbf{A} + \mathbf{D} \wedge \mathbf{A} = D_n A^n + \frac{1}{2}(D_m A_n - D_n A_m) \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n , \quad (38.82)$$

where the factor of  $\frac{1}{2}$  in the bivector term cancels the double counting of distinct pairs  $mn$ . Neither the potential  $\mathbf{A}$  nor its divergence  $\mathbf{D} \cdot \mathbf{A}$  is gauge-invariant, but its curl  $\mathbf{D} \wedge \mathbf{A}$  is gauge-invariant. Here gauge-invariant

means invariant under electromagnetic gauge transformations (38.24). The curl is also by construction coordinate and tetrad gauge-invariant. The gauge-invariant curl defines the bivector  $\mathbf{F}$  that is conventionally called the electromagnetic field,

$$\mathbf{F} \equiv -\mathbf{D} \wedge \mathbf{A} . \quad (38.83)$$

In components, the electromagnetic field  $\mathbf{F} = \frac{1}{2} F_{mn} \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n$  is

$$F_{mn} = D_n A_m - D_m A_n . \quad (38.84)$$

Since the electromagnetic Lagrangian must be gauge-invariant, it must be a scalar formed from the field  $\mathbf{F}$ ,

$$L = -\frac{1}{8\pi} \mathbf{F} \cdot \mathbf{F} + \mathbf{A} \cdot \mathbf{j} . \quad (38.85)$$

The term  $\mathbf{A} \cdot \mathbf{j}$  is the contribution to the Lagrangian from the interaction of the electromagnetic field with an electric current  $\mathbf{j}$ , consistent with equation (38.31). The  $1/(8\pi)$  factor holds when the electromagnetic field is expressed in Gaussian units. High-energy physicists commonly use SI or rationalized units, in which case the factor  $1/(8\pi)$  should be replaced by  $1/2$ . The factor of  $4\pi$  in Maxwell's equations (38.94) disappears in rationalized units. If the electromagnetic field had mass  $m$ , then the electromagnetic Lagrangian would contain a mass term  $-(1/8\pi)m^2 \mathbf{A} \cdot \mathbf{A}$ . However, such a mass term would violate electromagnetic gauge-invariance, spoiling the theory. REALLY? THIS IS PROCA THEORY. THEN  $\mathbf{D} \cdot \mathbf{A} = 0$  EMERGES AS EQ OF MOT? In index notation, the electromagnetic Lagrangian (38.85) is

$$L = -\frac{1}{16\pi} F_{mn} F^{mn} + A_n j^n . \quad (38.86)$$

To apply least action, the Lagrangian must written as a function  $L(\mathbf{A}, \mathbf{D} \wedge \mathbf{A})$  of the “field”  $\mathbf{A}$  and its velocity, which is, apparently, the curl  $\mathbf{D} \wedge \mathbf{A}$ ,

$$L = -\frac{1}{8\pi} (\mathbf{D} \wedge \mathbf{A}) \cdot (\mathbf{D} \wedge \mathbf{A}) + \mathbf{A} \cdot \mathbf{j} . \quad (38.87)$$

Canonically conjugate momenta are obtained by differentiating the Lagrangian with respect to the velocity. The derivative of a scalar product  $\mathbf{a} \cdot \mathbf{b}$  of two same-grade multivectors  $\mathbf{a} \equiv a^A \boldsymbol{\gamma}_A$  and  $\mathbf{b} \equiv b^B \boldsymbol{\gamma}_B$  with respect to the multivector  $\mathbf{a}$  is defined naturally by

$$\frac{\partial(\mathbf{a} \cdot \mathbf{b})}{\partial \mathbf{a}} \equiv \frac{\partial(a_B b^B)}{\partial a_A} \boldsymbol{\gamma}_A = b^A \boldsymbol{\gamma}_A = \mathbf{b} . \quad (38.88)$$

The momentum canonically conjugate to the electromagnetic vector potential  $\mathbf{A}$  is proportional to the field  $\mathbf{F}$ ,

$$\frac{\partial L}{\partial(\mathbf{D} \wedge \mathbf{A})} = -\frac{1}{4\pi} \mathbf{D} \wedge \mathbf{A} = \frac{1}{4\pi} \mathbf{F} . \quad (38.89)$$

Variation of the action (a careful variation follows equation (38.98)) yields the Euler-Lagrangian equations of motion for the electromagnetic field,

$$\frac{1}{4\pi} \mathbf{D} \cdot \mathbf{F} = \frac{\partial L}{\partial \mathbf{A}} = \mathbf{j} . \quad (38.90)$$

The covariant curl of  $\mathbf{F}$  vanishes by the rule (15.68) for  $\mathbf{D} \wedge \mathbf{D}$ ,

$$\mathbf{D} \wedge \mathbf{F} = -\mathbf{D} \wedge \mathbf{D} \wedge \mathbf{A} = 0 . \quad (38.91)$$

Equations (38.91) are the Bianchi identities for the electromagnetic field. The Bianchi identities follow also from the Jacobi identity applied to the gauge-covariant derivative  $\mathcal{D} \equiv \mathbf{D} + ie\mathbf{A}$ ,

$$[\mathcal{D}_k, [\mathcal{D}_l, \mathcal{D}_m]] = ieD_{[k}F_{lm]} = 0 . \quad (38.92)$$

The two sets of equations (38.90) and (38.91) form the source and source-free parts of Maxwell's equations:

$$\mathbf{D} \cdot \mathbf{F} = 4\pi\mathbf{j} , \quad (38.93a)$$

$$\mathbf{D} \wedge \mathbf{F} = 0 . \quad (38.93b)$$

The full set of Maxwell's equations can be expressed in an impressively economical fashion by the single multivector equation

$$\boxed{\mathbf{DF} = 4\pi\mathbf{j}} . \quad (38.94)$$

The left hand side of equation (38.94) contains two contributions, a vector and trivector, which constitute respectively the source and source-free parts of Maxwell's equations. In components, Maxwell's equations (38.93) are

$$D_m F^{mn} = 4\pi j^n , \quad (38.95a)$$

$$D_k F_{mn} + D_m F_{nk} + D_n F_{km} = 0 . \quad (38.95b)$$

### 38.7.2 Electromagnetic super-Hamiltonian

The electromagnetic super-Hamiltonian  $H(\mathbf{A}, \mathbf{F})$  is defined by

$$H \equiv \frac{1}{4\pi} \mathbf{F} \cdot (\mathbf{D} \wedge \mathbf{A}) - L , \quad (38.96)$$

which expressed as required in terms of the “field”  $\mathbf{A}$  and its momentum  $\mathbf{F}$  is

$$H = -\frac{1}{8\pi} \mathbf{F} \cdot \mathbf{F} - \mathbf{A} \cdot \mathbf{j} . \quad (38.97)$$

The electromagnetic part of the super-Hamiltonian (38.97) is the same as the Lagrangian, but the interaction part has opposite sign to the Lagrangian. Hamilton's equations are obtained by extremizing the action

$$S = \int \left[ \frac{1}{4\pi} \mathbf{F} \cdot (\mathbf{D} \wedge \mathbf{A}) - H \right] d^4x . \quad (38.98)$$

Variation of the first term in the integrand on the right hand side of equation (38.98) gives

$$\delta(\mathbf{F} \cdot (\mathbf{D} \wedge \mathbf{A})) = \mathbf{F} \cdot (\mathbf{D} \wedge \delta\mathbf{A}) + (\delta\mathbf{F}) \cdot (\mathbf{D} \wedge \mathbf{A}) = D_m(F^{mn}\delta A_n) - (D_m F^{mn})\delta A_n + (\delta\mathbf{F}) \cdot (\mathbf{D} \wedge \mathbf{A}) . \quad (38.99)$$

Varying the action (38.98) with respect to the potential  $\mathbf{A}$  and its conjugate momentum  $\mathbf{F}$  gives

$$\delta S = \frac{1}{4\pi} \oint F_{mn} \delta A^n d^3x^m + \int \left[ -\left( \frac{1}{4\pi} \mathbf{D} \cdot \mathbf{F} + \frac{\partial H}{\partial \mathbf{A}} \right) \cdot \delta \mathbf{A} + \delta \mathbf{F} \cdot \left( \frac{1}{4\pi} \mathbf{D} \wedge \mathbf{A} - \frac{\partial H}{\partial \mathbf{F}} \right) \right] d^4x . \quad (38.100)$$

Hamilton's equations are thus

$$\mathbf{D} \cdot \mathbf{F} = -4\pi \frac{\partial H}{\partial \mathbf{A}} = 4\pi \mathbf{j} , \quad \mathbf{D} \wedge \mathbf{A} = 4\pi \frac{\partial H}{\partial \mathbf{F}} = -\mathbf{F} , \quad (38.101)$$

reproducing equations (38.94) from the Lagrangian approach.

### 38.7.3 Electric and magnetic fields

The electromagnetic field  $\mathbf{F}$  is a bivector, and as such has an intrinsic complex (with respect to  $I$ ) structure. The real and imaginary parts of this complex structure define the electric and magnetic field bivectors  $\mathbf{E} = E^a \iota_a$  and  $\mathbf{B} = B^a \iota_a$ ,

$$\mathbf{F} = I(\mathbf{E} + i\mathbf{B}) . \quad (38.102)$$

Since the pseudoscalar  $I$  flips sign under spatial inversion, the electric field bivector  $\mathbf{E}$  changes sign under spatial inversion, whereas the magnetic field bivector  $\mathbf{B}$  does not.

In flat space with a Minkowski metric, the source-free Maxwell's equations (38.93b) reduce to the traditional form

$$\nabla \cdot \mathbf{B} = 0 , \quad \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 , \quad (38.103)$$

where  $\mathbf{E}$  and  $\mathbf{B}$  are interpreted as 3-vectors, and  $\nabla$  is the spatial 3-gradient. The sourced Maxwell's equations (38.93a) reduce to

$$\nabla \cdot \mathbf{E} = 4\pi q , \quad \nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} = 4\pi \mathbf{j} , \quad (38.104)$$

where the electric charge density  $q$  and the electric 3-current density  $\mathbf{j}$  are the time and space components of the electric 4-current  $j^n$ ,

$$j^n = \{q, \mathbf{j}\} . \quad (38.105)$$

### 38.7.4 Chiral decomposition of the electromagnetic field

The electromagnetic field  $\mathbf{F}$  decomposes into right- and left-handed components  $\mathbf{F}_R$  and  $\mathbf{F}_L$

$$\mathbf{F} = \mathbf{F}_R + \mathbf{F}_L , \quad (38.106)$$

defined to be the chiral projections of the field  $\mathbf{F}$

$$\mathbf{F}_R = \frac{1}{2}(1 \pm \gamma_5)\mathbf{F} = \frac{1}{2}(1 \mp iI)\mathbf{F} . \quad (38.107)$$

The chiral decomposition is Lorentz-invariant, since the pseudoscalar  $I$  is Lorentz-invariant. In the absence of an electric current, the chiral components of the free electromagnetic field propagate independently, without mixing. In terms of the electric and magnetic fields  $\mathbf{E}$  and  $\mathbf{B}$ , the chiral components are

$$\mathbf{F}_{\text{R}} = \frac{1}{2}I(1 \pm \gamma_5)(\mathbf{E} + I\mathbf{B}) = \frac{1}{2}I(1 \pm \gamma_5)(\mathbf{E} \pm i\mathbf{B}) . \quad (38.108)$$

The chiral components can be expressed in terms of the field  $\mathbf{F}$  and its Hodge dual  ${}^*\mathbf{F} \equiv I\mathbf{F}$  as

$$\mathbf{F}_{\text{R}} = \frac{1}{2}(\mathbf{F} \mp i{}^*\mathbf{F}) . \quad (38.109)$$

In index notation,

$$\mathbf{F}_{\text{R}} = \frac{1}{4} \left( F^{kl} \pm \frac{i}{2} \varepsilon^{kl}{}_{mn} F^{mn} \right) \boldsymbol{\gamma}_k \wedge \boldsymbol{\gamma}_l , \quad (38.110)$$

the factor and sign flips coming from  $I\boldsymbol{\gamma}_k \wedge \boldsymbol{\gamma}_l = -\frac{1}{2}\varepsilon_{kl}{}^{mn}\boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n$ .

In terms of the chiral components of the field, the electromagnetic Lagrangian takes the form

$$L = -\frac{1}{8\pi} (\mathbf{F}_{\text{R}} \cdot \mathbf{F}_{\text{R}} + \mathbf{F}_{\text{L}} \cdot \mathbf{F}_{\text{L}}) . \quad (38.111)$$

The product of right- and left-handed electromagnetic fields vanishes

$$\mathbf{F}_{\text{R}} \cdot \mathbf{F}_{\text{L}} = \frac{1}{4}(\mathbf{F} - iI\mathbf{F})(\mathbf{F} + iI\mathbf{F}) = \frac{1}{4}(\mathbf{F}^2 - \mathbf{F}^2) = 0 . \quad (38.112)$$

Maxwell's equations (38.94) can be written

$$\mathbf{D} \cdot \mathbf{F}_{\text{R}} = 2\pi \mathbf{j} , \quad (38.113)$$

or in index notation,

$$\frac{1}{2} D_k \left( F^{kl} + \frac{i}{2} \varepsilon^{kl}{}_{mn} F^{mn} \right) = 2\pi j^l . \quad (38.114)$$

In the Dirac or chiral representations of the orthonormal  $\gamma$ -matrices, the coefficients  $F^{kl}$  are real, and the totally antisymmetric matrix  $\varepsilon^{klmn}$  is also real. In the Dirac or chiral representations, the real and imaginary (with respect to  $i$ ) parts of equations (38.114) constitute respectively the source and source-free parts of Maxwell's equations (38.93). The electric current vector  $j^l$  is pure real. The imaginary part of  $j^l$  would correspond to magnetic charge, which has not (yet) been observed to exist. Whereas equations (38.94) contained all of Maxwell's equations as a combination of a vector and a pseudovector, equations (38.114) contain all of Maxwell's equations as a complex (with respect to  $i$ ) vector.

**Exercise 38.1** The geometric product of the electromagnetic field  $\mathbf{F}$  with itself contains a scalar and a pseudoscalar part, the bivector part vanishing identically,

$$\mathbf{FF} = \mathbf{F} \cdot \mathbf{F} + \mathbf{F} \wedge \mathbf{F} . \quad (38.115)$$

Consider forming a Lagrangian from the pseudoscalar part,

$$IL = -\frac{1}{8\pi} \mathbf{F} \wedge \mathbf{F} . \quad (38.116)$$

Using the associativity of the wedge product, equation (13.31), the commutation rule (15.41) for the covariant curl, and the rule (15.68) for  $\mathbf{D} \wedge \mathbf{D}$ , argue that the pseudoscalar  $\mathbf{F} \wedge \mathbf{F}$  can be expressed as a covariant curl,

$$\mathbf{F} \wedge \mathbf{F} = \mathbf{D} \wedge (\mathbf{A} \wedge \mathbf{D} \wedge \mathbf{A}) . \quad (38.117)$$

Applying Stokes' theorem (15.88), conclude that the variation of the action leads to a pure surface term

$$\delta S = \int IL d^4x^{0123} = -\frac{1}{8\pi} \oint (\mathbf{A} \wedge \mathbf{D} \wedge \mathbf{A})_B d^3x^B , \quad (38.118)$$

with  $B$  implicitly summed over distinct antisymmetric sequences of 3 indices.

**Exercise 38.2** Is a purely chiral Lagrangian

$$L = -\frac{1}{4\pi} \mathbf{F}_R \cdot \mathbf{F}_R \quad (38.119)$$

a viable electromagnetic Lagrangian?

**Solution.** Yes. In a Dirac or chiral representation, the chiral Lagrangian is complex, like the traditional Dirac Lagrangian (38.1). The imaginary part integrates to a surface term, so has no effect on the electromagnetic equations of motion. REALLY? YES. OTHER PART IS PSEUDOSCALAR.

---

### 38.7.5 Newman-Penrose components of the electromagnetic field

The chiral decomposition of the electromagnetic field  $\mathbf{F} = \frac{1}{2} F^{mn} \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n$  is most transparent in a Newman-Penrose tetrad, where the right- and left-handed basis bivectors are (notation  $\boldsymbol{\gamma}_{mn} \equiv \boldsymbol{\gamma}_m \wedge \boldsymbol{\gamma}_n$ )

$$\begin{aligned} \{ \boldsymbol{\gamma}_{v+}, \frac{1}{2}(\boldsymbol{\gamma}_{vu} - \boldsymbol{\gamma}_{+-}), \boldsymbol{\gamma}_{u-} \} & \text{ right-handed ,} \\ \{ \boldsymbol{\gamma}_{v-}, \frac{1}{2}(\boldsymbol{\gamma}_{vu} + \boldsymbol{\gamma}_{+-}), \boldsymbol{\gamma}_{u+} \} & \text{ left-handed .} \end{aligned} \quad (38.120)$$

The right- and left-handed basis bivectors are conjugates of each other, equation (36.116),

$$C \begin{pmatrix} \boldsymbol{\gamma}_{v+} \\ \frac{1}{2}(\boldsymbol{\gamma}_{vu} - \boldsymbol{\gamma}_{+-}) \\ \boldsymbol{\gamma}_{u-} \end{pmatrix} C^{-1} = \begin{pmatrix} \boldsymbol{\gamma}_{v-} \\ \frac{1}{2}(\boldsymbol{\gamma}_{vu} + \boldsymbol{\gamma}_{+-}) \\ \boldsymbol{\gamma}_{u+} \end{pmatrix} . \quad (38.121)$$

Define the spin components  $E_{\pm}$  and  $B_{\pm}$  of the electric and magnetic fields to be

$$E_{\pm} \equiv E_1 \pm iE_2 , \quad B_{\pm} \equiv B_1 \pm iB_2 , \quad (38.122)$$

satisfying  $E_+^* = E_-$  and  $B_+^* = B_-$ . The coefficients of the right-handed chiral components of the electromagnetic field in a Newman-Penrose basis are

$$F_{v+}^R = F_{v+} = \frac{1}{2}(E_+ + iB_+) , \quad (38.123a)$$

$$F_{vu}^R = -F_{+-}^R = \frac{1}{2}(F_{vu} - F_{+-}) = -\frac{1}{2}(E_3 + iB_3) , \quad (38.123b)$$

$$F_{u-}^R = F_{u-} = \frac{1}{2}(E_- + iB_-) , \quad (38.123c)$$

while the coefficients of the left-handed chiral components are

$$F_{v-}^L = F_{v-} = \frac{1}{2}(E_- - iB_-) , \quad (38.124a)$$

$$F_{vu}^L = F_{+-}^L = \frac{1}{2}(F_{vu} + F_{+-}) = -\frac{1}{2}(E_3 - iB_3) , \quad (38.124b)$$

$$F_{u+}^L = F_{u+} = \frac{1}{2}(E_+ - iB_+) . \quad (38.124c)$$

The Newman-Penrose components are conjugated by taking the complex conjugate and flipping spin indices  $+ \leftrightarrow -$ , §36.7.4. The left-handed components are conjugates of the right-handed components,

$$F_{v+}^{R*} = F_{v-}^L , \quad (38.125a)$$

$$F_{vu}^{R*} = F_{vu}^L , \quad (38.125b)$$

$$F_{u-}^{R*} = F_{u+}^L . \quad (38.125c)$$

In the literature, the Newman-Penrose coefficients (38.123) are commonly denoted (Chandrasekhar, 1983)

$$F_{v+}^R = \phi_0 , \quad (38.126a)$$

$$F_{vu}^R = \phi_1 , \quad (38.126b)$$

$$F_{u-}^R = \phi_2 . \quad (38.126c)$$

The standard notation obscures the boost and spin weight properties, and the bivector structure, of the electromagnetic field. The coefficients  $\phi_n$  are commonly referred to as Newman-Penrose scalars, a confusing designation for quantities that are the components of a bivector, an antisymmetric tensor of rank 2.

### 38.8 A scalar product of derivatives of multivectors of different grade is a total divergence

Let  $\mathbf{a}$  be a multivector of grade  $p+2$ , and  $\mathbf{b}$  be a multivector of grade  $p$ . Then the product  $(\mathbf{D}\mathbf{a})(\mathbf{D}\mathbf{b})$  of their torsion-free covariant spacetime derivatives contains a scalar part,

$$\begin{aligned} \langle(\mathbf{D}\mathbf{a})(\mathbf{D}\mathbf{b})\rangle_0 &= (\mathbf{D} \cdot \mathbf{a}) \cdot (\mathbf{D} \wedge \mathbf{b}) \\ &= (1/p!)(D_k a^{klm\dots n})(D_{[l} b_{m\dots n]}) \\ &= (1/p!) [D_k (a^{klm\dots n} D_{[l} b_{m\dots n]}) - a^{klm\dots n} D_{[k} D_{l} b_{m\dots n]}] \\ &= \mathbf{D} \cdot (\mathbf{a} \cdot (\mathbf{D} \wedge \mathbf{b})) - \mathbf{a} \cdot (\mathbf{D} \wedge \mathbf{D} \wedge \mathbf{b}) \\ &= \mathbf{D} \cdot (\mathbf{a} \cdot (\mathbf{D} \wedge \mathbf{b})) , \end{aligned} \quad (38.127)$$

which is a total covariant divergence. Consequently any potential contribution to a Lagrangian from a product of derivatives of multivectors of different grades yields a total divergence, and therefore has no effect on the equations of motion. In a Lagrangian, only products of derivatives of multivectors of the same grade can affect equations of motion.

## 38.9 Yang-Mills field

### 38.9.1 Yang-Mills gauge symmetries

Electromagnetism is the simplest example of a Yang-Mills field. Electromagnetism postulates that the Lagrangian is unchanged under a local transformation of charged fields by a complex phase. Yang-Mills theory postulates that the Lagrangian is invariant under a larger group of local transformations. After the example of electromagnetism, Yang-Mills transformations are commonly called gauge transformations, and Yang-Mills fields are called gauge fields.

The standard model of physics is based on Yang-Mills theory with the gauge group

$$U(1) \times SU(2) \times SU(3) . \quad (38.128)$$

The group  $U(N)$  is the group of  $N \times N$  unitary complex matrices. A unitary matrix  $U$  is a complex  $N \times N$  matrix satisfying

$$U^\dagger U = 1 , \quad (38.129)$$

where  $U^\dagger \equiv U^{*\top}$  is its Hermitian conjugate. A unitary matrix has a complex determinant of unit modulus. The special unitary group  $SU(N)$  is the group of  $N \times N$  complex unitary matrices with unit determinant. In the standard model, the group  $U(1) \times SU(2)$  describes the unified electroweak group, which at familiar low energies is spontaneously broken to the electromagnetic group  $U(1)_{\text{em}}$ , which is not the same as the  $U(1)$  part of the electroweak group  $U(1) \times SU(2)$ . The group  $SU(3)$  describes the strong, or colour force, which is unbroken at low energies. The group  $U(1)$  is abelian, meaning that the group product is commutative,

$ab = ba$  for any two elements  $a$  and  $b$  of the group. For  $N > 1$ , the groups  $U(N)$  and  $SU(N)$  are non-abelian, meaning that the product is not commutative.

A Yang-Mills group is generated by a set of matrices  $T_a$ , called the generators of the group. Any element of the group can be expressed as the exponential of the generators of the group. The generators of  $SU(2)$  are the three Pauli matrices. The generators of  $SU(3)$  are the eight Gell-Mann matrices. The  $U(N)$  group has  $N^2$  generators, while the  $SU(N)$  group has  $N^2 - 1$  generators. The generators of a unitary group are necessarily Hermitian, and the generators of a special unitary group are in addition traceless.

Yang-Mills theory postulates that the Lagrangian is unchanged under a gauge transformation that changes spinor fields  $\psi$  that feel the force by

$$\psi \rightarrow e^{-ig\theta^a T_a} \psi, \quad \bar{\psi} \rightarrow e^{ig\theta^a T_a} \bar{\psi}, \quad (38.130)$$

where  $T_a$  are the generators of the gauge group, and  $\theta^a(x)$  are arbitrary functions of spacetime. For the transformation (38.130) to make sense, the field  $\psi$  must be a column vector acted on by the matrices  $T_a$ . The coupling parameter  $g$  is analogous to the electric charge  $e$  of the electromagnetic field. The transformation law (38.130) shows that if the field  $\psi$  has Yang-Mills charge  $g$ , then the  $C$ -conjugate field  $\bar{\psi}$  has charge  $-g$ . The coupling parameter  $g$  is dimensionless, and is a fundamental property of the Yang-Mills field.

To ensure that the Lagrangian remains invariant under the gauge transformation, the derivative  $\mathbf{D}$  must be replaced by a gauge-covariant derivative  $\mathbf{D} + ig\mathbf{A}^a T_a$  which, when acting on the spinor field  $\psi$ , transforms under the gauge transformation (38.130) as

$$(\mathbf{D} + ig\mathbf{A}^a T_a)\psi \rightarrow e^{-ig\theta^a T_a} (\mathbf{D} + ig\mathbf{A}^a T_a)\psi. \quad (38.131)$$

The derivative  $\mathbf{D}$  must be interpreted as a unit matrix with respect to the Yang-Mills group. The potential  $\mathbf{A}^a T_a = \boldsymbol{\gamma}^m A_m^a T_a$  transforms as a vector under Lorentz transformations, but is also a matrix that transforms under gauge transformations of the Yang-Mills group. The gauge-covariant derivative transforms correctly provided that the gauge field  $\mathbf{A}^a T_a$  transforms under the Yang-Mills gauge transformation (38.130) as DO INFINITESIMAL VERSION? SIGN?

$$\mathbf{A}^a T_a \rightarrow e^{ig\theta^a T_a} \mathbf{A}^a T_a e^{-ig\theta^a T_a} + \mathbf{D}\theta^a T_a. \quad (38.132)$$

It is often convenient to absorb the Yang-Mills indices of the gauge field  $\mathbf{A}^a T_a$  into abbreviated symbols

$$\mathbf{A} \equiv \mathbf{A}^a T_a, \quad A_m \equiv A_m^a T_a. \quad (38.133)$$

### 38.9.2 Properties of the generators of a Yang-Mills group

The commutators of the Yang-Mills generators  $T_a$  are themselves generators. The commutators define the structure coefficients  $f_{abc}$  of the group,

$$[T_a, T_b] = i f_{abc} T_c. \quad (38.134)$$

The structure coefficients  $f_{abc}$  are manifestly antisymmetric in  $ab$ . It can be proven that if the group is finite dimensional and compact (for example, the 3D rotation group is compact, but the Lorentz group, which

involves boosts, is not compact), then a basis of generators can always be found such that the structure coefficients are totally antisymmetric,

$$f_{abc} = f_{[abc]} . \quad (38.135)$$

For example, the three generators of the  $SU(2)$  electroweak group are conventionally half the Pauli matrices, equation (13.97),  $T_a = \frac{1}{2}\sigma_a$ , which satisfy

$$\left[ \frac{\sigma_a}{2}, \frac{\sigma_b}{2} \right] = i\varepsilon_{abc} \frac{\sigma_c}{2} , \quad (38.136)$$

with  $\varepsilon_{abc} \equiv [abc]$  the totally antisymmetric symbol. Similarly the eight Gell-Mann generators for the  $SU(3)$  colour group have, by construction, totally antisymmetric structure coefficients (though with various different numeric values, unlike Pauli matrices).

The Yang-Mills generators also possess a scalar product, as is essential for a scalar Yang-Mills Lagrangian to be defined. Again, it can be proven that if the group is finite dimensional and compact, then by a suitable unitary transformation, generators with totally antisymmetric structure coefficients can be found such that the scalar product defined by the trace of their products,

$$T_a \cdot T_b \equiv \text{Tr}(T_a T_b) , \quad (38.137)$$

is proportional to the Euclidean metric

$$T_a \cdot T_b \propto \delta_{ab} . \quad (38.138)$$

Again, the assumption of compactness is essential for the metric to be Euclidean; the Lorentz group is not compact, and its metric is not Euclidean. For example, the three generators of the  $SU(2)$  electroweak group are half the Pauli matrices,  $T_a = \frac{1}{2}\sigma_a$ , which satisfy

$$\frac{\sigma_a}{2} \cdot \frac{\sigma_b}{2} \equiv \text{Tr} \left( \frac{\sigma_a}{2} \frac{\sigma_b}{2} \right) = \frac{1}{2} \delta_{ab} . \quad (38.139)$$

The factor on the right hand side is  $\frac{1}{2}$  not  $\frac{1}{4}$  because the self-product of each Pauli matrix is the unit  $2 \times 2$  matrix, whose trace is 2. The eight Gell-Mann matrices  $\sigma_a$  of the  $SU(3)$  colour group satisfy the same relations (38.139).

In what follows, the Yang-Mills structure coefficients will be taken to be antisymmetric, equation (38.135), and the Yang-Mills metric Euclidean, equation (38.138) (proportionality to the Euclidean metric suffices for all equations below).

### 38.9.3 Yang-Mills covariant derivative

Suppose that  $B = B^a T_a$  is a linear combination of Yang-Mills generators  $T_a$ . The matrix  $B$  acts on the spinor  $\psi$ . The matrix  $B$  is gauge-covariant if  $B\psi$  transforms like (38.130) under Yang-Mills gauge transformations. To preserve gauge covariance, the matrix  $B$  must transform under a gauge transformation (38.130) as

$$B \rightarrow e^{-ig\theta^a T_a} B e^{ig\theta^a T_a} . \quad (38.140)$$

The Yang-Mills gauge-covariant derivative of  $B\psi$  is

$$(D_m + igA_m)B\psi = B(D_m\psi + igA_m\psi) + (D_mB + ig[A_m, B])\psi . \quad (38.141)$$

To ensure the Leibniz rule, the Yang-Mills gauge-covariant derivative of  $B$ , conveniently denoted by a script  $\mathcal{D}_m$ , must therefore be

$$\mathcal{D}_m B \equiv D_mB + ig[A_m, B] , \quad (38.142)$$

which resembles the behaviour (15.15) of the covariant derivative in general relativity. The Yang-Mills gauge-covariant spacetime derivative  $\mathbf{D} \equiv \boldsymbol{\gamma}^m \mathcal{D}_m$  is

$$\mathbf{D}B \equiv DB + ig[\mathbf{A}, B] . \quad (38.143)$$

More explicitly, with the generators  $T_a$  restored, the Yang-Mills gauge-covariant derivative of  $B = B^a T_a$  is

$$\mathcal{D}_m B = D_mB^a T_a + ig[A_m^b T_b, B^c T_c] , \quad (38.144)$$

which involves the commutators  $[T_b, T_c]$  of the generators. In terms of the structure coefficients  $f_{abc}$ , the gauge-covariant derivative is

$$\mathcal{D}_m B = (D_mB^a - gf_{bca}[A_m^b, B^c]) T_a . \quad (38.145)$$

### 38.9.4 Yang-Mills field

FIX

The Yang-Mills field  $\mathbf{F}$  is defined in terms of the commutator of the gauge-covariant derivative by

$$\mathbf{F} \equiv -(ig)^{-1}[\mathbf{D} + ig\mathbf{A}, \mathbf{D} + ig\mathbf{A}] = -\mathbf{D} \wedge \mathbf{A} - ig\mathbf{A} \wedge \mathbf{A} . \quad (38.146)$$

ACTING ON SOMETHING? The components of the Yang-Mills field  $\mathbf{F} = \frac{1}{2}F_{mn}^a \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n T_a$  are the real quantities

$$F_{mn}^a = -D_m A_n^a + D_n A_m^a + gf_{bca} A_m^b A_n^c . \quad (38.147)$$

The field  $F_{mn}^a$  is antisymmetric in  $mn$ . The antisymmetry of the structure coefficients  $f_{bca}$  ensures that the product  $A_m^b A_n^c$  in last term on the right hand side of equation (38.147) is antisymmetric in both the spacetime indices  $mn$  and the Yang-Mills indices  $bc$ . The potential is a grade-1 multivector  $\mathbf{A} = A_m \boldsymbol{\gamma}^m$ . The commutator  $[\mathbf{A}, \mathbf{A}]$  can thus also be written as a wedge product

$$[\mathbf{A}, \mathbf{A}] = \frac{1}{2}[A_m, A_n] \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n = A_m A_n \boldsymbol{\gamma}^m \wedge \boldsymbol{\gamma}^n = \mathbf{A} \wedge \mathbf{A} . \quad (38.148)$$

The factor of  $\frac{1}{2}$  in the second expression of (38.148) cancels the double-counting of indices  $mn$ . REALLY? There is no factor of  $\frac{1}{2}$  in the relation between commutator and the wedge product. The commutator  $\mathbf{A} \wedge \mathbf{A}$  does not vanish for a non-abelian Yang-Mills field.

The Yang-Mills field  $\mathbf{F}$  is gauge-covariant but not gauge-invariant under Yang-Mills transformations, since it is a sum  $\mathbf{F} = \mathbf{F}^a T_a$  of Yang-Mills generators  $T_a$ , and thus varies under Yang-Mills transformations

as (38.140). However, the scalar product of the Yang-Mills field with itself is gauge-invariant, and provides a suitable Lagrangian.

The Yang-Mills potential  $\mathbf{A}$  is the connection of the Yang-Mills group, analogous to the connection  $\mathbf{\Gamma}$  of general relativity. The Yang-Mills potential  $\mathbf{A}$  transforms under Yang-Mills transformations as (38.132), not as (38.140). Thus  $\mathbf{A}$  is not, to use the language of general relativity, a Yang-Mills tensor, just as the connection  $\mathbf{\Gamma}$  is not a general relativistic tensor.

In particular, the gauge-covariant derivative of the potential  $\mathbf{A}$

$$\mathcal{D}\mathbf{A} = \mathbf{D}\mathbf{A} + ig[\mathbf{A}, \mathbf{A}] \quad (38.149)$$

is a Yang-Mills tensor. The gauge-covariant derivative  $\mathcal{D}\mathbf{A}$  is gauge-covariant but not gauge-invariant. The Yang-Mills Lagrangian is required to be a gauge-invariant scalar. The covariant spacetime derivative  $\mathbf{D}\mathbf{A}$  is a sum of a divergence and a curl, equation (15.37). No scalar can be constructed from the divergence, which can be made equal identically to zero by a suitable Yang-Mills gauge transformation. The curl is by itself not a Yang-Mills scalar, but a gauge-invariant scalar can be constructed from it. The gauge-covariant curl of the Yang-Mills potential  $\mathbf{A}$  defines the Yang-Mills field  $\mathbf{F}$ ,

$$\mathbf{F} \equiv -\mathcal{D} \wedge \mathbf{A} = -\mathbf{D} \wedge \mathbf{A} - ig[\mathbf{A}, \mathbf{A}] . \quad (38.150)$$

### 38.9.5 Yang-Mills Bianchi identities

$$\mathcal{D} \wedge \mathbf{F} = 0 ?? \quad (38.151)$$

### 38.9.6 Yang-Mills Lagrangian

In terms of the Yang-Mills field  $\mathbf{F}$ , the Yang-Mills Lagrangian is the gauge-invariant scalar

$$L = -\frac{1}{2}\mathbf{F} \cdot \mathbf{F} . \quad (38.152)$$

The factor  $1/2$  in the Yang-Mills Lagrangian (38.152) holds for SI units, as commonly used in high-energy physics. The factor would be  $1/(8\pi)$  in Gaussian units. The scalar product in the Yang-Mills Lagrangian (38.152) is with respect to Yang-Mills as well as tetrad components.

Written, as required, as a function  $L(\mathbf{A}, \mathbf{D} \wedge \mathbf{A})$  of the potential  $\mathbf{A}$  and its velocity  $\mathbf{D} \wedge \mathbf{A}$ , the Yang-Mills Lagrangian (38.152) is

$$L = -\frac{1}{2}(\mathbf{D} \wedge \mathbf{A} + ig\mathbf{A} \wedge \mathbf{A}) \cdot (\mathbf{D} \wedge \mathbf{A} + ig\mathbf{A} \wedge \mathbf{A}) . \quad (38.153)$$

The Yang-Mills Lagrangian (38.153) is real despite the seeming presence of the imaginary  $i$ , because  $\mathbf{A} \wedge \mathbf{A}$  is imaginary.

The momentum canonically conjugate to the Yang-Mills potential  $\mathbf{A}$  is the Yang-Mills field  $\mathbf{F}$ ,

$$\frac{\partial L}{\partial(\mathbf{D} \wedge \mathbf{A})} = \mathbf{F} . \quad (38.154)$$

Variation of the Lagrangian with respect to  $\mathbf{A}$  gives

$$\delta L = -ig \delta(\mathbf{A} \wedge \mathbf{A}) \cdot \mathbf{F} = -ig [T_a, T_b] \cdot T_c (\delta \mathbf{A}^a) \cdot (\mathbf{A}^b \cdot \mathbf{F}^c) = g (\delta \mathbf{A}) \cdot (f_{abc} T_a (\mathbf{A}^b \cdot \mathbf{F}^c)) = -ig (\delta \mathbf{A}) \cdot [\mathbf{A}, \mathbf{F}] . \quad (38.155)$$

The penultimate step of equations (38.155) holds because the Yang-Mills metric is Euclidean, equation (38.138), while the final step follows because  $f_{abc}$  is totally antisymmetric, equation (38.135). The Euler-Lagrange equation of motion for the Yang-Mills field is

$$\mathbf{D} \cdot \mathbf{F} = \frac{\partial L}{\partial \mathbf{A}} = -ig[\mathbf{A}, \mathbf{F}] , \quad (38.156)$$

which rearranges to

$$\mathbf{D} \cdot \mathbf{F} + ig[\mathbf{A}, \mathbf{F}] = 0 . \quad (38.157)$$

The dot product  $\mathbf{D} \cdot \mathbf{F}$  on the left hand side of equation (38.157) signifies that the lowest grade component of the multivector product is to be taken, which here is the vector (grade 1) part. The left hand side of the Euler-Lagrange equation of motion (38.157) can be recognized as the Yang-Mills gauge covariant derivative of the field  $\mathbf{F}$ , so the equation of motion reduces to

$$\mathbf{D} \cdot \mathbf{F} = 0 . \quad (38.158)$$

In component form, the Yang-Mills Euler-Lagrange equation of motion (38.158) is

$$D^m F_{mn}^a - g f_{abc} A^{bm} F_{mn}^c = 0 . \quad (38.159)$$

### 38.9.7 Yang-Mills super-Hamiltonian

The Yang-Mills super-Hamiltonian  $H(\mathbf{A}, \mathbf{F})$  is defined to be

$$H \equiv \mathbf{F} \cdot (\mathbf{D} \wedge \mathbf{A}) - L , \quad (38.160)$$

which expressed as required in terms of the “field”  $\mathbf{A}$  and the momentum  $\mathbf{F}$  is

$$H = -\frac{1}{2} \mathbf{F} \cdot \mathbf{F} - ig(\mathbf{A} \wedge \mathbf{A}) \cdot \mathbf{F} . \quad (38.161)$$

Hamilton’s equations are

$$\mathbf{D} \cdot \mathbf{F} = -\frac{\partial H}{\partial \mathbf{A}} = -ig[\mathbf{A}, \mathbf{F}] , \quad \mathbf{D} \wedge \mathbf{A} = \frac{\partial H}{\partial \mathbf{F}} = -\mathbf{F} - ig\mathbf{A} \wedge \mathbf{A} , \quad (38.162)$$

reproducing the Euler-Lagrange equations (38.157) and the relation (38.150) between the field  $\mathbf{F}$  and the derivative  $\mathbf{D} \wedge \mathbf{A}$ .

### 38.9.8 Simple harmonic Yang-Mills super-Hamiltonian

The non-abelian Yang-Mills super-Hamiltonian (38.160) is not quadratic in  $\mathbf{A}$  and its momentum  $\mathbf{F}$ , and so does not have the structure of a simple harmonic oscillator. The super-Hamiltonian can be made to look like a simple harmonic oscillator by regarding the velocity of  $\mathbf{A}$  not as the spacetime curl  $\mathcal{D} \wedge \mathbf{A}$ , but rather as the gauge-covariant spacetime curl  $\mathcal{D} \wedge \mathbf{A}$ , which is covariant with respect to both Yang-Mills and general relativistic (coordinate and tetrad) transformations. The momentum conjugate to  $\mathbf{A}$  is still the field  $\mathbf{F}$ , but the super-Hamiltonian defined by

$$H \equiv \mathbf{F} \cdot (\mathcal{D} \wedge \mathbf{A}) - L , \quad (38.163)$$

is quadratic in the field  $\mathbf{F}$ , like that (38.97) of electromagnetism,

$$H = -\frac{1}{2} \mathbf{F} \cdot \mathbf{F} . \quad (38.164)$$

Hamilton's equations are

$$\mathcal{D} \cdot \mathbf{F} = -\frac{\partial H}{\partial \mathbf{A}} = 0 , \quad \mathcal{D} \wedge \mathbf{A} = \frac{\partial H}{\partial \mathbf{F}} = -\mathbf{F} , \quad (38.165)$$

which reproduce the earlier equations. Equations (38.164) and (38.165) constitute an elegant expression of the dynamics of a Yang-Mills field.

## 38.10 Real scalar field

With the possible exception of the electroweak Higgs field, no fundamental scalar field has yet been observed in nature. However, scalar fields are predicted by theoretical physicists and cosmologists to play a central role in the Universe. The Higgs field is postulated to be a scalar field. The field that drives inflation in cosmology is postulated to be a scalar field. Supersymmetry predicts that fundamental spin- $\frac{1}{2}$  fields should have fundamental spin-0 superpartners. The cold dark matter that dominates matter in the Universe could perhaps be a scalar particle.

A scalar field  $\varphi$  has spin 0, so it is invariant under Lorentz transformations. One might pity a scalar field, because how can it know about the structure of spacetime if it has no sense of direction? The solution to this conundrum is that a scalar field has a vector derivative  $\partial\varphi$ , which gives it a sense of direction.

### 38.10.1 Real scalar Lagrangian

For a scalar field  $\varphi$ , the torsion-free covariant derivative of the field is the same as its directed derivative,  $\mathcal{D}\varphi = \partial\varphi$ , or in components  $D_n\varphi = \partial_n\varphi$ . The general relativistic Lagrangian  $L(\varphi, \partial\varphi)$  of a real scalar field  $\varphi$  of mass  $m$  is

$$L = -\frac{1}{2} [(\partial\varphi) \cdot (\partial\varphi) + m^2\varphi^2] . \quad (38.166)$$

In index notation, the scalar Lagrangian (38.166) is

$$L = -\frac{1}{2} [\gamma^{mn}(\partial_m\varphi)(\partial_n\varphi) + m^2\varphi^2] , \quad (38.167)$$

where  $\gamma_{mn}$  is the tetrad metric.

The momentum  $\boldsymbol{\pi}$  conjugate to the field  $\varphi$  is defined to be the derivative of the Lagrangian with respect to the velocity  $\partial\varphi$  of the field,

$$\boldsymbol{\pi} \equiv \frac{\partial L}{\partial(\partial\varphi)} = -\partial\varphi , \quad (38.168)$$

the derivative being defined by (38.88). Whereas the field  $\varphi$  is a scalar, its conjugate momentum  $\boldsymbol{\pi}$  is a vector in the spacetime algebra.

The Euler-Lagrange equations of motion are

$$\mathbf{D} \cdot \boldsymbol{\pi} = \frac{\partial L}{\partial\varphi} = -m^2\varphi , \quad (38.169)$$

which, when combined with (38.168), gives the Klein-Gordon wave equation

$$(\mathbf{D} \cdot \mathbf{D} - m^2)\varphi = 0 . \quad (38.170)$$

The Klein-Gordon equation (38.170) can also be written

$$(\square - m^2)\varphi = 0 , \quad (38.171)$$

where  $\square \equiv \mathbf{D} \cdot \mathbf{D} = D^n D_n$  is the d'Alembertian.

### 38.10.2 Real scalar super-Hamiltonian

The scalar super-Hamiltonian  $H(\varphi, \boldsymbol{\pi})$  is defined by

$$H \equiv \boldsymbol{\pi} \cdot \partial\varphi - L . \quad (38.172)$$

The resulting Hamiltonian looks like a simple harmonic oscillator,

$$H = \frac{1}{2}(-\boldsymbol{\pi} \cdot \boldsymbol{\pi} + m^2\varphi^2) . \quad (38.173)$$

Hamilton's equations are obtained by varying the action  $S$

$$S = \int (\boldsymbol{\pi} \cdot \partial\varphi - H) d^4x \quad (38.174)$$

with respect to the field  $\varphi$  and its conjugate momentum  $\boldsymbol{\pi}$ . The variation of the first term in the integrand of the action (38.174) is

$$\delta(\boldsymbol{\pi} \cdot \partial\varphi) = \boldsymbol{\pi} \cdot \partial(\delta\varphi) + (\delta\boldsymbol{\pi}) \cdot \partial\varphi = D_n(\pi^n \delta\varphi) - (\mathbf{D} \cdot \boldsymbol{\pi}) \delta\varphi + (\delta\boldsymbol{\pi}) \cdot \partial\varphi . \quad (38.175)$$

Whereas the field  $\varphi$  is scalar and therefore its covariant derivative equals its ordinary derivative,  $\mathbf{D}\varphi = \partial\varphi$ , the momentum  $\boldsymbol{\pi}$  is a vector and consequently its general relativistic covariant divergence  $\mathbf{D} \cdot \boldsymbol{\pi}$  includes a connection, equation (15.15). Varying the action (38.174) with respect to the field  $\varphi$  and its conjugate momentum  $\boldsymbol{\pi}$  gives

$$\delta S = \oint \pi^n \delta\varphi d^3x_n + \int \left[ -\left( \mathbf{D} \cdot \boldsymbol{\pi} + \frac{\partial H}{\partial\varphi} \right) \delta\varphi + \delta\boldsymbol{\pi} \cdot \left( \partial\varphi - \frac{\partial H}{\partial\boldsymbol{\pi}} \right) \right] d^4x . \quad (38.176)$$

Hamilton's equations are thus

$$\mathbf{D} \cdot \boldsymbol{\pi} = -\frac{\partial H}{\partial \varphi} = -m^2 \varphi, \quad \partial \varphi = \frac{\partial H}{\partial \boldsymbol{\pi}} = -\boldsymbol{\pi}. \quad (38.177)$$

Combining equations (38.177) recovers the Klein-Gordon wave equation (38.170).

## 38.11 Complex scalar field

### 38.11.1 Complex scalar Lagrangian

A complex scalar field  $\varphi$  is equivalent to two real scalar fields  $\varphi_R$  and  $\varphi_I$  of the same mass,  $\varphi = \frac{1}{\sqrt{2}}(\varphi_R + i\varphi_I)$ , except that the complex field has an additional symmetry of rotation of the fields by a complex phase. It is most elegant to treat the field  $\varphi$  and its complex conjugate  $\varphi^*$  as two independent fields, to be varied independently in applying the least action principle, as opposed to treating the real and imaginary parts as the two independent components. The Lagrangian  $L(\varphi, \varphi^*, \partial\varphi, \partial\varphi^*)$  of a massive complex scalar field  $\varphi$  is the real scalar

$$L = -[(\partial\varphi^*) \cdot (\partial\varphi) + m^2 \varphi^* \varphi]. \quad (38.178)$$

The momenta  $\boldsymbol{\pi}$  and  $\boldsymbol{\pi}^*$  conjugate to the fields  $\varphi$  and  $\varphi^*$  are

$$\boldsymbol{\pi} \equiv \frac{\partial L}{\partial(\partial\varphi)} = -\partial\varphi^*, \quad \boldsymbol{\pi}^* \equiv \frac{\partial L}{\partial(\partial\varphi^*)} = -\partial\varphi. \quad (38.179)$$

Notice that the momentum of the complex conjugate field is the complex conjugate of the momentum of the field, which is true because the Lagrangian is real.

### 38.11.2 Complex scalar super-Hamiltonian

The super-Hamiltonian  $H(\varphi, \varphi^*, \boldsymbol{\pi}, \boldsymbol{\pi}^*)$  of the massive complex scalar field is defined by

$$H \equiv \boldsymbol{\pi} \cdot \partial\varphi + \boldsymbol{\pi}^* \cdot \partial\varphi^* - L, \quad (38.180)$$

giving

$$H = -\boldsymbol{\pi}^* \cdot \boldsymbol{\pi} + m^2 \varphi^* \varphi. \quad (38.181)$$

Hamilton's equations are

$$\mathbf{D} \cdot \boldsymbol{\pi} = -\frac{\partial H}{\partial \varphi} = -m^2 \varphi^*, \quad \partial \varphi = \frac{\partial H}{\partial \boldsymbol{\pi}} = -\boldsymbol{\pi}^*, \quad (38.182)$$

and their complex conjugates. Combining equations (38.182) yields the Klein-Gordon wave equation

$$(\mathbf{D} \cdot \mathbf{D} - m^2)\varphi = 0, \quad (38.183)$$

and its complex conjugate.

### 38.11.3 Current conservation for a complex scalar field

The Lagrangian of a complex scalar field is invariant under the global transformation

$$\varphi \rightarrow e^{-i\epsilon} \varphi, \quad \varphi^* \rightarrow e^{i\epsilon} \varphi^*. \quad (38.184)$$

For infinitesimal  $\epsilon$ , the transformation (38.184) takes the form (16.14) with

$$\delta\varphi = -i\varphi, \quad \delta\varphi^* = i\varphi^*. \quad (38.185)$$

The Noether current  $j^n$  defined by

$$j^n \equiv \frac{i}{2} (\varphi \partial^n \varphi^* - \varphi^* \partial^n \varphi) \quad (38.186)$$

is then covariantly conserved, equation (16.18),

$$D_n j^n = 0. \quad (38.187)$$

For a plane wave  $\varphi \propto e^{ip_n x^n}$ , the current equals the energy-momentum density of the wave,

$$j^n = |\varphi|^2 p^n. \quad (38.188)$$

### 38.11.4 Chiral decomposition of a complex scalar field

Let the field  $\varphi$  be a sum of scalar and pseudoscalar fields  $A$  and  $IB$ ,

$$\varphi = A + IB, \quad (38.189)$$

where  $A$  and  $B$  could be either real or complex. If the fields  $A$  and  $B$  have the same mass  $m$ , their Lagrangian is

$$L = - [(\partial A^*) \cdot (\partial A) + m^2 A^* A] - [(\partial B^*) \cdot (\partial B) + m^2 B^* B]. \quad (38.190)$$

The chiral components of the scalar/pseudoscalar field  $\varphi$  are

$$\varphi_R \equiv \frac{1}{2}(1 \pm \gamma_5)(A + IB) = \frac{1}{2}(1 \pm \gamma_5)(A \pm iB). \quad (38.191)$$

The complex conjugates of the chiral components are, in a Dirac or chiral representation where  $\gamma_5$  is real and  $I \equiv i\gamma_5$  is imaginary,

$$\varphi_R^* = \frac{1}{2}(1 \pm \gamma_5)(A^* - IB^*) = \frac{1}{2}(1 \pm \gamma_5)(A^* \mp iB^*). \quad (38.192)$$

Since products of opposite chiral components vanish,  $\varphi_R \varphi_L = 0$ , the Lagrangian in terms of the chiral components is

$$L = - [(\partial \varphi_R^*) \cdot (\partial \varphi_R) + m^2 \varphi_R^* \varphi_R] - [(\partial \varphi_L^*) \cdot (\partial \varphi_L) + m^2 \varphi_L^* \varphi_L]. \quad (38.193)$$

### 38.11.5 Conjugate scalar field

The conjugate  $\bar{\varphi}$  scalar field is defined analogously to the conjugate Dirac spinor (36.95) and conjugate electromagnetic field (??) by

$$\bar{\varphi} = A^* - IB^* , \quad (38.194)$$

which complex conjugates not only the coefficients  $A$  and  $B$  but also the basis scalar 1 and pseudoscalar  $I$ . Conjugation flips the chirality of the scalar field. NO IT DOES NOT. Hence

$$\varphi_R^* = \overline{\varphi_L} . \quad (38.195)$$

Lagrangian is

$$L = - [(\partial\overline{\varphi_L}) \cdot (\partial\varphi_R) + m^2 \overline{\varphi_L}\varphi_R] - [(\partial\overline{\varphi_R}) \cdot (\partial\varphi_L) + m^2 \overline{\varphi_R}\varphi_L] . \quad (38.196)$$

### 38.12 Conformally coupled scalar field

Consider the conformally rescaled metric

$$g_{\mu\nu} = a^2(x) g_{\mu\nu}^{(c)} , \quad (38.197)$$

where the conformal factor  $a(x)$  is some arbitrary function of spacetime. The null geodesic structure of spacetime is unchanged by such a conformal rescaling, since the null condition  $ds = 0$  is unchanged by rescaling.

The d'Alembertian of a scalar field  $\varphi$  in the conformally rescaled frame is related to that in the original frame by

$$(\square - \frac{1}{6}R) \varphi = a^{-3} \square^{(c)} a \varphi , \quad (38.198)$$

where  $R$  is the Ricci scalar. Equation (38.198) is most easily proved by taking the original frame to be Minkowski,  $g_{\mu\nu}^{(c)} = \eta_{\mu\nu}$ , and invoking general covariance to conclude that the equation is true in general.

A conformally coupled scalar field is one satisfying

$$(\square - \frac{1}{6}R - m^2) \varphi = 0 . \quad (38.199)$$

In particular, a conformally coupled massless scalar field satisfies

$$(\square - \frac{1}{6}R) \varphi = 0 . \quad (38.200)$$

In the presence of a cosmological constant  $\Lambda$ , for example, the Ricci scalar is the constant  $R = \Lambda/4$ . A positive Ricci scalar  $R$  would give the scalar field a positive effective mass, while a negative Ricci scalar  $R$  would give the scalar field an imaginary effective mass. The imaginary mass does not lead to superluminal propagation; rather, it makes the scalar field unstable. For example, the inflaton field that drives inflation has an imaginary effective mass, making it unstable to decay.

A massless scalar field remains massless in the presence of a finite Ricci scalar only if the field is conformally

coupled. For this reason it is sometimes argued that conformal coupling, as opposed to the minimal coupling of the Klein-Gordon equation, is most natural for a massless scalar field.

### 38.13 Spin- $\frac{3}{2}$ field

#### 38.13.1 Spin- $\frac{3}{2}$ particles

No fundamental spin- $\frac{3}{2}$  field is known. But if supersymmetry is a local symmetry of nature, then the connection field of local supersymmetry, a particle called the gravitino, must be a spin- $\frac{3}{2}$  field  $\psi^{ka}$  with both vector  $k$  and spinor  $a$  indices. The spin- $\frac{3}{2}$  field  $\psi$  may be written variously, with indices either explicit or suppressed,

$$\psi = \psi^{ka} \gamma_k \otimes \gamma_a = \psi^a \otimes \gamma_a = \gamma_k \otimes \psi^k . \quad (38.201)$$

The product  $\gamma_k \otimes \gamma_a$  is a tensor product of vector and spinor basis elements, not a multivector product. Vector indices  $k$  are lowered and raised with the tetrad tensor  $\gamma_{kl}$  and its inverse, while spinor indices  $a$  are raised and lowered with the spinor metric  $\varepsilon_{ab}$  and its inverse.

Contraction of  $\psi$  with a vector (grade-1 multivector)  $a$  yields a spin- $\frac{1}{2}$  spinor,

$$a \cdot \psi = a_k \psi^k = a_k \psi^{ka} \gamma_a , \quad (38.202)$$

In equation (38.202), the scalar product is to be interpreted as contracting the vector  $a$  with the vector components of  $\psi$ . To ensure that the gravitino  $\psi$  is pure spin- $\frac{3}{2}$ , all such scalar products must vanish. If that is not already true, the spin- $\frac{3}{2}$  components of  $\psi$  can be projected out by subtracting the spin- $\frac{1}{2}$  part,

$$\psi \rightarrow \psi - 1 \otimes (\gamma_k \psi^k) = \psi^{ka} (\gamma_k \otimes \gamma_a - 1 \otimes \gamma_k \gamma_a) , \quad (38.203)$$

where 1 is the multivector scalar (a unit matrix). The vector spinor  $\psi$  started with  $4 \times 4 = 16$  complex components, but removing the spin- $\frac{1}{2}$  part eliminates 4 complex components, leaving the spin- $\frac{3}{2}$  gravitino  $\psi$  with 12 complex components.

The gravitino  $\psi$  has right- and left-handed chiral components

$$\psi = \psi_R + \psi_L , \quad \psi_R = \psi^a \otimes \frac{1}{2}(1 \pm \gamma_5) \gamma_a . \quad (38.204)$$

#### 38.13.2 Spin- $\frac{3}{2}$ particles in a spinor basis

The structure of the gravitino is more transparent when expressed with respect to a spinor basis. The gravitino  $\psi$  is a 3-tensor of spinors,

$$\psi = \psi^{abc} \gamma_a \otimes \gamma_b \otimes \gamma_c . \quad (38.205)$$

Flipping the indices changes the sign in accordance with equation (36.58),

$$\psi = -\psi_{abc} \gamma^a \otimes \gamma^b \otimes \gamma^c . \quad (38.206)$$

The first two indices  $ab$  may be taken without loss of generality to represent the vector part of the vector-spinor  $\psi$ , while the last index  $c$  represents the spinor part (in a moment the spinor 3-tensor will be found to be totally symmetric, so the ordering of indices is actually irrelevant). According to the correspondence (??) between vectors and outer products of spinors, the indices  $a$  and  $b$  must have opposite chirality, and  $\psi^{abc}$  must be symmetric in  $ab$  TRUE ONLY FOR DIRAC. The third index  $c$  must have the same chirality as one of the two indices  $a$  or  $b$ , since there are only two chiralities, right R and left L. Without loss of generality, suppose that  $c$  has the same chirality as  $b$ . The spinor tensor  $\psi^{abc}$  can be decomposed into parts that are respectively symmetric and antisymmetric in  $bc$ . Since the spinor metric  $\varepsilon_{bc}$  is antisymmetric, and non-zero only for  $b$  and  $c$  of the same chirality, contraction over the antisymmetric part  $[bc]$  yields a spin- $\frac{1}{2}$  spinor,

$$\varepsilon_{bc}\psi^{abc} = \psi^{ab}{}_b . \quad (38.207)$$

In fact if the spinor tensor  $\psi^{abc}$  had any part that was antisymmetric in any two indices, then that part would, according to the super spacetime algebra relations (36.62)–(??), correspond to either a scalar, a pseudoscalar, or a pseudovector, all of which are excluded by the requirement that  $\psi$  be a vector of spinors. Thus the spin- $\frac{3}{2}$  field  $\psi$  is a spinor tensor that is symmetric in all its 3 indices

$$\psi^{abc} = \psi^{(abc)} , \quad (38.208)$$

and one of whose indices has opposite chirality to the other two.

The right- and left-handed chiralities of the gravitino  $\psi^{abc}$  are embodied respectively in its LRR and RLL components. Each chirality has  $2 \times 3 = 6$  complex components, so the spin- $\frac{3}{2}$  gravitino  $\psi$  has a total of 12 distinct complex components, in agreement with the earlier vector-spinor accounting.

### 38.13.3 Spin- $\frac{3}{2}$ field

The covariant spacetime derivative  $D\psi$  of the gravitino decomposes into a divergence and a curl,

$$D\psi = D \cdot \psi + D \wedge \psi . \quad (38.209)$$

The divergence is spin- $\frac{1}{2}$ , while the curl is spin- $\frac{3}{2}$ . In a spinor basis, the covariant spacetime derivative  $D = \gamma_k D^k$  is the multivector

$$D = (\gamma_a \gamma_b \cdot) D^{ab} . \quad (38.210)$$

Because  $D$  is a vector (a grade-1 multivector), the indices  $a$  and  $b$  must have opposite chirality, and  $D^{ab}$  must be symmetric in  $ab$ , in accordance with the correspondence (??). In a spinor basis, the covariant spacetime spin- $\frac{1}{2}$  divergence and spin- $\frac{3}{2}$  curl are

$$D \cdot \psi = D_{ab}\psi^{abc} \gamma_c , \quad D \wedge \psi = -D_d^a \psi^{dbc} \gamma_a \otimes \gamma_b \otimes \gamma_c . \quad (38.211)$$

There is no spin- $\frac{5}{2}$  contribution because the spacetime covariant derivative  $D$  is a multivector: the trailing dot on  $\gamma_a \gamma_b \cdot$  in equation (38.210), symbolic of the spinor metric tensor  $\varepsilon$ , forces at least one contraction over spinor indices.

The spin- $\frac{3}{2}$  gravitino field  $\Psi$  is defined to be minus the covariant spacetime curl of  $\psi$ ,

$$\Psi \equiv -\mathbf{D} \wedge \psi . \quad (38.212)$$

In spinor index notation,

$$\Psi^{abc} = D_d^{(a} \psi^{dbc)} . \quad (38.213)$$

Like the gravitino spinor tensor  $\psi^{abc}$ , the spinor tensor  $\Psi^{abc}$  is totally symmetric in all its indices,

$$\Psi^{abc} = \Psi^{(abc)} . \quad (38.214)$$

Whereas the spin- $\frac{3}{2}$  gravitino  $\psi$  has only vector-spinor components, the spin- $\frac{3}{2}$  field  $\Psi$  has both vector-spinor  $\underline{\Psi}$  and bivector-spinor  $\underline{\Psi}$  components,

$$\Psi = \underline{\Psi} + \underline{\Psi} . \quad (38.215)$$

One of the spinor indices of the vector-spinor  $\underline{\Psi}$  has opposite chirality to the other two, LRR or RLL, while all the spinor indices of the bivector-spinor  $\underline{\Psi}$  have the same chirality, RRR or LLL. In multivector-spinor index notation, the vector-spinor and bivector-spinor parts are

$$\Psi = \underline{\Psi}^{ka} \gamma_k \otimes \gamma_a + \frac{1}{2} \underline{\Psi}^{kla} (\gamma_k \wedge \gamma_l) \otimes \gamma_a , \quad (38.216)$$

where the  $\frac{1}{2}$  in the bivector part cancels the double summation over antisymmetric pairs  $kl$  of indices. The vector-spinor  $\underline{\Psi}$  and bivector-spinor  $\underline{\Psi}$  parts of the gravitino field are analogous to the compressive (trace-full) and tidal (trace-free, Weyl) parts of the Riemann tensor of general relativity.

The chiralities of the field  $\Psi$  are related to those of the gravitino  $\psi$  by

$$-\mathbf{D} \wedge \psi_R = \underline{\Psi}_L + \underline{\Psi}_R . \quad (38.217)$$

The right- and left-handed components of the vector-spinor  $\underline{\Psi}$  have spinor indices LRR and RLL respectively, while the right- and left-handed components of the bivector-spinor  $\underline{\Psi}$  have spinor indices RRR and LLL.

### 38.13.4 Comparison to electromagnetism

The electromagnetic potential is the spinor tensor

$$\mathbf{A} = A^{ab} \gamma_a \otimes \gamma_b , \quad (38.218)$$

where, in accordance with the correspondence (??), the indices  $a$  and  $b$  have opposite chirality, and  $A^{ab}$  is symmetric in  $ab$ . The covariant spacetime derivative

$$\mathbf{D}\mathbf{A} = \mathbf{D} \cdot \mathbf{A} + \mathbf{D} \wedge \mathbf{A} , \quad (38.219)$$

$$\mathbf{D} \cdot \mathbf{A} = D_{ab} A^{ab} , \quad \mathbf{D} \wedge \mathbf{A} = -D_c^a A^{cb} \gamma_a \otimes \gamma_b . \quad (38.220)$$

### 38.13.5 Supersymmetric gauge transformation

The gravitino arises as the connection field of supersymmetry. Under a supersymmetric gauge transformation by a spin- $\frac{1}{2}$  field  $\lambda$ , the gravitino  $\psi$  transforms as

$$\psi^k \rightarrow \psi^k + D^k \lambda . \quad (38.221)$$

In spinor index notation,

$$\psi^{abc} \rightarrow \psi^{abc} + D^{(ab} \lambda^{c)} . \quad (38.222)$$

Notice that the gauge change in  $\psi$  is spin- $\frac{3}{2}$ , a symmetric spinor 3-tensor, as it must be. The gauge change is not  $D\lambda$ , which is spin- $\frac{1}{2}$ . The gauge transformation of the gravitino  $\psi$  may be written in abstract notation

$$\psi \rightarrow \psi + D \otimes \lambda . \quad (38.223)$$

The gauge ambiguity in the gravitino  $\psi$  means that 4 of its 12 complex components are gauge degrees of freedom, leaving 8 physical complex degrees of freedom. The vector-spinor part  $\underline{\Psi}$  of the field transforms under the supersymmetric gauge transformation, but the bivector-spinor part  $\underline{\Psi}$  is gauge-invariant,

$$\underline{\Psi} \rightarrow \underline{\Psi} - D \wedge (D \otimes \lambda) , \quad \underline{\Psi} \rightarrow \underline{\Psi} . \quad (38.224)$$

The vector-spinor field  $\underline{\Psi}$  contains 12 complex components, of which 4 are gauge degrees of freedom, and 8 are physical. The bivector-spinor field  $\underline{\Psi}$  contains 8 complex components, all of which are gauge-invariant. The vector-spinor  $\underline{\Psi}$  and bivector-spinor  $\underline{\Psi}$  each independently capture the 8 physical degrees of freedom of the gravitino  $\psi$ , but only the bivector-spinor  $\underline{\Psi}$  captures them in a gauge-invariant way.

### 38.13.6 Gravitino scalar product

The notions of row spinor, scalar product, and complex conjugation for spin- $\frac{1}{2}$  fields carry over to the spin- $\frac{3}{2}$  field  $\psi$ . Corresponding to the column gravitino  $\psi$  is a row gravitino  $\psi^\cdot$  defined by, analogously to the definition (36.55) for spinors,

$$\psi^\cdot \equiv \psi^\top \varepsilon = (\psi^{abc} \gamma_a \otimes \gamma_b \otimes \gamma_c)^\top \varepsilon , \quad (38.225)$$

where  $\varepsilon$  is the gravitino metric tensor, satisfying, analogously to (36.43),

$$(\gamma_a \otimes \gamma_b \otimes \gamma_c)^\top \varepsilon \gamma_d \otimes \gamma_e \otimes \gamma_f = \varepsilon_{ad} \varepsilon_{be} \varepsilon_{cf} . \quad (38.226)$$

Because spin- $\frac{3}{2}$  fields are always symmetric in their 3 spinor indices, the ordering of indices  $abc$  or  $def$  makes no difference.

The product of a row gravitino  $\psi^\cdot$  with a column gravitino  $\chi$  is a scalar, defining their scalar product

$$\psi \cdot \chi = \psi^\top \varepsilon \chi = \psi^{abc} \chi_{abc} . \quad (38.227)$$

Like the scalar product of spinors, the scalar product of gravitinos is antisymmetric,

$$\psi \cdot \chi = -\chi \cdot \psi . \quad (38.228)$$

In vector-spinor index notation, the scalar product is

$$\psi \cdot \chi = \psi^k \cdot (\gamma_k \wedge \gamma_l) \chi^l . \quad (38.229)$$

The right hand side of equation (38.229) is to be interpreted as, first evaluate the multivector in parentheses, then pre- and post-multiply by the row and column spinors  $\psi^k \cdot$  and  $\chi^l$ . When the multivectors in the scalar product (38.229) are expanded in a spinor basis, the scalar product includes terms that are contractions of  $\psi$  on pairs of its own indices, and likewise for  $\chi$ , but such contractions are spin- $\frac{1}{2}$ , and vanish as long as  $\psi$  and  $\chi$  are pure spin- $\frac{3}{2}$ , with no spin- $\frac{1}{2}$  parts.

### 38.13.7 Conjugate gravitino

The conjugate gravitino field  $\bar{\psi}$  is defined analogously to the definition (36.95) for spinors,

$$\bar{\psi} \equiv C\psi^* = (\psi^{abc})^* C\gamma_a \otimes C\gamma_b \otimes C\gamma_c . \quad (38.230)$$

The conjugate field  $\bar{\psi}$  transforms in the same way as the field  $\psi$  under Lorentz transformations. In vector-spinor index notation,

$$\bar{\psi} = C(\gamma_k \otimes \psi^k)^* = \gamma_k \otimes C(\psi^k)^* = \gamma_k \otimes \bar{\psi}^k . \quad (38.231)$$

### 38.13.8 Spin- $\frac{3}{2}$ Lagrangian

The spin- $\frac{3}{2}$  Rarita-Schwinger Lagrangian comes in Dirac and Majorana varieties. This section covers the Dirac version.

The Lagrangian for a massive spin- $\frac{3}{2}$  particle is the Rarita-Schwinger Lagrangian (Rarita and Schwinger, 1941)

$$L = \bar{\psi} \cdot (\mathbf{D} \wedge \psi + m\psi) . \quad (38.232)$$

In spinor index notation, the Rarita-Schwinger Lagrangian is

$$L = \bar{\psi}^{abc} \left( D_{(a}^d \psi_{dbc)} + m\psi_{abc} \right) . \quad (38.233)$$

In vector-spinor notation, the Lagrangian is

$$L = \bar{\psi}_k \cdot (\gamma^{kmn} D_m + m\gamma^{kn}) \psi_n . \quad (38.234)$$

where  $\gamma^{kn}$  and  $\gamma^{kmn}$  are shorthand for

$$\gamma^{kn} \equiv \gamma^k \wedge \gamma^n , \quad \gamma^{kmn} \equiv \gamma^k \wedge \gamma^m \wedge \gamma^n . \quad (38.235)$$

The right hand side of equation (38.234) is to be interpreted as, first evaluate the multivector in parentheses, then pre- and post-multiply by the row and column spinors  $\bar{\psi}_k \cdot$  and  $\psi_n$ .

Consider more closely the chiral content of the Rarita-Schwinger Lagrangian (38.232). The right-handed (say) chiral component of the gravitino field  $\psi$  has spinor indices LRR. The derivative  $\mathbf{D}$  has spinor indices RL, so the derivative  $\mathbf{D} \wedge \psi$  is a sum of a vector-spinor with indices RLL and a bivector-spinor with indices

RRR. The conjugate  $\bar{\psi}$  has spinor indices RLL opposite to those of  $\psi$ . Since the scalar product links spinor indices only of the same chirality, only the vector-spinor part of  $\mathbf{D} \wedge \psi$  contributes to the Lagrangian, not the bivector-spinor part. This is similar to the situation in general relativity, whose Lagrangian involves only the trace-full Ricci part of the Riemann tensor, not the trace-free Weyl part.

Since the gravitino and its conjugate have opposite chiralities, the mass  $m$  term in the Rarita-Schwinger Lagrangian can be non-vanishing only if the gravitino  $\psi$  contains both chiralities LRR and RLL, similar to the situation for Dirac spinors.

Symmetrized with its complex conjugate, the Rarita-Schwinger Lagrangian (38.232) is

$$L = \frac{1}{2}\bar{\psi} \cdot (\mathbf{D} \wedge \psi + m\psi) - \frac{1}{2}\psi \cdot (\mathbf{D} \wedge \bar{\psi} + m\bar{\psi}) . \quad (38.236)$$

The momenta canonically conjugate to the gravitino  $\psi$  and its conjugate  $\bar{\psi}$  are

$$\frac{1}{2}\boldsymbol{\pi} \cdot = \frac{\partial L}{\partial \mathbf{D} \wedge \psi} = \frac{1}{2}\bar{\psi} \cdot , \quad -\frac{1}{2}\bar{\boldsymbol{\pi}} \cdot = \frac{\partial L}{\partial \mathbf{D} \wedge \bar{\psi}} = -\frac{1}{2}\psi \cdot , \quad (38.237)$$

which shows that

$$\boldsymbol{\pi} = \bar{\psi} . \quad (38.238)$$

Varying the action with respect to the fields  $\psi$  and  $\bar{\psi}$  and their velocities  $\mathbf{D} \wedge \psi$  and  $\mathbf{D} \wedge \bar{\psi}$  gives

$$\delta S = \oint \frac{1}{2} (\boldsymbol{\pi} \cdot \boldsymbol{\gamma}_n \wedge \delta\psi - \bar{\boldsymbol{\pi}} \cdot \boldsymbol{\gamma}_n \wedge \delta\bar{\psi}) d^3x^n + \int \left[ \left( \frac{\partial L}{\partial \psi} + \frac{1}{2}\mathbf{D} \wedge \boldsymbol{\pi} \cdot \right) \delta\psi + \left( \frac{\partial L}{\partial \bar{\psi}} - \frac{1}{2}\mathbf{D} \wedge \bar{\boldsymbol{\pi}} \cdot \right) \delta\bar{\psi} \right] d^4x = 0 . \quad (38.239)$$

The Euler-Lagrange equations of motion for the gravitino field  $\psi$  are, analogously to the Dirac equations (38.11),

$$\frac{1}{2}\mathbf{D} \wedge \boldsymbol{\pi} \cdot = -\frac{\partial L}{\partial \psi} = -\frac{1}{2}\mathbf{D} \wedge \bar{\psi} \cdot - m\bar{\psi} \cdot , \quad \frac{1}{2}\mathbf{D} \wedge \bar{\boldsymbol{\pi}} \cdot = \frac{\partial L}{\partial \bar{\psi}} = -\frac{1}{2}\mathbf{D} \wedge \psi \cdot - m\psi \cdot . \quad (38.240)$$

The resulting Rarita-Schwinger equations of motion are

$$\mathbf{D} \wedge \psi + m\psi = 0 , \quad (38.241a)$$

$$\mathbf{D} \wedge \bar{\psi} + m\bar{\psi} = 0 . \quad (38.241b)$$

However, all these equations involve only the vector-spinor part of  $\mathbf{D} \wedge \psi$ .

### 38.13.9 Bianchi identity

I think the supersymmetric covariant derivative for any element  $a$  of the superspacetime algebra

$$\mathcal{D}_k a = (D_k + \psi_k)a = D_k a + [\psi_k, a] . \quad (38.242)$$

I also think

$$[\psi_k, \psi_l] = 0 \quad (38.243)$$

because the product of two column spinors is zero. So I think

If so, then

$$\mathbf{D} \wedge \Psi = 0 . \quad (38.244)$$

$$D_{[k} \Psi_{lm]} = 0 . \quad (38.245)$$

### 38.13.10 Spin- $\frac{3}{2}$ super-Hamiltonian

The Rarita-Schwinger super-Hamiltonian  $H$  is derived by

$$H = \frac{1}{2} \boldsymbol{\pi} \cdot (\mathbf{D} \wedge \boldsymbol{\psi}) - \frac{1}{2} \bar{\boldsymbol{\pi}} \cdot (\mathbf{D} \wedge \bar{\boldsymbol{\psi}}) - L . \quad (38.246)$$

Same issues as Dirac.

$$H = -\frac{1}{2} m (-\bar{\boldsymbol{\pi}} \cdot \boldsymbol{\pi} + \bar{\boldsymbol{\psi}} \cdot \boldsymbol{\psi}) . \quad (38.247)$$

## 38.14 (Super)-Hamiltonian

Define super-Hamiltonian by  $H(\varphi, \pi^n)$ ,

$$H = \pi^n D_n \varphi - L . \quad (38.248)$$

$$S = \int (\pi^n D_n \varphi - H) d^4x \quad (38.249)$$

$$\delta S = \oint \pi^n \delta \varphi \mathbf{D} \mathbf{x}_n^3 + \int \left\{ - \left( D_n \pi^n + \frac{\partial H}{\partial \varphi} \right) \delta \varphi + \delta \pi^n \left( D_n \varphi - \frac{\partial H}{\partial \pi^n} \right) \right\} d^4x \quad (38.250)$$

Hamilton's equations

$$D_n \pi^n = -\frac{\partial H}{\partial \varphi}, \quad D_n \varphi = \frac{\partial H}{\partial \pi^n} .$$

(38.251)

## 38.15 Conventional Hamiltonian

$H(\varphi, \pi^0)$

$$S = \int (\pi^0 D_0 \varphi - H) d^4x \quad (38.252)$$

$$\delta S = \oint \pi^0 \delta \varphi \mathbf{D} \mathbf{x}_0^3 + \int \left\{ - \left( D_0 \pi^0 + \frac{\partial H}{\partial \varphi} \right) \delta \varphi + \delta \pi^0 \left( D_0 \varphi - \frac{\partial H}{\partial \pi^0} \right) \right\} d^4x \quad (38.253)$$

Hamilton's equations

$$D_0 \pi^0 = -\frac{\partial H}{\partial \varphi}, \quad D_0 \varphi = \frac{\partial H}{\partial \pi^0} . \quad (38.254)$$



---

## Bibliography

- Abramowicz, Marek A. and P. Chris Fragile (2013). “Foundations of Black Hole Accretion Disk Theory”. In: *Living Rev. Rel.* 16, p. 1. DOI: [10.12942/lrr-2013-1](https://doi.org/10.12942/lrr-2013-1). arXiv: [1104.5499 \[astro-ph.HE\]](https://arxiv.org/abs/1104.5499) (cit. on p. 622).
- Ade P. A. R. et al. (Planck Collaboration, 260 authors) (2015). “Planck 2015 results. XIII. Cosmological parameters”. In: arXiv: [1502.01589 \[astro-ph.CO\]](https://arxiv.org/abs/1502.01589) (cit. on pp. 231–233, 238, 243, 244, 690, 702, 715, 744, 747, 756, 785, 786, 844).
- Ade P. A. R. et al. (Planck Collaboration, 263 authors) (2014). “Planck 2013 results. XVI. Cosmological parameters”. In: *Astron. Astrophys.* DOI: [10.1051/0004-6361/201321591](https://doi.org/10.1051/0004-6361/201321591). arXiv: [1303.5076 \[astro-ph.CO\]](https://arxiv.org/abs/1303.5076) (cit. on p. 251).
- Ade P. A. R. et al. (Planck Collaboration, 276 authors) (2013). “Planck 2013 results. I. Overview of products and scientific results”. In: arXiv: [1303.5062 \[astro-ph.CO\]](https://arxiv.org/abs/1303.5062) (cit. on p. 220).
- Ade P. A. R. et al. (BICEP2 Collaboration, 47 authors) (2014). “BICEP2 I: Detection Of B-mode Polarization at Degree Angular Scales”. In: arXiv: [1403.3985 \[astro-ph.CO\]](https://arxiv.org/abs/1403.3985) (cit. on pp. 702, 844).
- Amos, D. E. (1986). “A Portable Package for Bessel Functions of a Complex Argument and Nonnegative Order”. In: *Trans. Math. Software* 12, pp. 265–273 (cit. on p. 836).
- Anderson, Lauren et al. (BOSS Collaboration, 65 authors) (2014). “The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Release 10 and 11 galaxy samples”. In: *Mon. Not. Roy. Astron. Soc.* 441, pp. 24–62. DOI: [10.1093/mnras/stu523](https://doi.org/10.1093/mnras/stu523). arXiv: [1312.4877 \[astro-ph.CO\]](https://arxiv.org/abs/1312.4877) (cit. on pp. 745, 747, 783, 802).
- Anderson, Lauren et al. (BOSS Collaboration, 77 authors) (2013). “The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Release 9 Spectroscopic Galaxy Sample”. In: *Mon. Not. Roy. Astron. Soc.* 428, pp. 1036–1054. DOI: [10.1093/mnras/sts084](https://doi.org/10.1093/mnras/sts084). arXiv: [1203.6594 \[astro-ph.CO\]](https://arxiv.org/abs/1203.6594) (cit. on p. 221).
- Arnold, Peter, Guy D. Moore, and Laurence G. Yaffe (2000). “Transport coefficients in high temperature gauge theories: (I) Leading-log results”. In: *JHEP* 11, p. 001. arXiv: [hep-ph/0010177](https://arxiv.org/abs/hep-ph/0010177) (cit. on p. 552).
- Arnowitt, R., S. Deser, and C. W. Misner (1959). “Dynamical Structure and Definition of Energy in General Relativity”. In: *Phys. Rev.* 116, pp. 1322–1330. DOI: [10.1103/PhysRev.116.1322](https://doi.org/10.1103/PhysRev.116.1322) (cit. on p. 446).
- (1963). “The dynamics of general relativity”. In: *Gravitation: an introduction to current research*. Ed. by L. Witten. John Wiley & Sons, pp. 227–265 (cit. on pp. 446, 451).

- Ashtekar, A. (1987). "New Hamiltonian Formulation of General Relativity". In: *Phys. Rev.* D36, pp. 1587–1602. DOI: [10.1103/PhysRevD.36.1587](https://doi.org/10.1103/PhysRevD.36.1587) (cit. on pp. 434, 435).
- Babichev, E. et al. (2008). "Ultra-hard fluid and scalar field in the Kerr-Newman metric". In: *Phys. Rev.* D78, p. 104027. DOI: [10.1103/PhysRevD.78.104027](https://doi.org/10.1103/PhysRevD.78.104027). arXiv: [0807.0449 \[gr-qc\]](https://arxiv.org/abs/0807.0449) (cit. on p. 550).
- Baker, John G. et al. (2006a). "Binary black hole merger dynamics and waveforms". In: *Phys. Rev.* D73, p. 104002. DOI: [10.1103/PhysRevD.73.104002](https://doi.org/10.1103/PhysRevD.73.104002). arXiv: [gr-qc/0602026 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0602026) (cit. on p. 447).
- (2006b). "Gravitational wave extraction from an inspiraling configuration of merging black holes". In: *Phys. Rev. Lett.* 96, p. 111102. DOI: [10.1103/PhysRevLett.96.111102](https://doi.org/10.1103/PhysRevLett.96.111102). arXiv: [gr-qc/0511103 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0511103) (cit. on p. 447).
- Balbus, Steven A. (2003). "Enhanced Angular Momentum Transport in Accretion Disks". In: *Ann. Rev. Astron. Astrophys.* 41, pp. 555–597. DOI: [10.1146/annurev.astro.41.081401.155207](https://doi.org/10.1146/annurev.astro.41.081401.155207). arXiv: [astro-ph/0306208](https://arxiv.org/abs/astro-ph/0306208) (cit. on p. 623).
- Balbus, Steven A and John F. Hawley (1998). "Instability, turbulence, and enhanced transport in accretion disks". In: *Rev. Mod. Phys.* 70, pp. 1–53 (cit. on p. 623).
- Barbero G., J. Fernando (1995). "Real Ashtekar variables for Lorentzian signature space times". In: *Phys. Rev.* D51, pp. 5507–5510. DOI: [10.1103/PhysRevD.51.5507](https://doi.org/10.1103/PhysRevD.51.5507). arXiv: [gr-qc/9410014 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9410014) (cit. on p. 437).
- Bardeen, J. M. (1970). "Kerr Metric Black Holes". In: *Nature* 226, pp. 64–65. DOI: [10.1038/226064a0](https://doi.org/10.1038/226064a0) (cit. on pp. 623, 624).
- Barrabès, C., W. Israel, and E. Poisson (1990). "Collision of light-like shells and mass inflation in rotating black holes". In: *Class. Quant. Grav.* 7.12, pp. L273–L278 (cit. on p. 638).
- Baumgarte, Thomas W. and Stuart L. Shapiro (1999). "On the numerical integration of Einstein's field equations". In: *Phys. Rev.* D59, p. 024007. DOI: [10.1103/PhysRevD.59.024007](https://doi.org/10.1103/PhysRevD.59.024007). arXiv: [gr-qc/9810065 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9810065) (cit. on pp. 433, 447, 480).
- Belinski, Vladimir A. (2014). "On the cosmological singularity". In: *Proc. XIII Marcel Grossmann Meeting, Stockholm 2012*. arXiv: [1404.3864 \[gr-qc\]](https://arxiv.org/abs/1404.3864) (cit. on pp. 447, 465, 472).
- Belinskii, Vladimir A. and Isaak M. Khalatnikov (1971). "General solution of the gravitational equations with a physical oscillatory singularity". In: *Sov. Phys. JETP* 32, pp. 169–172 (cit. on pp. 447, 472).
- Belinskii, Vladimir A., Isaak M. Khalatnikov, and Evgeny M. Lifshitz (1970). "Oscillatory approach to a singular point in the relativistic cosmology". In: *Advances in Physics* 19, pp. 525–573. DOI: [10.1080/00018737000101171](https://doi.org/10.1080/00018737000101171) (cit. on pp. 447, 472, 480).
- (1972). "Construction of a General Cosmological Solution of the Einstein Equation with a Time Singularity". In: *Sov. Phys. JETP* 35, pp. 838–841 (cit. on pp. 447, 472).
- (1982). "A general solution of the Einstein equations with a time singularity". In: *Advances in Physics* 31, pp. 639–667 (cit. on pp. 447, 465, 472, 473, 478).
- Berger, Beverly K. (2002). "Numerical Approaches to Spacetime Singularities". In: *Living Rev. Rel.* 5, p. 1. arXiv: [gr-qc/0201056](https://arxiv.org/abs/gr-qc/0201056). URL: <http://www.livingreviews.org/lrr-2002-1> (cit. on p. 472).
- Bernardis, P. de et al. (2000). "A flat Universe from high-resolution maps of the cosmic microwave background radiation". In: *Nature* 404, pp. 955–959. eprint: [astro-ph/0004404](https://arxiv.org/abs/astro-ph/0004404) (cit. on p. 822).
- Bertschinger, Edmund (1993). "Cosmological dynamics: Course 1, 1993 Les Houches Lectures". In: arXiv: [astro-ph/9503125 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9503125) (cit. on p. 662).

- Betoule M. et al. (SDSS Collaboration, 63 authors) (2014). “Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples”. In: *Astron. Astrophys.* arXiv: [1401.4064 \[astro-ph.CO\]](#) (cit. on pp. 217, 218).
- 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)”. In: *Soc. Ital. Sci. Mem. di Mat.* 11, p. 267 (cit. on p. 465).
- Blagojević, Milutin and Friedrich W. Hehl (2013). *Gauge Theories of Gravitation*. Imperial College Press (cit. on p. 396).
- Bradley, James (1728). “An account of a new discovered motion of the fixed stars”. In: *Phil. Trans.* 35, pp. 637–661 (cit. on p. 43).
- Brillouin, Léon (1926). “La m’ecanique ondulatoire de Schrödinger: une m’ethode g’en’erale de resolution par approximations successives”. In: *Comptes Rendus de l’Academie des Sciences* 187, pp. 24–26 (cit. on p. 789).
- Campanelli, Manuela, C. O. Lousto, and Y. Zlochower (2006). “The Last orbit of binary black holes”. In: *Phys. Rev.* D73, p. 061501. doi: [10.1103/PhysRevD.73.061501](#). arXiv: [gr-qc/0601091 \[gr-qc\]](#) (cit. on p. 447).
- Campanelli, Manuela et al. (2006). “Accurate evolutions of orbiting black-hole binaries without excision”. In: *Phys. Rev. Lett.* 96, p. 111101. doi: [10.1103/PhysRevLett.96.111101](#). arXiv: [gr-qc/0511048 \[gr-qc\]](#) (cit. on p. 447).
- Carroll, Sean M., William H. Press, and Edwin L. Turner (1992). “The Cosmological constant”. In: *Ann. Rev. Astron. Astrophys.* 30, pp. 499–542. doi: [10.1146/annurev.aa.30.090192.002435](#) (cit. on p. 742).
- Cartan, Élie (1904). “Sur la structure des groupes infinis de transformations”. In: *Annales Scientifiques de l’École Normale Supérieure* 21, pp. 153–206 (cit. on pp. 292, 373, 413).
- Carter, Brandon (1968a). “Global structure of the Kerr family of gravitational fields”. In: *Phys. Rev.* 174, pp. 1559–1571 (cit. on p. 206).
- (1968b). “Hamilton-Jacobi and Schrödinger separable solutions of Einstein’s equations”. In: *Commun. Math. Phys.* 10, pp. 280–310 (cit. on pp. 579, 582, 585, 639).
- Chandrasekhar, Subrahmanyan (1983). *The mathematical theory of black holes*. Oxford, England: Clarendon Press (cit. on pp. 304, 935).
- Chluba, J. and R. M. Thomas (2011). “Towards a complete treatment of the cosmological recombination problem”. In: *Mon. Not. Roy. Astron. Soc.* 412, pp. 748–764. doi: [10.1111/j.1365-2966.2010.17940.x](#). arXiv: [1010.3631 \[astro-ph.CO\]](#) (cit. on p. 773).
- Christodoulou, Demetrios (1984). “Violation of cosmic censorship in the gravitational collapse of a dust cloud”. In: *Commun. Math. Phys.* 93, pp. 171–195. doi: [10.1007/BF01223743](#) (cit. on p. 542).
- Clifford, W. K. (1878). “Applications of Grassmann’s extensive algebra”. In: *Am. J. Math.* 1, pp. 350–358 (cit. on pp. 307, 309).
- Das, Sudeep and 40 authors Sherwin Blake D. et al. (ACT Collaboration (2011). “Detection of the Power Spectrum of Cosmic Microwave Background Lensing by the Atacama Cosmology Telescope”. In: *Phys. Rev. Lett.* 107, p. 021301. doi: [10.1103/PhysRevLett.107.021301](#). arXiv: [1103.2124 \[astro-ph.CO\]](#) (cit. on p. 220).

- Dicke, R. H. et al. (1965). "Cosmic Black-Body Radiation". In: *Astrophys. J. Lett.* 142, pp. 414–419. DOI: [10.1086/148306](https://doi.org/10.1086/148306) (cit. on p. 219).
- Diener, Peter et al. (2006). "Accurate evolution of orbiting binary black holes". In: *Phys. Rev. Lett.* 96, p. 121101. DOI: [10.1103/PhysRevLett.96.121101](https://doi.org/10.1103/PhysRevLett.96.121101). arXiv: [gr-qc/0512108 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0512108) (cit. on p. 447).
- Dirac, Paul A. M. (1964). *Lectures on Quantum Mechanics*. New York: Belfer Graduate School of Science (cit. on pp. 382, 389).
- Doran, Chris (2000). "A new form of the Kerr solution". In: *Phys. Rev. D* 61, p. 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 p. 210).
- Droste, Johannes (1916). "Het veld van een enkel centrum in Einstein's theorie der zwaarerekracht end de beweging van een stoffelijk punt in dat veld". In: *Verslagen van de Gewone Vergaderingen der Wis- en Natuurkundige Afdeeling, Koninklijke Akademie van Wetenschappen te Amsterdam* 25, pp. 163–180 (cit. on pp. 128, 145).
- Eddington, Arthur Stanley (1924). "A comparison of Whitehead's Einstein's formulae". In: *Nature* 113, p. 192 (cit. on p. 146).
- Einstein, Albert (1905). "Zur Elektrodynamik bewegter Körper". In: *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 gezogene Folgerungen". In: *Jahrbuch der Radioaktivität und Elektronik* 4 (cit. on p. 51).
- (1915). "Die Feldgleichungen der Gravitation". In: *Königlich Preussische Akademie der Wissenschaften (Berlin), Sitzungsberichte*, pp. 844–847 (cit. on p. 52).
- Eisenhauer, Frank et al. (2005). "SINFONI in the Galactic Center: young stars and IR flares in the central light month". In: *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. 162).
- Finkelstein, David (1958). "Past-Future Asymmetry of the Gravitational Field of a Point Particle". In: *Phys. Rev.* 110, pp. 965–967. DOI: [10.1103/PhysRev.110.965](https://doi.org/10.1103/PhysRev.110.965) (cit. on p. 146).
- Fixsen, D. J. (2009). "The Temperature of the Cosmic Microwave Background". In: *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 \[astro-ph.CO\]](https://arxiv.org/abs/0911.1955) (cit. on pp. 219, 785).
- Forero, D. V., M. Tortola, and J. W. F. Valle (2012). "Global status of neutrino oscillation parameters after Neutrino-2012". In: *Phys. Rev. D* 86, p. 073012. DOI: [10.1103/PhysRevD.86.073012](https://doi.org/10.1103/PhysRevD.86.073012). arXiv: [1205.4018 \[hep-ph\]](https://arxiv.org/abs/1205.4018) (cit. on p. 254).
- Friedmann, Alexander (1922). "Über die Krümmung des Raumes". In: *Zeitschrift für Physik* A10, pp. 377–386 (cit. on p. 223).
- (1924). "Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes". In: *Zeitschrift für Physik* A21, pp. 326–332 (cit. on p. 223).
- Gebhardt, Karl et al. (2011). "The Black-Hole Mass in M87 from Gemini/NIFS Adaptive Optics Observations". In: *Astrophys. J.* 729, p. 119. DOI: [10.1088/0004-637X/729/2/119](https://doi.org/10.1088/0004-637X/729/2/119). arXiv: [1101.1954 \[astro-ph.CO\]](https://arxiv.org/abs/1101.1954) (cit. on p. 123).
- Geroch, R., A. Held, and R. Penrose (1973). "A space-time calculus based on pairs of null directions". In: *J. Math. Phys.* 14, pp. 874–881 (cit. on p. 868).

- Ghez, A. M. et al. (2005). “Stellar Orbits Around the Galactic Center Black Hole”. In: *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. 162).
- Ghez, A. M. et al. (2008). “Measuring Distance and Properties of the Milky Way’s Central Supermassive Black Hole with Stellar Orbits”. In: *Astrophys. J.* 689, pp. 1044–1062. DOI: [10.1086/592738](https://doi.org/10.1086/592738). arXiv: [0808.2870 \[astro-ph\]](https://arxiv.org/abs/0808.2870) (cit. on p. 123).
- Gillessen, S. et al. (2009). “Monitoring stellar orbits around the Massive Black Hole in the Galactic Center”. In: *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 \[astro-ph\]](https://arxiv.org/abs/0810.4674) (cit. on p. 123).
- Goldberg, J. N. et al. (1967). “Spin- $s$  spherical harmonics and  $\delta$ ”. In: *J. Math. Phys.* 8, pp. 2155–61 (cit. on p. 868).
- Goldman, Martin (1984). *The Demon in the Aether: The Story of James Clerk Maxwell*. Adam Hilger (cit. on p. 10).
- Grassmann, Hermann (1862). *Die Ausdehnungslehre. Vollständig und in strenger Form begründet*. Berlin: Enslin (cit. on pp. 307, 308).
- (1877). “Der Ort der Hamilton’schen Quaternionen in der Ausdehnungslehre”. In: *Math. Ann.* 12, pp. 375–386 (cit. on pp. 307, 309).
- Graves, John C. and Dieter R. Brill (1960). “Oscillatory Character of Reissner-Nordström Metric for an Ideal Charged Wormhole”. In: *Phys. Rev.* 120, pp. 1507–1513. DOI: [10.1103/PhysRev.120.1507](https://doi.org/10.1103/PhysRev.120.1507) (cit. on p. 181).
- Gullstrand, Allvar (1922). “Allgemeine Lösung des statischen Einkörperproblems in der Einsteinschen Gravitationstheorie”. In: *Arkiv. Mat. Astron. Fys.* 16(8), pp. 1–15 (cit. on pp. 137, 511).
- Guth, A. H. (1981). “Inflationary universe: A possible solution to the horizon and flatness problems”. In: *Phys. Rev.* D23, pp. 347–356 (cit. on p. 246).
- Hamilton, Andrew J. S. and Jason P. Lisle (2008). “The river model of black holes”. In: *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. 125).
- Hamilton, Andrew J. S. and Gavin Polhemus (2010). “Stereoscopic visualization in curved spacetime: seeing deep inside a black hole”. In: *New J. Phys.* 12, p. 123027. DOI: [10.1088/1367-2630/12/12/123027](https://doi.org/10.1088/1367-2630/12/12/123027). arXiv: [1012.4043 \[gr-qc\]](https://arxiv.org/abs/1012.4043) (cit. on pp. 160, 161).
- Hammond, Richard T. (2002). “Torsion gravity”. In: *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. 396).
- Harrison, E. R. (1970). “Fluctuations at the Threshold of Classical Cosmology”. In: *Phys. Rev. D* 1 (10), pp. 2726–2730. DOI: [10.1103/PhysRevD.1.2726](https://doi.org/10.1103/PhysRevD.1.2726) (cit. on p. 743).
- Hawking, Stephen W. and George F. R. Ellis (1973). *The large scale structure of space-time*. Cambridge University Press (cit. on pp. 156, 161, 245, 489, 505).
- Hehl, Friedrich W. (2012). “Gauge Theory of Gravity and Spacetime”. In: arXiv: [1204.3672 \[gr-qc\]](https://arxiv.org/abs/1204.3672) (cit. on p. 396).
- Hehl, Friedrich W., Paul von der Heyde, and G. David Kerlick (1976). “General relativity with spin and torsion: Foundations and prospects”. In: *Rev. Mod. Phys.* 48, pp. 393–416 (cit. on p. 293).

- Hehl, Friedrich W. et al. (1995). "Metric affine gauge theory of gravity: Field equations, Noether identities, world spinors, and breaking of dilation invariance". In: *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 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9402012) (cit. on pp. 399, 422).
- Held, A. (1974). "A formalism for the investigation of algebraically special metrics. I". In: *Commun. Math. Phys.* 37, pp. 311–326 (cit. on p. 300).
- Hestenes, David (1966). *Space-Time Algebra*. Gordon & Breach (cit. on p. 307).
- Hestenes, David and Garret Sobczyk (1987). *Clifford Algebra to Geometric Calculus*. D. Reidel Publishing Company (cit. on p. 307).
- Hilbert, David (1915). "Die Grundlagen der Physik". In: *Konigl. Gesell. d. Wiss. Göttingen, Nachr., Math.-Phys. Kl.* Pp. 395–407 (cit. on pp. 52, 291, 373, 390).
- Hinshaw, G. et al. (2012). "Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results". In: arXiv: [1212.5226 \[astro-ph.CO\]](https://arxiv.org/abs/1212.5226) (cit. on pp. 220, 232, 233, 243, 244).
- Holst, Soren (1996). "Barbero's Hamiltonian derived from a generalized Hilbert-Palatini action". In: *Phys. Rev. D* 53, pp. 5966–5969. DOI: [10.1103/PhysRevD.53.5966](https://doi.org/10.1103/PhysRevD.53.5966). arXiv: [gr-qc/9511026 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9511026) (cit. on p. 437).
- Hu, Wayne and Martin J. White (1997). "CMB anisotropies: Total angular momentum method". In: *Phys. Rev. D* 56, pp. 596–615. DOI: [10.1103/PhysRevD.56.596](https://doi.org/10.1103/PhysRevD.56.596). arXiv: [astro-ph/9702170 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9702170) (cit. on pp. 849, 850).
- Hubble, Edwin P. (1929). "A relation between distance and radial velocity among extra-galactic nebulae". In: *Proc. Nat. Acad. Sci.* 15, pp. 168–173 (cit. on p. 217).
- 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". In: *Mon. Not. Roy. Astron. Soc.* 268, pp. 109–112 (cit. on p. 774).
- Hummer, D. G. and P. J. Storey (1998). "Recombination of helium-like ions – I. Photoionization cross-sections and total recombination and cooling coefficients for atomic helium". In: *Mon. Not. Roy. Astron. Soc.* 297, pp. 1073–1078. DOI: [10.1046/j.1365-8711.1998.2970041073.x](https://doi.org/10.1046/j.1365-8711.1998.2970041073.x) (cit. on p. 774).
- Immirzi, Giorgio (1997). "Real and complex connections for canonical gravity". In: *Class. Quant. Grav.* 14, pp. L177–L181. DOI: [10.1088/0264-9381/14/10/002](https://doi.org/10.1088/0264-9381/14/10/002). arXiv: [gr-qc/9612030 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9612030) (cit. on p. 437).
- Jacobson, Ted and Lee Smolin (1988). "Nonperturbative Quantum Geometries". In: *Nucl. Phys.* B299, p. 295. DOI: [10.1016/0550-3213\(88\)90286-6](https://doi.org/10.1016/0550-3213(88)90286-6) (cit. on p. 437).
- Jarosik, N. et al. (2011). "Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results". In: *Astrophys. J. Suppl.* 192, p. 14. DOI: [10.1088/0067-0049/192/2/14](https://doi.org/10.1088/0067-0049/192/2/14). arXiv: [1001.4744 \[astro-ph.CO\]](https://arxiv.org/abs/1001.4744) (cit. on p. 220).
- Kagramanova, Valeria et al. (2010). "Analytic treatment of complete and incomplete geodesics in Taub-NUT space-times". In: *Phys. Rev. D* 81, p. 124044. DOI: [10.1103/PhysRevD.81.124044](https://doi.org/10.1103/PhysRevD.81.124044). arXiv: [1002.4342 \[gr-qc\]](https://arxiv.org/abs/1002.4342) (cit. on pp. 589, 594).
- Kaiser, N. (1984). "On the spatial correlations of Abell clusters". In: *Astrophys. J. Lett.* 284, pp. L9–L12. DOI: [10.1086/184341](https://doi.org/10.1086/184341) (cit. on p. 746).
- Kasner, Edward (1921). "Geometrical Theorems on Einstein's Cosmological Equations". In: *Am. J. Math.* 43, pp. 217–221. DOI: [10.2307/2370192](https://doi.org/10.2307/2370192) (cit. on pp. 472, 475).
- Keisler, R. and 49 authors) Reichardt C. L. et al. (SPT Collaboration (2011). "A Measurement of the Damping Tail of the Cosmic Microwave Background Power Spectrum with the South Pole Telescope". In:

- Astrophys. J.* 743, p. 28. doi: [10.1088/0004-637X/743/1/28](https://doi.org/10.1088/0004-637X/743/1/28). arXiv: [1105.3182 \[astro-ph.CO\]](https://arxiv.org/abs/1105.3182) (cit. on p. 220).
- Kerr, Roy P. (1963). “Gravitational field of a spinning mass as an example of algebraically special metrics”. In: *Phys. Rev. Lett.* 11, pp. 237–238 (cit. on p. 200).
- (2009). “Discovering the Kerr and Kerr-Schild metrics”. In: *The Kerr spacetime: rotating black holes in general relativity*. Ed. by David L. Wiltshire, Matt Visser, and Susan Scott. Cambridge University Press, pp. 38–72. arXiv: [0706.1109 \[gr-qc\]](https://arxiv.org/abs/0706.1109) (cit. on p. 200).
- Kolb, Edward W. and Michael S. Turner (1990). “The Early universe”. In: *Front. Phys.* 69, pp. 1–547 (cit. on p. 252).
- Kormendy, John and Karl Gebhardt (2001). “Supermassive Black Holes in Nuclei of Galaxies”. In: *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. 123).
- Kramers, H. A. (1926). “Wellenmechanik und halbzählig Quantisierung”. In: *Zeitschrift für Physik* 39, pp. 828–840. doi: [10.1007/BF01451751](https://doi.org/10.1007/BF01451751) (cit. on p. 789).
- Kruskal, Martin D. (1960). “Maximal extension of Schwarzschild metric”. In: *Phys. Rev.* 119, pp. 1743–1745 (cit. on p. 147).
- Landau, L. D. and E. M. Lifshitz (1975). *The Classical Theory of Fields, Fourth revised English Edition*. Pergamon Press (cit. on p. 344).
- Leavitt, Henrietta S. and Edward C. Pickering (1912). “Periods of 25 Variable Stars in the Small Magellanic Cloud”. In: *Harvard College Observatory Circular* 173, pp. 1–3 (cit. on p. 217).
- Lemaître, Georges (1927). “Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extragalactiques”. In: *Ann. Soc. Sci. Brux.* 47A, p. 49 (cit. on pp. 217, 223).
- (1931). “Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae”. In: *Mon. Not. Roy. Astron. Soc.* 91, pp. 483–490 (cit. on p. 223).
- Lense, Josef and Hans Thirring (1918). “Ueber den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie”. In: *Phys. Z.* 19, pp. 156–163 (cit. on p. 674).
- Lorentz, Hendrik Antoon (1904). “Electromagnetic phenomena in a system moving with any velocity smaller than that of light”. In: *Proceedings of the Royal Netherlands Academy of Arts and Sciences* 6, pp. 809–831 (cit. on p. 11).
- Luzum, Brian et al. (2009). “The IAU 2009 system of astronomical constants: the report of the IAU working group on numerical standards for Fundamental Astronomy”. In: *Celestial Mechanics and Dynamical Astronomy* 110, pp. 293–304 (cit. on p. 674).
- Ma, Chung-Pei and Edmund Bertschinger (1995). “Cosmological perturbation theory in the synchronous and conformal Newtonian gauges”. In: *Astrophys.J.* 455, pp. 7–25. doi: [10.1086/176550](https://doi.org/10.1086/176550). arXiv: [astro-ph/9506072 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9506072) (cit. on p. 841).
- Mangano, G. et al. (2002). “A Precision calculation of the effective number of cosmological neutrinos”. In: *Phys. Lett.* B534, 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 \[astro-ph\]](https://arxiv.org/abs/astro-ph/0111408) (cit. on p. 786).

- McClintock, Jeffrey E. et al. (2011). “Measuring the Spins of Accreting Black Holes”. In: *Class. Quant. Grav.* 28, p. 114009. DOI: [10.1088/0264-9381/28/11/114009](https://doi.org/10.1088/0264-9381/28/11/114009). arXiv: [1101.0811 \[astro-ph.HE\]](https://arxiv.org/abs/1101.0811) (cit. on p. 123).
- Mellinger, Axel (2009). “A Color All-Sky Panorama Image of the Milky Way”. In: *Pub. Astron. Soc. Pacific* 121, pp. 1180–1187 (cit. on p. 162).
- Michell, John (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”. In: *Phil. Trans. Roy. Soc. London* 74, pp. 35–57 (cit. on p. 125).
- Minkowski, Hermann (1909). “Raum und Zeit”. In: *Physikalische Zeitschrift* 10, pp. 75–88 (cit. on p. 13).
- Misner, Charles W. and David H. Sharp (1964). “Relativistic equations for adiabatic, spherically symmetric gravitational collapse”. In: *Phys. Rev.* 136, B571–B576 (cit. on p. 527).
- Misner, Charles W., Kip S. Thorne, and John A. Wheeler (1973). *Gravitation*. W. H. Freeman and Co. (cit. on pp. 106, 295).
- Newman, E. T. and R. Penrose (1962). “An Approach to Gravitational Radiation by a Method of Spin Coefficients”. In: *J. Math. Phys.* 3, pp. 566–579 (cit. on pp. 300, 868, 885).
- (2009). “Spin-coefficient formalism”. In: *Scholarpedia* 4(6), p. 7445. URL: [http://www.scholarpedia.org/article/Newman\\_Penrose\\_formalism](http://www.scholarpedia.org/article/Newman_Penrose_formalism) (cit. on p. 300).
- Newman, E. T., L. A. Tamburino, and T. Unti (1963). “Empty-space generalization of the Schwarzschild metric”. In: *J. Math. Phys.* 4, pp. 915–923 (cit. on pp. 588, 594).
- Newman, E. T. et al. (1965). “Metric of a rotating, charged mass”. In: *J. Math. Phys.* 6, pp. 918–919 (cit. on p. 200).
- Nieuwenhuizen, Peter van (1981). “Supergravity”. In: *Phys. Rept.* 68, pp. 189–398 (cit. on p. 67).
- Noether, E. (1918). “Invariante Variationsprobleme”. In: *Nachr. D. König. Gesellsch. D. Wiss. Zu Göttingen, Math-phys. Klasse* 1918, pp. 235–257 (cit. on pp. 376, 381).
- Nolte, David D. (2010). “The tangled tale of phase space”. In: *Physics Today* 63 (4), pp. 33–38 (cit. on p. 118).
- Nordström, Gunnar (1918). “On the energy of the gravitational field in Einstein’s theory”. In: *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen* 20, pp. 1238–1245 (cit. on p. 181).
- O’Donnell, S. (1983). *Portrait of a Prodigy*. Dublin: Boole Press (cit. on p. 322).
- Oppenheimer, J. R. and H. Snyder (1939). “On continued gravitational contraction”. In: *Phys. Rev.* 56, pp. 455–459. DOI: [10.1103/PhysRev.56.455](https://doi.org/10.1103/PhysRev.56.455) (cit. on pp. 156, 157, 541, 542).
- Padmanabhan, T. (2010). *Gravitation: Foundations and Frontiers*. Cambridge University Press (cit. on p. 405).
- Painlevé, Paul (1921). “La mécanique classique et la théorie de la relativité”. In: *Comptes Rendus de l’Académie des sciences (Paris)* 173, pp. 677–680 (cit. on pp. 137, 146, 511).
- Palatini, Attilio (1919). “Deduzione invariantiva delle equazioni gravitazionali dal principio di Hamilton”. In: *Rend. Circ. Mat. Palermo* 43, pp. 203–212 (cit. on p. 393).
- Peebles, P. J. E. (1968). “Recombination of the Primeval Plasma”. In: *Astrophys. J.* 153, pp. 1–11. DOI: [10.1086/149628](https://doi.org/10.1086/149628) (cit. on pp. 756, 768, 769, 773).

- Penrose, Roger (1959). "The Apparent shape of a relativistically moving sphere". In: *Proc. Cambridge Phil. Soc.* 55, pp. 137–139. DOI: [10.1017/S0305004100033776](https://doi.org/10.1017/S0305004100033776) (cit. on p. 42).
- (1965). "Gravitational collapse and spacetime singularities". In: *Phys. Rev. Lett.* 14, pp. 57–59 (cit. on pp. 489, 504, 636).
- (2011). "Conformal treatment of infinity". In: *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. 153).
- Penzias, Arno A. and Robert W. Wilson (1965). "A Measurement of Excess Antenna Temperature at 4080 Mc/s". In: *Astrophys. J. Lett.* 142, pp. 419–421. DOI: [10.1086/148307](https://doi.org/10.1086/148307) (cit. on p. 219).
- Perlmutter S. et al. (Supernova Cosmology Project, 32 authors) (1999). "Measurements of Omega and Lambda from 42 high redshift supernovae". In: *Astrophys. J.* 517, pp. 565–586. DOI: [10.1086/307221](https://doi.org/10.1086/307221). arXiv: [astro-ph/9812133 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9812133) (cit. on p. 219).
- Peskin, Michael E. and Daniel V. Schroeder (1995). *An Introduction to Quantum Field Theory*. Reading, MA: Perseus Books (cit. on pp. 764, 811).
- Poincaré, Henri (1905). "On the Dynamics of the Electron". In: *Comptes Rendus* 140, pp. 1504–1508 (cit. on p. 11).
- Poisson, E. and W. Israel (1990). "Internal structure of black holes". In: *Phys. Rev.* D41, pp. 1796–1809. DOI: [10.1103/PhysRevD.41.1796](https://doi.org/10.1103/PhysRevD.41.1796) (cit. on pp. 509, 559, 636, 638).
- Pretorius, Frans (2005a). "Evolution of binary black hole spacetimes". In: *Phys. Rev. Lett.* 95, p. 121101. DOI: [10.1103/PhysRevLett.95.121101](https://doi.org/10.1103/PhysRevLett.95.121101). arXiv: [gr-qc/0507014 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0507014) (cit. on p. 447).
- (2005b). "Numerical relativity using a generalized harmonic decomposition". In: *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 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0407110) (cit. on pp. 447, 484, 485).
- (2006). "Simulation of binary black hole spacetimes with a harmonic evolution scheme". In: *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 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0602115) (cit. on p. 447).
- Pretorius, Frans and Werner Israel (1998). "Quasispherical light cones of the Kerr geometry". In: *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 \[gr-qc\]](https://arxiv.org/abs/gr-qc/9803080) (cit. on pp. 634–636).
- Rarita, William and Julian Schwinger (1941). "On a Theory of Particles with Half-Integral Spin". In: *Phys. Rev.* 60 (1), pp. 61–61. DOI: [10.1103/PhysRev.60.61](https://doi.org/10.1103/PhysRev.60.61). URL: <http://link.aps.org/doi/10.1103/PhysRev.60.61> (cit. on p. 951).
- Raychaudhuri, Amalkumar (1955). "Relativistic cosmology. I". In: *Phys. Rev.* 98, pp. 1123–1126. DOI: [10.1103/PhysRev.98.1123](https://doi.org/10.1103/PhysRev.98.1123) (cit. on p. 491).
- Regge, T. and John A. Wheeler (1957). "Stability of the Schwarzschild singularity". In: *Phys. Rev.* 108, pp. 1063–1069 (cit. on p. 147).
- Reissner, Hans (1916). "Über die Eigengravitation des electrischen Feldes nach der Einsteinschen Theorie". In: *Annalen der Physik (Leipzig)* 50, pp. 106–120 (cit. on p. 181).
- Riess, Adam G. and 20 authors) Filippenko Alexei V. et al. (Supernova Search Team (1998). "Observational evidence from supernovae for an accelerating universe and a cosmological constant". In: *Astron. J.* 116, pp. 1009–1038. DOI: [10.1086/300499](https://doi.org/10.1086/300499). arXiv: [astro-ph/9805201 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9805201) (cit. on p. 219).

- Riess, Adam G. et al. (2011). “A 3% Solution: Determination of the Hubble Constant with the Hubble Space Telescope and Wide Field Camera 3”. In: *Astrophys. J.* 730, p. 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://arxiv.org/abs/1103.2976). arXiv: [1103.2976 \[astro-ph.CO\]](https://arxiv.org/abs/1103.2976) (cit. on pp. 230, 233).
- Robertson, Howard Percy (1935). “Kinematics and world structure”. In: *Astrophys. J.* 82, pp. 284–301 (cit. on p. 223).
- (1936a). “Kinematics and world structure II”. In: *Astrophys. J.* 83, pp. 187–201 (cit. on p. 223).
- (1936b). “Kinematics and world structure III”. In: *Astrophys. J.* 83, pp. 257–271 (cit. on p. 223).
- Rovelli, Carlo (2007). *Quantum Gravity*. Cambridge University Press (cit. on p. 97).
- Sachs, R. K. (1961). “Gravitational waves in general relativity. 6. The outgoing radiation condition”. In: *Proc. Roy. Soc. London A* 264, pp. 309–338 (cit. on p. 496).
- Sachs, R. K. and A. M. Wolfe (1967). “Perturbations of a cosmological model and angular variations of the microwave background”. In: *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. 838).
- Schwarzschild, Karl (1916a). “Über das Gravitationsfeld eines Kugel aus inkompressibler Flüssigkeit nach der Einsteinschen Theorie”. In: *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. 540).
- (1916b). “Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie”. In: *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. 128).
- Scott, D. and A. Moss (2009). “Matter temperature during cosmological recombination”. In: *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 \[astro-ph.CO\]](https://arxiv.org/abs/0902.3438) (cit. on p. 767).
- Seager, Sara, Dimitar D. Sasselov, and Douglas Scott (1999). “A new calculation of the recombination epoch”. In: *Astrophys. J.* 523, pp. L1–L5. DOI: [10.1086/312250](https://doi.org/10.1086/312250). arXiv: [astro-ph/9909275 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9909275) (cit. on pp. 773, 774).
- (2000). “How exactly did the universe become neutral?” In: *Astrophys. J. Suppl.* 128, pp. 407–430. DOI: [10.1086/313388](https://doi.org/10.1086/313388). arXiv: [astro-ph/9912182 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9912182) (cit. on pp. 773, 774).
- Seljak, Uros and Matias Zaldarriaga (1996). “A Line of sight integration approach to cosmic microwave background anisotropies”. In: *Astrophys. J.* 469, pp. 437–444. DOI: [10.1086/177793](https://doi.org/10.1086/177793). arXiv: [astro-ph/9603033 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9603033) (cit. on p. 822).
- (1997). “Signature of gravity waves in polarization of the microwave background”. In: *Phys. Rev. Lett.* 78, pp. 2054–2057. DOI: [10.1103/PhysRevLett.78.2054](https://doi.org/10.1103/PhysRevLett.78.2054). arXiv: [astro-ph/9609169 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9609169) (cit. on p. 849).
- Senovilla, José M. M. (1997). “Singularity theorems and their consequences”. In: *Gen. Rel. Grav.* 29, pp. 701–848 (cit. on pp. 489, 504).
- Shapiro, I. I. (1964). “Fourth test of general relativity”. In: *Phys. Rev. Lett.* 13, pp. 789–791. DOI: [10.1103/PhysRevLett.13.789](https://doi.org/10.1103/PhysRevLett.13.789) (cit. on p. 84).
- Shibata, Masaru and Takashi Nakamura (1995). “Evolution of three-dimensional gravitational waves: Harmonic slicing case”. In: *Phys. Rev. D* 52, pp. 5428–5444. DOI: [10.1103/PhysRevD.52.5428](https://doi.org/10.1103/PhysRevD.52.5428) (cit. on pp. 433, 447, 480).

- Shinkai, Hisa-aki (2009). "Formulations of the Einstein equations for numerical simulations". In: *J. Korean Phys. Soc.* 54, pp. 2513–2528. DOI: [10.3938/jkps.54.2513](https://doi.org/10.3938/jkps.54.2513). arXiv: [0805.0068 \[gr-qc\]](https://arxiv.org/abs/0805.0068) (cit. on p. 480).
- Sitter, W. de (1916). "On Einstein's theory of gravitation and its astronomical consequences. Second paper". In: *Mon. Not. Roy. Astron. Soc.* 77, pp. 155–184 (cit. on p. 674).
- Smarr, Larry and Jr. York James W. (1978). "Kinematical conditions in the construction of space-time". In: *Phys. Rev. D* 17, pp. 2529–2551. DOI: [10.1103/PhysRevD.17.2529](https://doi.org/10.1103/PhysRevD.17.2529) (cit. on p. 451).
- Sopuerta, Carlos F., Ulrich Sperhake, and Pablo Laguna (2006). "Hydro-without-hydro framework for simulations of black hole-neutron star binaries". In: *Class. Quant. Grav.* 23, S579–S598. DOI: [10.1088/0264-9381/23/16/S15](https://doi.org/10.1088/0264-9381/23/16/S15). arXiv: [gr-qc/0605018 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0605018) (cit. on p. 447).
- Stephani, Hans et al. (2003). *Exact solutions of Einstein's field equations, 2nd edition*. Cambridge, England: Cambridge University Press (cit. on pp. 579, 589, 594).
- Susskind, Leonard (2003). "Black Holes and the Information Paradox". In: *Sci. Am.* 13, pp. 18–23 (cit. on p. 125).
- Susskind, Leonard, Lárus Thorlacius, and John Uglum (1993). "The stretched horizon and black hole complementarity". In: *Phys. Rev.* D48, pp. 3743–3761. DOI: [10.1103/PhysRevD.48.3743](https://doi.org/10.1103/PhysRevD.48.3743). arXiv: [hep-th/9306069](https://arxiv.org/abs/hep-th/9306069) (cit. on pp. 161, 164).
- Szekeres, George (1960). "On the singularities of a Riemann manifold". In: *Publ. Mat. Debrecen* 7, pp. 285–301 (cit. on p. 147).
- Taub, A. H. (1951). "Empty space-times admitting a three parameter group of motions". In: *Ann. Math.* 53, pp. 472–490 (cit. on pp. 588, 594).
- Terrell, James (1959). "Invisibility of the Lorentz Contraction". In: *Phys. Rev.* 116, pp. 1041–1045. DOI: [10.1103/PhysRev.116.1041](https://doi.org/10.1103/PhysRev.116.1041) (cit. on p. 42).
- Thirring, Hans (1918). "Republication of: On the formal analogy between the basic electromagnetic equations and Einstein's gravity equations in first approximation". In: *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. 674).
- Thorne, Kip S. (1974). "Disk-accretion onto a black hole. II. Evolution of the hole". In: *Astrophys. J.* 191, pp. 507–519. DOI: [10.1086/152991](https://doi.org/10.1086/152991) (cit. on p. 624).
- (1994). *Black holes and time warps: Einstein's outrageous legacy*. W. W. Norton (cit. on pp. 10, 128).
- Unruh, W. G. (1976). "Notes on black hole evaporation". In: *Phys. Rev. D* 14, p. 870. DOI: [10.1103/PhysRevD.14.870](https://doi.org/10.1103/PhysRevD.14.870) (cit. on p. 165).
- Walker, Arthur Geoffrey (1937). "On Milne's theory of world-structure". In: *Proc. London Math. Soc.* 2 42, pp. 90–127 (cit. on p. 223).
- Weisberg, Joel M. and Joseph H. Taylor (2005). "Relativistic binary pulsar B1913+16: Thirty years of observations and analysis". In: *ASP Conf. Ser.* 328, pp. 25–31. arXiv: [astro-ph/0407149 \[astro-ph\]](https://arxiv.org/abs/astro-ph/0407149) (cit. on p. 681).
- Wentzel, Gregor (1926). "Eine Verallgemeinerung der Quantenbedingungen für die Zwecke der Wellenmechanik". In: *Zeitschrift für Physik* 38, pp. 518–529. DOI: [10.1007/BF01397171](https://doi.org/10.1007/BF01397171) (cit. on p. 789).
- Weyl, Hermann (1917). "Zur Gravitationstheorie". In: *Ann. Phys. (Berlin)* 54, pp. 117–145. DOI: [10.1002/andp.19173591804](https://doi.org/10.1002/andp.19173591804) (cit. on p. 181).

- White, Simon D. M., C. S. Frenk, and M. Davis (1983). “Clustering in a Neutrino Dominated Universe”. In: *Astrophys. J.* 274, pp. L1–L5. DOI: [10.1086/161425](https://doi.org/10.1086/161425) (cit. on p. 819).
- Will, Clifford M. (2005). “The confrontation between general relativity and experiment”. In: *Living Rev. Rel.* 9, p. 3. arXiv: [gr-qc/0510072](https://arxiv.org/abs/gr-qc/0510072) (cit. on p. 50).
- Wong, Wan Yan, Adam Moss, and Douglas Scott (2008). “How well do we understand cosmological recombination?” In: *Mon. Not. Roy. Astron. Soc.* 386, pp. 1023–1028. DOI: [10.1111/j.1365-2966.2008.13092.x](https://doi.org/10.1111/j.1365-2966.2008.13092.x). arXiv: [0711.1357 \[astro-ph\]](https://arxiv.org/abs/0711.1357) (cit. on p. 773).
- York Jr., J. W. (1979). “Kinematics and dynamics of general relativity”. In: *Sources of Gravitational Radiation*. Ed. by L. L. Smarr. Cambridge, England: Cambridge University Press, pp. 83–126 (cit. on p. 451).
- Zaldarriaga, Matias and Uros Seljak (1997). “An all sky analysis of polarization in the microwave background”. In: *Phys. Rev. D* 55, pp. 1830–1840. DOI: [10.1103/PhysRevD.55.1830](https://doi.org/10.1103/PhysRevD.55.1830). arXiv: [astro-ph/9609170 \[astro-ph\]](https://arxiv.org/abs/astro-ph/9609170) (cit. on p. 849).
- Zeldovich, Y. B. (1972). “A hypothesis, unifying the structure and the entropy of the Universe”. In: *Mon. Not. Roy. Astron. Soc.* 160, 1P–3P (cit. on p. 743).

---

## Index

Globally inertial frame, 8  
Postulates of special relativity, 8  
Special relativity, 7