

Elements of Astrophysics

Nick Kaiser

April 21, 2002

Preface

These are the notes that have grown out of a introductory graduate course I have given for the past few years at the IfA. They are meant to be a ‘primer’ for students embarking on a Ph.D. in astronomy. The level is somewhat shallower than standard textbook courses, but quite a broad range of material is covered. The goal is to get the student to the point of being able to make meaningful order-of-magnitude calculations — and a number of problems are included — and to give the students a fairly uniform base in the relevant physics that they can use as a starting point and introduction to the more detailed textbooks they will need to use when they come to address serious problems. The books that I have drawn upon extensively in devising this course are Rybicki and Lightman; Shu’s two-volume series; Longair’s two-volume series; various sections of Landau and Lifshitz (Classical theory of fields; Mechanics and Fluid Mechanics in particular); Huang’s statistical mechanics, and Binney and Tremaine. Some sections here are rather terse overviews of the relevant parts of these texts, but there are some other areas which I felt were not adequately covered, where I have tried to give more elaborate coverage. The reader is strongly encouraged to consult these texts along with the present work and particularly to attempt the relevant problems contained in many of these.

The book is organised in the following sections:

- **Preliminaries** — We review aspects special relativity, Lagrangian and Hamiltonian dynamics, and the mathematics of random processes.
- **Radiation** — The course follows quite closely the first few chapters of Rybicki and Lightman. We review the macroscopic properties of electromagnetic radiation; we briefly review the concepts of radiative transfer and then consider the properties of thermal radiation and show the relation connection between the Planck spectrum and Einstein’s discovery of stimulated emission. The treatment of polarization in chapter is done somewhat differently and more attention is given to radiation propagation both in the geometric optics limit and via diffraction theory. This section concludes with a general discussion of radiation from moving charges, followed by specific chapters for the important radiation mechanisms of bremsstrahlung, synchrotron radiation and Compton scattering.
- **Field Theory** Initially an informal introduction to the ‘matter’ section, this has now expanded to become a major part of the book.
- **Matter** Starting with the reaction cross sections as computed from field theory we develop kinetic theory and the Boltzmann transport equation, which in turn forms the basis for fluid dynamics. The goal is to show how the approximate, macroscopic theory is based on fundamental physics. We then consider ideal fluids; viscous fluids; fluid instabilities and supersonic flows and shocks. Also covered here is the propagation of electromagnetic waves in a plasma.
- **Gravity** — We start with a brief review of Newtonian gravity and review properties of simple spherical model systems. We then consider collisionless dynamics, with particular emphasis on their use for determining masses of astronomical systems.
- **Cosmology** We then consider cosmology, cosmological fluctuations and gravitational lensing (TBD).
- **Appendices** — In an attempt to make the course self-contained I have included some basic results from vector calculus and Fourier transform theory. There is a brief and simple review of the Boltzmann formula, and appendices on dispersive waves, the relativistic covariance of electromagnetism, and complex analysis.

There are still some major holes in the syllabus. Little attention is given to neutron stars and black holes, for instance, nor to accretion disk theory or MHD. These shortcomings reflect the interests of the author.

I am continually finding errors in the text, and am grateful to the students who have suffered through the course who have pointed out many other errors.

Contents

I	Preliminaries	17
1	Special Relativity	19
1.1	Time Dilation	19
1.2	Length Contraction	21
1.3	Lorentz Transformation	23
1.4	Four-vectors	24
1.5	The 4-velocity	26
1.6	The 4-acceleration	27
1.7	The 4-momentum	27
1.8	Doppler Effect	29
1.9	Relativistic Beaming	29
1.10	Relativistic Decays	30
1.11	Invariant Volumes and Densities	31
1.11.1	Space-Time Volume Element	31
1.11.2	Momentum-Space Volume Element	32
1.11.3	Momentum-Space Density	32
1.11.4	Spatial Volume and Density	32
1.11.5	Phase-Space Density	33
1.11.6	Specific Intensity	33
1.12	Emission from Relativistic Particles	34
1.13	Problems	35
1.13.1	Speed and Velocity Transformation	35
1.13.2	Four-Acceleration	35
1.13.3	Geometry of Minkowski space	35
1.13.4	Relativistic decays	36
2	Dynamics	37
2.1	Lagrangian Dynamics	37
2.1.1	Generalized Coordinates	37
2.1.2	The Lagrangian and the Action	37
2.1.3	The Principle of Least Action	37
2.1.4	The Euler-Lagrange Equations	38
2.1.5	Example Lagrangians	39
2.2	Conservation Laws	39
2.2.1	Energy Conservation	39
2.2.2	Momentum Conservation	40
2.3	Coordinate Transformations	40
2.4	Hamilton's Equations	41
2.5	Adiabatic Invariance	42
2.6	Problems	43
2.6.1	Extremal paths	43
2.6.2	Schwarzschild Trajectories	44

2.6.3	Lagrangian electrodynamics	44
3	Random Fields	45
3.1	Descriptive Statistics	45
3.1.1	N -Point Distribution Functions	45
3.1.2	N -Point Correlation Functions	46
3.2	Two-point Correlation Function	46
3.3	Power Spectrum	46
3.4	Measuring the Power Spectrum	48
3.5	Moments of the Power Spectrum	50
3.6	Variance of Smoothed Fields	51
3.7	Power Law Power Spectra	51
3.8	Projections of Random Fields	53
3.9	Gaussian Random Fields	55
3.9.1	Central Limit Theorem	55
3.9.2	Multi-Variate Central Limit Theorem	56
3.9.3	Gaussian Fields	57
3.10	Gaussian N -point Distribution Functions	57
3.11	Gaussian Conditional Probabilities	57
3.12	Ricean Calculations	58
3.13	Variance of the Median	59
3.14	Problems	60
3.14.1	Radiation autocorrelation function	60
II	Radiation	61
4	Properties of Electromagnetic Radiation	63
4.1	Electromagnetic Spectrum	63
4.2	Macroscopic Description of Radiation	64
4.2.1	The Specific Intensity	64
4.2.2	Energy Flux	64
4.2.3	Momentum Flux	65
4.2.4	Inverse Square Law for Energy Flux	65
4.2.5	Specific Energy Density	65
4.2.6	Radiation Pressure	66
4.3	Constancy of Specific Intensity	67
5	Thermal Radiation	71
5.1	Thermodynamics of Black Body Radiation	71
5.1.1	Stefan-Boltzmann Law	71
5.1.2	Entropy of Black-Body Radiation	73
5.1.3	Adiabatic Expansion Laws	73
5.2	Planck Spectrum	73
5.2.1	Density of States	74
5.2.2	Mean Energy per State	75
5.2.3	Occupation Number in the Planck Spectrum	75
5.2.4	Specific Energy Density and Brightness	76
5.3	Properties of the Planck Spectrum	76
5.3.1	Raleigh-Jeans Law	76
5.3.2	Wien Law	77
5.3.3	Monotonicity with Temperature	77
5.3.4	Wien Displacement Law	77
5.3.5	Radiation Constants	77

5.4	Characteristic Temperatures	77
5.4.1	Brightness Temperature	77
5.4.2	Color Temperature	77
5.4.3	Effective Temperature	78
5.5	Bose-Einstein Distribution	78
5.6	Problems	78
5.6.1	Thermodynamics of Black-Body Radiation	78
5.6.2	Black body radiation and adiabatic invariance	78
5.6.3	Black-body radiation	79
5.6.4	Planck Spectrum	79
6	Radiative Transfer	81
6.1	Emission	81
6.2	Absorption	82
6.3	The Equation of Radiative Transfer	82
6.4	Kirchoff's Law	83
6.5	Mean Free Path	83
6.6	Radiation Force	83
6.7	Random Walks	84
6.8	Combined Scattering and Absorption	85
6.9	Rosseland Approximation	85
6.10	The Eddington Approximation	87
6.11	Einstein A , B Coefficients	88
6.11.1	Einstein Relations	88
6.11.2	Emission and Absorption Coefficients	89
6.12	Problems	90
6.12.1	Main sequence	90
6.12.2	Radiative Transfer.	90
6.12.3	Eddington luminosity	90
6.12.4	Poissonian statistics	91
7	Radiation Fields	93
7.1	Lorentz Force Law	93
7.2	Field Energy Density	93
7.3	Maxwell's Equations	93
7.4	Electromagnetic Waves	95
7.5	Radiation Power Spectrum	96
7.5.1	Polarization of Planar Waves	96
7.5.2	Polarization of Quasi-Monochromatic Waves	97
7.5.3	Power Spectrum Tensor and Stokes Parameters	97
7.6	Problems	99
7.6.1	Maxwell's equations	99
7.6.2	Energy and momentum of radiation field	100
8	Geometric Optics	101
8.1	Caustics	102
8.2	Random Caustics	104
8.2.1	Probability for Amplification	105
8.2.2	Caustics from Gaussian Deflections	105

9	Diffraction Theory	107
9.1	Fresnel Diffraction	108
9.2	Fraunhofer Diffraction	110
9.2.1	Babinet's Principle	111
9.3	Telescope Resolution	111
9.3.1	The Optical Transfer Function	112
9.3.2	Properties of the Telescope PSF and OTF	113
9.3.3	Random Phase Errors	113
9.4	Image Wander	118
9.5	Occultation Experiments	119
9.6	Scintillation	119
9.7	Transition to Geometric Optics	121
9.8	Problems	123
9.8.1	Diffraction.	123
9.8.2	Fraunhofer	123
9.8.3	Telescope PSF from Fresnel Integral	124
10	Radiation from Moving Charges	127
10.1	Electromagnetic Potentials	127
10.2	Retarded Potentials	128
10.3	Lienard-Wiechart Potentials	129
10.4	Dipole Radiation	132
10.5	Larmor's Formula	133
10.6	General Multi-pole Expansion	133
10.7	Thomson Scattering	135
10.8	Radiation Reaction	137
10.9	Radiation from Harmonically Bound Particles	138
10.10	Scattering by Bound Charges	138
10.11	Problems	139
10.11.1	Antenna beam pattern	139
10.11.2	Multipole radiation 1	139
10.11.3	Multipole radiation 2	140
10.11.4	Electron scattering	141
10.11.5	Thomson drag	141
10.11.6	Polarisation	141
11	Cerenkov Radiation	143
11.1	Retarded Potential	144
11.2	LW Potentials	145
12	Bremsstrahlung	149
12.1	Radiation from a Single Collision	149
12.2	Photon Discreteness	150
12.3	Single-Speed Electron Stream	151
12.4	Thermal Bremsstrahlung	151
12.5	Thermal Bremsstrahlung Absorption	152
12.6	Relativistic Bremsstrahlung	153
12.7	Applications of Thermal Bremsstrahlung	154
12.7.1	Low Frequency Emission from Ionized Gas Clouds	154
12.7.2	Clusters of Galaxies	154
12.7.3	Bremsstrahlung from High Energy Electrons	154
12.8	Problems	155
12.8.1	Bremsstrahlung	155

13 Synchrotron Radiation	157
13.1 Equations of Motion	157
13.2 Total Power Radiated	158
13.3 Synchrotron Cooling	158
13.4 Spectrum of Synchrotron Radiation	158
13.4.1 Pulse Profile	160
13.4.2 Low-Frequency Power Spectrum	163
13.5 Power-Law Electrons	165
14 Compton Scattering	167
14.1 Kinematics of Compton Scattering	167
14.2 Inverse Compton Effect	169
14.3 Inverse Compton Power	169
14.4 Compton vs Inverse Compton Scattering	172
14.5 The Compton y -Parameter	172
14.6 Repeated Scatterings	173
14.6.1 Non-Relativistic, High Optical Depth	173
14.6.2 Highly-Relativistic, Low Optical Depth	173
14.7 The Sunyaev-Zel'dovich Effect	173
14.8 Compton Cooling and Compton Drag	174
14.9 Problems	174
14.9.1 Compton scattering 2	174
14.9.2 Inverse Compton Effect	175
14.9.3 Compton y -parameter	175
III Field Theory	177
15 Field Theory Overview	179
16 Classical Field Theory	183
16.1 The BRS Model	183
16.2 The Continuum Limit	185
16.3 Conservation of Wave-Momentum	189
16.4 Energy and Momentum in the BRS Model	191
16.5 Covariance of the BRS Model	191
16.6 Interactions in Classical Field Theory	192
16.7 Wave-Momentum Puzzles	195
16.8 Conservation of 'Charge'	196
16.9 Conservation of Particle Number	199
16.10 Particle Number Conservation at Low Energies	201
16.11 Ideal Fluid Limit of Wave Mechanics	204
16.11.1 Local Average Stress-Energy Tensor	205
16.11.2 Evolution Equations	207
16.12 Discussion	209
17 Quantum Fields	211
17.1 The Simple Harmonic Oscillator	211
17.2 The Interaction Picture	213
17.2.1 The S -Matrix Expansion	214
17.2.2 Example: A Forced Oscillator	214
17.3 Free Fields	215
17.3.1 Discrete 1-Dimensional Lattice Model	215
17.3.2 Continuous 3-Dimensional Field	217
17.4 Interactions	217

17.4.1	Scattering off an Impurity	218
17.4.2	Self Interactions	219
17.4.3	Second Order Scattering	221
17.4.4	Contour Integral Formalism	226
17.4.5	Feynman Rules	227
17.4.6	Kinematics of Scattering	227
17.4.7	Discussion	227
17.5	Problems	228
17.5.1	Ladder Operators	228
18	Relativistic Field Theory	229
18.1	The Klein-Gordon Field	229
18.2	Quantum Electrodynamics	231
18.3	Connection to Kinetic Theory	232
18.4	The Scalar Field in an Expanding Universe	233
18.5	Non-Relativistic Scalar Fields	235
18.6	Problems	236
18.6.1	Stress-Energy Tensor	236
18.6.2	Klein-Gordon Field	236
18.6.3	Scalar Field Pressure	237
18.6.4	Domain Walls and Strings	237
IV	Matter	239
19	Kinetic Theory	241
19.1	The Collisionless Boltzmann Equation	241
19.2	The Boltzmann Transport Equation	242
19.3	Applications of the Transport Equation	244
19.3.1	Equilibrium Solutions	244
19.3.2	Boltzmann's H -Theorem	245
19.4	Conserved Quantities	246
19.4.1	Mass Conservation	246
19.4.2	Momentum Conservation	247
19.4.3	Energy Conservation	247
19.5	Fluid Equations	248
19.6	Problems	248
19.6.1	Boltzmann distribution	248
19.6.2	Kinetic theory and entropy	249
19.6.3	Kinetic theory	249
19.6.4	Massive neutrinos	249
20	Ideal Fluids	251
20.1	Adiabatic Flows	251
20.2	Hydrostatic Equilibrium	252
20.3	Convective Stability	252
20.4	Bernoulli's Equation	252
20.5	Kelvin's Circulation Theorem	253
20.6	Potential Flows	255
20.7	Incompressible Potential Flows	255
20.8	Gravity Waves	256
20.9	Sound Waves	258
20.10	Problems	259
20.10.1	Ideal fluids	259

20.10.2 Potential flows	259
20.10.3 Hydrostatic equilibrium	259
21 Viscous Fluids	261
21.1 Transport Coefficients	261
21.2 Damping of Sound Waves	263
21.3 Reynold's Number	264
21.4 Problems	264
21.4.1 Viscous hydrodynamics	264
21.4.2 Damping of Sound Waves	264
21.4.3 Sound waves	265
22 Fluid Instabilities	267
22.1 Rayleigh-Taylor and Kelvin-Helmholtz	267
22.2 Gravitational Instability	267
22.3 Thermal Instability	268
22.4 Turbulence	269
22.4.1 Kolmogorov Spectrum	269
22.4.2 Passive Additives	270
22.4.3 Inner Scale	270
22.4.4 Atmospheric Seeing	270
22.4.5 Stability	271
22.5 Problems	272
22.5.1 Kolmogorov turbulence	272
23 Supersonic Flows and Shocks	273
23.1 The de Laval Nozzle	273
23.2 Shock Waves	273
23.2.1 The Shock Tube	274
23.2.2 Vorticity Generation	276
23.2.3 Taylor-Sedov Solution	276
23.3 Problems	276
23.3.1 Collisional shocks	276
24 Plasma	277
24.1 Time and Length Scales	277
24.1.1 Plasma Frequency	277
24.1.2 Relaxation Time	278
24.1.3 Debye Length	278
24.2 Electromagnetic Waves in a Plasma	279
24.2.1 Dispersion in a Cold Plasma	279
24.2.2 Faraday Rotation	280
24.3 Problems	283
24.3.1 Dispersion Measure	283
V Gravity	285
25 The Laws of Gravity	287
25.1 General Relativity	287
25.2 Newtonian Gravity	287
25.3 Spherical Systems	288
25.3.1 Newton's Theorems	288
25.3.2 Circular and Escape Speed	288
25.3.3 Useful Spherical Models	289

26 Collisionless Systems	291
26.1 Relaxation Time	291
26.2 Jeans Equations	292
26.3 The Virial Theorem	292
26.4 Applications of the Virial Theorem	293
26.4.1 Spherical Collapse Model	293
26.4.2 Galaxy Cluster Mass to Light Ratios	293
26.4.3 Flat Rotation Curve Halos	293
26.5 Masses from Kinematic Tracers	294
26.6 The Oort Limit	295
26.7 Problems	295
26.7.1 Two-body Relaxation.	295
27 Evolution of Gravitating Systems	297
27.1 Negative Specific Heats	297
27.2 Phase Mixing	298
27.3 Violent Relaxation	299
27.4 Dynamical Friction	299
27.5 Collisions Between Galaxies	300
27.6 Tidal Stripping	300
VI Cosmology	301
28 Friedmann-Robertson-Walker Models	303
28.1 Newtonian Cosmology	303
28.2 Solution of the Energy Equation	304
28.3 Asymptotic Behavior	306
28.4 The Density Parameter	307
28.5 The Cosmological Redshift	308
28.6 The Horizon Problem	308
28.7 Cosmology with Pressure	309
28.8 Radiation Dominated Universe	312
28.9 Number of Quanta per Horizon Volume	312
28.10 Curvature of Space-Time	313
28.11 Problems	316
28.11.1 Energy of a Uniform Expanding Sphere	316
28.11.2 Solution of the FRW Energy Equation	316
29 Inflation	319
29.1 Problems with the FRW Models	319
29.2 The Inflationary Scenario	319
29.3 Chaotic Inflation	321
29.4 Discussion	325
29.5 Problems	327
29.5.1 Inflation	327
30 Observations in FRW Cosmologies	329
30.1 Distances in FRW Cosmologies	329
30.1.1 Scale Factor vs Hubble Parameter	329
30.1.2 Redshift vs Comoving Distance	329
30.2 Angular Diameter and Luminosity Distances	331
30.3 Magnitudes and Distance Moduli	334
30.4 K-Corrections	334

31 Linear Cosmological Perturbation Theory	337
31.1 Perturbations of Zero-Pressure Models	337
31.1.1 The Spherical ‘Top-Hat’ Perturbation	337
31.1.2 General Perturbations	340
31.2 Non-zero Pressure and the Jeans Length	343
31.2.1 Matter Dominated Era	343
31.2.2 Radiation Dominated Era	344
31.2.3 Super-Horizon Scale Perturbations	347
31.2.4 Isocurvature vs Isentropic Perturbations	347
31.2.5 Diffusive Damping and Free-Streaming	348
31.3 Scenarios	349
31.3.1 The Adiabatic-Baryonic Model	349
31.3.2 The Hot-Dark-Matter Model	351
31.3.3 The Cold Dark Matter Model	351
32 Origin of Cosmological Structure	355
32.1 Spontaneous Generation of Fluctuations	355
32.2 Fluctuations from Inflation	358
32.3 Self-Ordering Fields	361
32.3.1 Domain Walls	361
32.3.2 Cosmic Strings	365
33 Probes of Large-Scale Structure	369
33.1 Introduction	369
33.2 Galaxy Clustering	369
33.2.1 Redshift Surveys	370
33.2.2 Poisson Sample Model	370
33.2.3 Correlation Functions	371
33.2.4 The Power Spectrum	373
33.2.5 Redshift Space Distortion	374
33.2.6 Angular Clustering Surveys	376
33.3 Bulk-Flows	378
33.3.1 Measuring Bulk-Flows	378
33.4 Microwave Background Anisotropies	380
33.4.1 Recombination and the Cosmic Photosphere	380
33.4.2 Large-Angle Anisotropies	380
33.4.3 Small-Angle Anisotropies	381
33.4.4 Polarization of the CMB	381
33.5 Weak Lensing	382
34 Non-Linear Cosmological Structure	385
34.1 Spherical Collapse Model	385
34.2 Gunn-Gott Spherical Accretion Model	387
34.3 The Zel’dovich Approximation	388
34.4 Press-Schechter Mass Function	390
34.5 Biased Clustering	390
34.6 Self-Similar Clustering	391
34.7 Davis and Peebles Scaling Solution	393
34.8 Cosmic Virial Theorem	394

VII Appendices	395
A Vector Calculus	397
A.1 Vectors	397
A.2 Vector Products	397
A.3 Div, Grad and Curl	397
A.4 The Divergence Theorem	398
A.5 Stokes' Theorem	398
A.6 Problems	398
A.6.1 Vector Calculus Identities	398
B Fourier Transforms	401
B.1 Discrete Fourier Transform	401
B.2 Continuous Fourier Transform	402
B.3 Parseval's Theorem	403
B.4 Convolution Theorem	403
B.5 Wiener-Khinchin Theorem	403
B.6 Fourier Transforms of Derivatives and Integrals	403
B.7 Fourier Shift Theorem	404
B.8 Utility of Fourier Transforms	404
B.9 Commonly Occurring Transforms	404
B.10 The Sampling Theorem	405
B.11 Problems	407
B.11.1 Fourier Transforms	407
C The Boltzmann Formula	409
D Dispersive Waves	411
D.1 The Group Velocity	411
D.2 Wave Packets	413
D.3 Evolution of Dispersive Waves	415
E Relativistic Covariance of Electromagnetism	421
E.1 EM Field of a Rapidly Moving Charge	422
F Complex Analysis	425
F.1 Complex Numbers and Functions	425
F.2 Analytic Functions	425
F.3 Analytic Continuation	426
F.4 Contour Integration	427

List of Figures

1.1	Time Dilation	20
1.2	Length Contraction	22
1.3	Lorentz Transformation	23
1.4	Rotated Lorentz Transformation	24
1.5	Four Momentum	28
1.6	Relativistic Beaming	29
1.7	Relativistic Decays	31
1.8	Spatial Volume Element	33
2.1	Euler-Lagrange Equations	38
2.2	Adiabatic Invariance	42
3.1	Power-Spectrum and Auto-Correlation Function	48
3.2	Measuring the Power Spectrum	49
3.3	Fields with Power-Law Spectra	52
4.1	Specific Intensity	64
4.2	Net Flux	65
4.3	Radiation Energy Density	66
4.4	Radiation Pressure	67
4.5	Invariance of the Intensity	68
4.6	Invariance of the Intensity in a Telescope	68
5.1	A Cylinder Containing Radiation	72
5.2	Density of States	74
5.3	Planck Function	76
6.1	Einstein Coefficients	88
7.1	Maxwell's Equations	94
8.1	Formation of a Caustic	103
8.2	Generic Form of a Caustic	104
9.1	Huygens Wavelets	107
9.2	Fresnel $\cos(x^2)$ Function	109
9.3	Fresnel Knife-Edge	110
9.4	Babinet's Principle	111
9.5	Refracting Telescope	112
9.6	PSF for a Square Pupil	114
9.7	Circular Pupil PSF	115
9.8	Wavefront Deformation from Turbulence	116
9.9	Speckly PSF	117
9.10	Scintillation	121

9.11 Telescope with Aberration	122
9.12 Geometric vs Wave Optics	124
10.1 Source for Retarded Potentials	128
10.2 Lienard-Wiechart Potentials	131
10.3 Dipole Radiation	132
10.4 Bounded Charge Distribution	134
10.5 Quadrupole Radiation	135
10.6 Thomson Scattering	136
11.1 Conical Shock Wave	143
11.2 Cerenkov Radiation Geometry	145
11.3 Light Cone-Particle Intersection	146
11.4 Pulse of Cerenkov Radiation	147
12.1 Electron-Ion Collision	150
12.2 Thermal Bremsstrahlung Spectra	152
13.1 Critical Frequency for Synchrotron Radiation	159
13.2 Geometry for Synchrotron Spectrum	160
13.3 Observer Time vs Retarded Time	161
13.4 Retarded vs Observer Time	162
13.5 Potential for a Pulse of Synchrotron Radiation	163
13.6 Field for a Pulse of Synchrotron Radiation	163
13.7 Synchrotron Power Spectrum	164
14.1 Compton Scattering	168
14.2 Inverse Compton Effect	170
16.1 Coupled Oscillators	184
16.2 Dispersion Relation	185
16.3 BRS Phase and Group Velocity	188
16.4 The $\lambda\phi^4$ Interaction	193
16.5 The $\alpha\phi^2\chi$ Interaction	194
16.6 Interacting Complex Scalar Field	198
17.1 First Order Phonon-Phonon Scattering	221
17.2 Chion Decay and Production	221
17.3 Second Order Phonon-Phonon Scattering	223
17.4 Phonon-Phonon Scattering Diagrams	224
17.5 Phonon-Phonon Scattering Diagrams	225
17.6 Twisted Phonon-Phonon Scattering Diagram	225
17.7 Contour Integral for the Chion Propoagator	226
18.1 W-Potential	231
18.2 Compton Scattering	232
19.1 Two-Body Collision	243
20.1 Bernoulli Effect	253
20.2 Kelvin's Circulation Theorem	254
20.3 Gravity Wave Geometry	257
21.1 Shear Viscosity	262
22.1 Thermal Instability	269

22.2 Atmospheric Wavefront Corrugation	271
23.1 Shock Tube	274
24.1 Plasma Frequency	277
24.2 Faraday Rotation	282
27.1 Phase Mixing	298
28.1 FRW Scale Factor (linear plot)	305
28.2 FRW Scale Factor (logarithmic plot)	306
28.3 Density Parameter	307
28.4 Causal Structure of the FRW Model	309
28.5 Microwave Background Photon World-Lines	310
28.6 Mass of FRW Closed Model	316
28.7 Embedding Diagram	317
29.1 Horizon-Scale in Inflation	322
29.2 Chaotic Inflation Potential Function	323
30.1 Comoving Distance	331
30.2 Angular Diameter Distance	332
30.3 Angular Diameter Distance	332
30.4 Luminosity Distance	333
31.1 Spherical Top-Hat Perturbation	338
31.2 Isocurvature and Isentropic Perturbations	348
31.3 Adiabatic-Baryonic Model	350
31.4 Adiabatic-Baryonic Power Spectrum	350
31.5 Hot Dark Matter Power Spectrum	352
31.6 Cold Dark Matter Power Spectrum	353
32.1 Monopole, Dipole and Quadrupole Perturbations	357
32.2 Multiple Quadrupole Perturbations	358
32.3 Scalar Field Potential	362
32.4 Domain Wall Profile	363
32.5 Spontaneous Symmetry Breaking	364
32.6 2 Dimensional Scalar Field Potential	365
33.1 The Tully-Fisher Relation	379
33.2 Large-Angle CMB Anisotropies	381
33.3 Weak Lensing	383
34.1 Spherical Collapse Model	386
34.2 Gunn-Gott Accretion Model	387
34.3 Biased Clustering	392
34.4 Self-Similar Evolution	393
B.1 Common Fourier Transforms	405
B.2 Sampling Theorem	406
C.1 Boltzmann Law	410
D.1 Beating of 2 Sinusoids	412
D.2 Evolution of a Wave Packet	414
D.3 Fourier Transform of a Dispersive Wave	415

D.4	Gravity Wave Chirp	418
D.5	Gravity Wave Swell	419
E.1	Electric Field of a Moving Charge	423
F.1	Analytic Continuation	426
F.2	Contour Integral	428

Part I

Preliminaries

Chapter 1

Special Relativity

A remarkable feature of Maxwell's equations is that they support waves with a unique velocity c , yet there is no 'underlying medium' with respect to which this velocity is defined (in contrast to say sound waves in a physical medium). An equally remarkable observational fact is that the velocity of propagation of light is indeed independent of the frame of reference of the observer or of the source (*Michelson and Morley experiment*). Searches for the expected *aether drift* proved unsuccessful. These results would seem to conflict with *Galilean relativity* in which there is a universal time, and universal Cartesian spatial coordinates such that each event can be assigned coordinates on which all observers can agree. Einstein's *special theory of relativity* makes sense of these results. The result is a consistent framework in which events in space-time are assigned coordinates, but where the coordinates depend on the state of motion of the observer. The situation is rather analogous to that in planar geometry, where the coordinates of a point depend on the origin and rotation of ones chosen frame of reference. However, one can also use vector notation to express relations between lines and point — e.g. $\mathbf{a} + \mathbf{b} = \mathbf{c}$ — which are valid for all frames of reference. In special relativity the fundamental quantities are points, or 'events', which are vectors in a 4-dimensional space-time. We will see how these '4-vectors' transform under changes in the observer's frame of reference, and how particle velocities, momenta and other physical quantities can be expressed in the language of 4-vector notation. Indeed the fundamental principle of relativity is that *all* of the laws of physics can be expressed in a frame invariant manner. The last part of the chapter deals with the transformation properties of distribution functions (e.g. the density of particles in space, or the distribution of particles over energy, velocity etc).

1.1 Time Dilation

An immediate consequence of the frame-independence of the speed of light is that observers in relative motion with respect to one another must assign different time separations to events.

Consider an observer A with a simple *gedanken* clock consisting of a photon bouncing between mirrors attached to the ends of a standard rod of length l_0 as illustrated in figure 1.1. One round trip of the photon takes an interval $\Delta t_0 = 2l_0/c$ in the 'rest-frame' of the clock. Now consider the same round trip as seen from the point of view of an observer B moving with some relative velocity v in a direction perpendicular to the rod.

First, note that A and B must assign the same length to the rod. To see this imagine B carries an identical rod, also perpendicular to his direction of motion, with pencils attached which make marks on A's rod as they pass. Since the situation is completely symmetrical, the marks on A's rod must have the same separation as the pencils on B's. Thus transverse spatial dimensions are independent of the frame of state of the observer.

From B's point of view then the distance traveled by the photon in one round trip must exceed $2l_0$, and consequently the time interval between the photon's departure and return is $\Delta t > \Delta t_0$. In this time, A's rod has moved a distance $v\Delta t$, so, by Pythagoras' theorem the total distance traveled

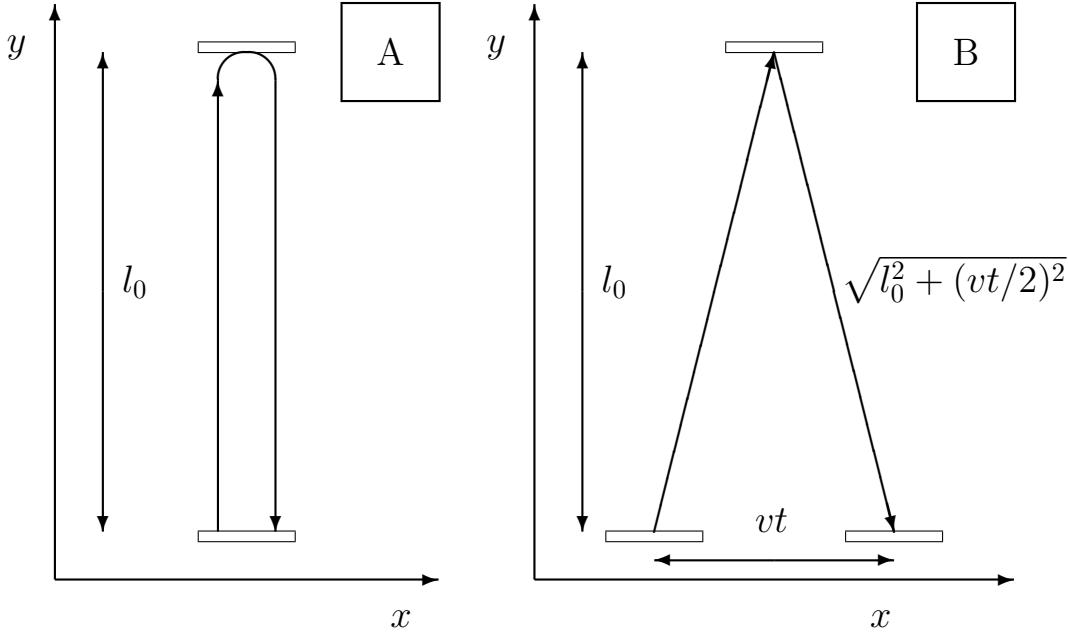


Figure 1.1: Illustration of time-dilation. Left panel shows one click of the gedanken clock in A's frame of reference. Right panel shows the path of the photon as seen by a moving observer. Evidently, if the clock is moving the photon has to travel further, so the interval between clicks for a moving clock is greater than if the clock is at rest. The *time dilation factor* is $\gamma = 1/\sqrt{1 - v^2/c^2}$.

by the photon is $2\sqrt{l_0^2 + (v\Delta t/2)^2} = c\Delta t$ and solving for Δt gives

$$\Delta t = \Delta t_0 / \sqrt{1 - v^2/c^2} \quad (1.1)$$

or, defining the *Lorentz gamma-factor*

$$\gamma \equiv 1/\sqrt{1 - v^2/c^2} \quad (1.2)$$

we have the *time dilation formula*

$$t = \gamma t_0. \quad (1.3)$$

Thus *moving clocks run slow* by the factor γ . This is a small correction for low velocities, but becomes very large as the velocity approaches c . This behavior is not paradoxical, since the situation is not symmetrical; the two events (departure and return) occur at the same point of space in A's frame of reference, whereas they occur at different positions in B's frame. The spatial separation of the two events in B's frame is $\Delta x = v\Delta t$, and the temporal separation is $\Delta t = \Delta t_0 / \sqrt{1 - v^2/c^2}$ and so we have

$$\Delta t^2 - \Delta x^2/c^2 = \Delta t_0^2. \quad (1.4)$$

The quantity

$$\Delta t_0 = \sqrt{\Delta t^2 - \Delta x^2/c^2} \quad (1.5)$$

is called the *proper time* interval between two events. It is *invariant*; ie it is the same for all observers in a state of constant relative motion, even though they assign different spatial and temporal intervals to the separation of the events. It is equal to the temporal separation of the events as measured

by an observer for whom the two events occur at the same point in space. Any other observer will assign a greater temporal separation.

1.2 Length Contraction

We can derive the *Lorentz-Fitzgerald length contraction* formula in a very similar fashion. Let us equip A with a pair of clocks mounted back to back so that a pair of photons repeatedly depart from A, travel equal and opposite distances l_0 , bounce off mirrors, and then return to A. A space-time diagram of one cycle of the clock as perceived by A is shown on the left hand side of figure 1.2.

The same set of events as perceived by an observer B now moving with constant velocity *parallel* to the arms of the clock is shown on the right for the case of $v = c/2$. Set the origin of coordinates at the emission point, and let the length of the arms in B's rest frame be l . The two photons propagate along trajectories $x = \pm ct$ until they reach the mirrors, which have world lines $x = \pm l + vt$. Solving for the reflection times t_{\pm} and locations gives

$$ct_{\pm} = \pm x_{\pm} = \frac{l}{1 \mp v/c}. \quad (1.6)$$

The return flight of each photon is the same as the outward flight of the other photon, so the total time Δt elapsed between departure and return to A satisfies

$$c\Delta t = c(t_+ + t_-) = l \left[\frac{1}{1 - v/c} + \frac{1}{1 + v/c} \right] = \frac{2l}{1 - v^2/c^2} = 2\gamma^2 l. \quad (1.7)$$

However, we know that $\Delta t = \gamma\Delta t_0 = 2\gamma l_0/c$ and hence we obtain the *Lorentz-Fitzgerald length contraction formula*

$$l = l_0/\gamma. \quad (1.8)$$

This is sometimes stated as *moving rods appear foreshortened*. More precisely, we have shown that two events which occur at the same time in one frame and have separation l in that frame will have a spatial separation in a relatively moving frame of $l_0 = \gamma l > l$. Consider, for example the reflection events. In A's frame these occur at the same time, so $\Delta t_0 = 0$, and have spatial separation $\Delta x_0 = 2l_0$. In B's frame however, they have temporal separation (times c) of

$$c\Delta t = c(t_+ - t_-) = 2\gamma^2 lv/c = 2\gamma l_0 v/c \quad (1.9)$$

and spatial separation

$$\Delta x = x_+ - x_- = 2\gamma^2 l = 2\gamma l_0. \quad (1.10)$$

Evidently

$$\Delta x^2 - c^2 \Delta t^2 = 4l_0^2 \gamma^2 (1 - v^2/c^2) = (2l_0)^2 = \Delta x_0^2 \quad (1.11)$$

is also an invariant. The quantity

$$\Delta x_0 = \sqrt{\Delta x^2 - c^2 \Delta t^2} \quad (1.12)$$

is known as the *proper distance* between the two events.

Finally, the area of the region enclosed by the photon world lines is, in A's frame, $A_0 = (\sqrt{2}l_0)^2$ whereas in B's frame

$$A = (\sqrt{2}x_+)(\sqrt{2}x_-) = \frac{2l_0^2}{\gamma^2(1 - v^2/c^2)} = A_0. \quad (1.13)$$

so this area is an invariant. Since transverse dimensions are invariant, this means that the *space-time 4-volume* is an invariant.

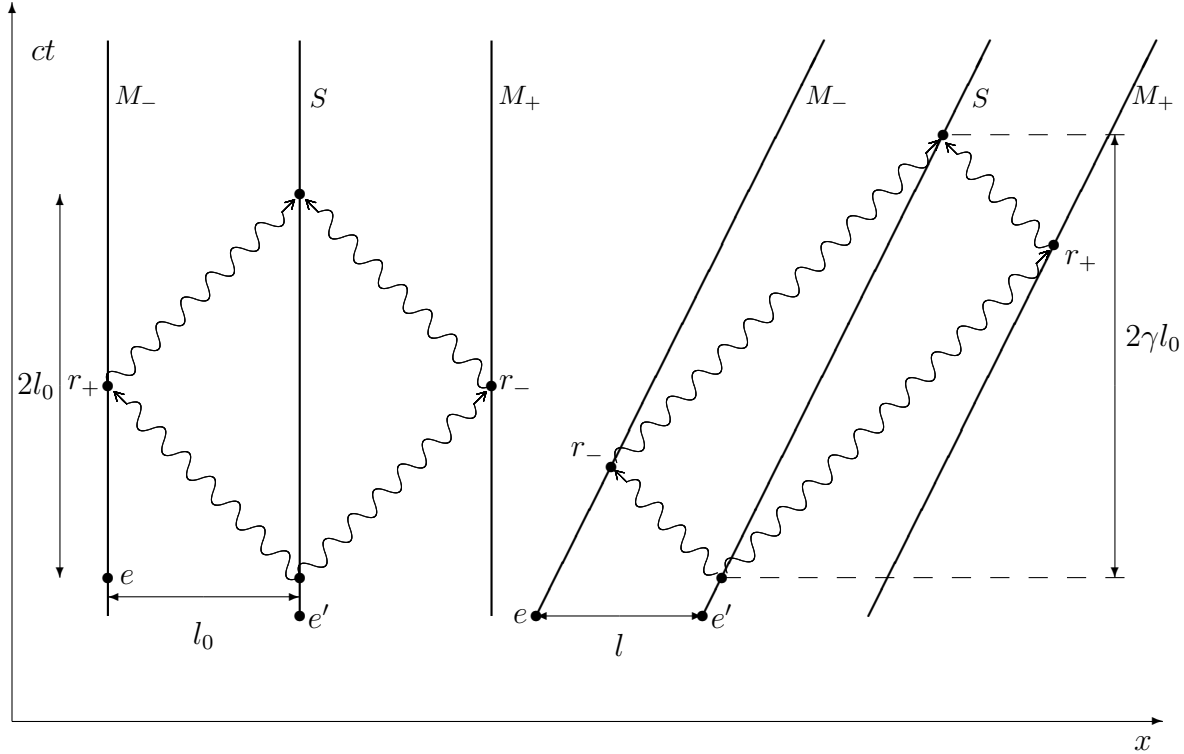


Figure 1.2: Space-time diagram used to derive length contraction. On the left are shown the photon trajectories (wiggly diagonal lines) departing from the source S (whose world-line is the central vertical line) and reflecting from mirrors M_- , M_+ (also with vertical world-lines) and returning to the source S . This is for a stationary clock, and the interval between clicks of the clock — the time between departure and return of the photons — is $c\Delta t = 2l_0$. The slanted lines on the right show the world lines for the source/receiver and mirrors for a clock which is moving at a constant velocity ($v = c/2$ in this case). Since the speed of light is invariant, the photons still move along 45-degree diagonal lines. Now we have already seen that the interval between clicks for a moving clock is larger than that for a stationary clock by a factor γ . This means that the time between emission and return for the moving clock is $c\Delta t = 2\gamma l_0$. It is then a matter of simple geometry (see text) to show that the distance between the source and the mirrors l is *smaller* than that for the stationary clock by a factor γ , or $l = l_0/\gamma$. This is the phenomenon of relativistic length contraction; if we have a metre rod moving in a direction parallel to its length then *at a given time* the distance between the ends of the rod is less than 1 metre by a factor $1/\gamma$. The above description is of two different clocks, viewed in a single coordinate system. There is a different, and illuminating, alternative way to view the above figure. We can think of these two pictures as being of the *same* clock — and indeed the very same set of emission, reflection and reception events — but as viewed from two different frames of reference. The left hand picture shows the events as recorded by an observer who sees the clock as stationary while the right hand picture is the events as recorded by an observer moving at velocity $v = c/2$ with respect to the clock. Now consider the reflection events, labelled r_- and r_+ . In the clock-frame these events have spatial separation $\Delta x = 2l_0$, while in the moving frame simple geometrical analysis shows that the spatial separation is $\Delta x = 2\gamma l_0$. Now we are saying that the separation of the mirrors is *larger* for the moving clock, whereas before we were saying the moving clock's rods were contracted. This sounds contradictory, or paradoxical, but it isn't really. The resolution of the apparent paradox is that the situation is again non-symmetrical between the two frames. The reflection events occur at the same time in the clock's rest frame, and the separation is the so-called 'proper-separation' $\Delta x_0 = 2l_0$. In the moving frame the two events have a time coordinate difference $\Delta t \neq 0$, and the spatial separation, as we show in the text, is now $\Delta x = \sqrt{\Delta x_0^2 + c^2 \Delta t^2}$. In the earlier discussion we were computing the distance between the two events e, e' which occur same time *in the moving frame*. These events in the rest frame do not occur at the same coordinate time.

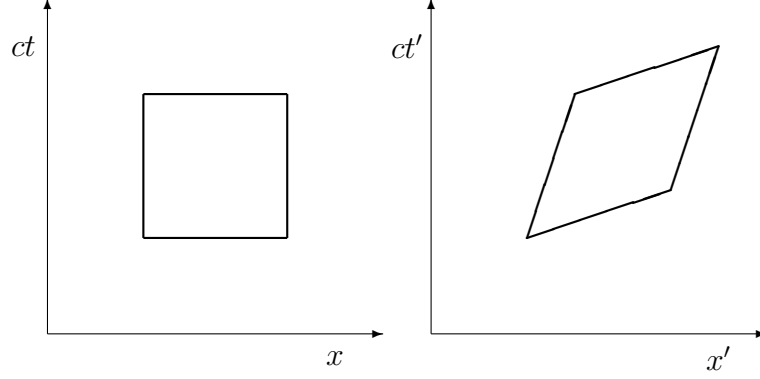


Figure 1.3: The Lorentz transformation causes a shearing in the $x-t$ space. This was shown above for an area bounded by null curves, but the result is true for arbitrary areas.

1.3 Lorentz Transformation

Figure 1.2 shows that the effect of a boost on the area in the $x-ct$ plane bounded by the photon paths is to squash it along one diagonal direction and to stretch it along the other. The same is true for any area, as illustrated in figure 1.3. In fact, one can write the transformation for the spatial coordinates as multiplication by a matrix, whose coefficients are a function of the boost velocity.

In this section it will prove convenient to work in units such that $c = 1$ (or equivalently let $t' = ct$ and drop the prime), so photon *world lines* are diagonals in $x-t$ space.

Let's now determine the form of this transformation matrix for the case of a boost along the x -axis. For such a boost, we know that the y and z -coordinates are unaffected, so we need only compute how the x and t coordinates are changed. Consider first what happens if we take the $x-t$ plane and rotate it by 45 degrees. Specifically, let's define new coordinates

$$\begin{bmatrix} T \\ X \end{bmatrix} = R(45^\circ) \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} \cos(45^\circ) & -\sin(45^\circ) \\ \sin(45^\circ) & \cos(45^\circ) \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix}. \quad (1.14)$$

Now we saw in the previous section that in this rotated frame the effect of a boost is just a stretch in the horizontal direction with scale factor $S_+ = x_+/l_0 = 1/\gamma(1-v)$ and a contraction in the vertical direction with scale factor $S_- = x_-/l_0 = 1/\gamma(1+v)$. The effect of the boost on position vectors in this 45-degree rotated system is just multiplication by the 2×2 matrix

$$M = \begin{bmatrix} S_+ & 0 \\ 0 & S_- \end{bmatrix}. \quad (1.15)$$

or, denoting coordinates in the boosted frame by superscript,

$$\begin{bmatrix} T' \\ X' \end{bmatrix} = \begin{bmatrix} S_+ & 0 \\ 0 & S_- \end{bmatrix} \begin{bmatrix} T \\ X \end{bmatrix} = \begin{bmatrix} S_+ T \\ S_- X \end{bmatrix}. \quad (1.16)$$

The effect of these linear transformations is illustrated in figure 1.4.

So far we have obtained the linear transformation matrix for transforming the rotated $X-T$ coordinates. What we really want is the matrix that transforms un-rotated $x-t$ coordinates. This is readily found since we have

$$\begin{bmatrix} t' \\ x' \end{bmatrix} = R^{-1} \begin{bmatrix} T' \\ X' \end{bmatrix} = R^{-1} M \begin{bmatrix} T \\ X \end{bmatrix} = R^{-1} M R \begin{bmatrix} t \\ x \end{bmatrix} \quad (1.17)$$

with $R = R(45^\circ)$ the rotation matrix for a 45 degree rotation. Evidently, the transformation from $x-t$ to boosted $x'-t'$ coordinates is effected by multiplying by the matrix $M' = R^{-1} M R$, or

$$M' = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} S_+ & 0 \\ 0 & S_- \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} S_+ + S_- & -S_+ + S_- \\ -S_+ + S_- & S_+ + S_- \end{bmatrix}. \quad (1.18)$$

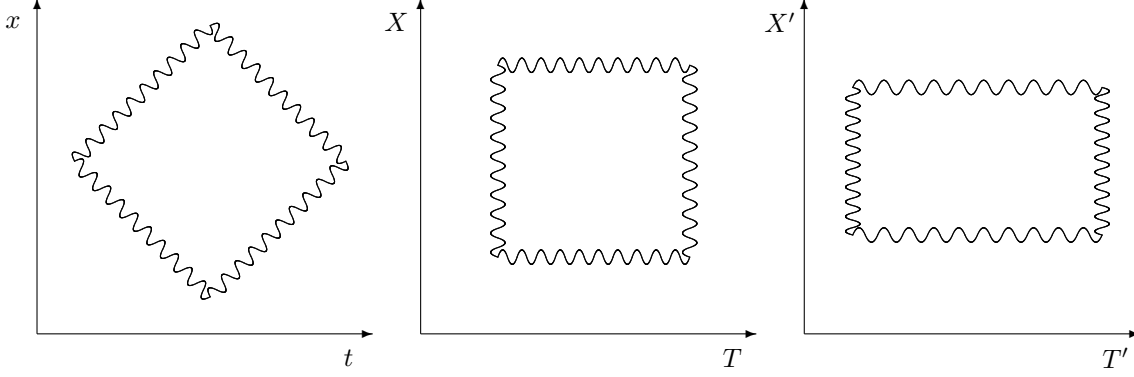


Figure 1.4: A region of 2-dimensional $x - t$ space-time bounded by photon world lines is shown in the left hand panel. The center panel shows the same region in $X - T$ coordinates, which are just $x - t$ coordinates rotated through 45° . The right panel shows the same region after applying a boost along the x -axis.

but $S_+ + S_- = 2\gamma$ and $S_+ - S_- = 2\gamma\beta$ where $\beta = v/c$. Therefore the transformation of $x - t$ coordinate vectors induced by a boost of dimensionless velocity β is

$$\begin{bmatrix} t' \\ x' \end{bmatrix} = M' \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} \gamma(t - \beta x) \\ \gamma(x - \beta t) \end{bmatrix}. \quad (1.19)$$

Finally, recalling that transverse dimensions y, z are unaffected by a boost in the x -direction we obtain the full transformation as a 4×4 matrix multiplication

$$\begin{bmatrix} t' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} t \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \gamma(t - \beta x) \\ \gamma(x - \beta t) \\ y \\ z \end{bmatrix}. \quad (1.20)$$

This is known as the *Lorentz transformation*.

1.4 Four-vectors

The *prototype 4-vector* is the separation between two space-time events

$$x^\mu = \begin{bmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{bmatrix} = \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} \quad (1.21)$$

which transforms under a boost $v/c = \beta$ as

$$x'^\mu = \Lambda^\mu{}_\nu x^\nu \quad (1.22)$$

where the 4×4 transformation matrix $\Lambda^\mu{}_\nu$ is

$$\Lambda^\mu{}_\nu = \begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \quad (1.23)$$

Summation of repeated indices is implied, and such summations should generally involve one subscript index and one superscript index.

Alternative notation for four vectors is

$$x^\mu = \vec{x} = (ct, \mathbf{x}) = (ct, x^i) \quad (1.24)$$

where $i = 1, 2, 3$.

We have seen that the Lorentz transformation matrix Λ corresponds to a diagonal shearing in the $x - t$ subspace. The determinant of Λ is unity, in accord with the invariance of space-time volume previously noted.

This is for a boost along the x -axis. The transformation law for a boost in another direction can be found by multiplying matrices for a spatial rotation and a boost. There are also generalizations of (1.20) which allow for reflections of the coordinates (including time). See any standard text for details.

The vector x^μ is a *contravariant* vector. It is also convenient to define a *covariant* 4-vector which is equivalent, but is defined as

$$x_\mu = \begin{bmatrix} -ct \\ x \\ y \\ z \end{bmatrix} \quad (1.25)$$

with a subscript index to distinguish it.

The two forms of 4-vector can be transformed into each other by multiplying by a 4×4 matrix called the *Minkowski metric*

$$\eta_{\mu\nu} = \eta^{\mu\nu} = \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}. \quad (1.26)$$

since clearly $x^\mu = \eta^{\mu\nu} x_\nu$ and $x_\mu = \eta_{\mu\nu} x^\nu$.

There is a version of the Lorentz matrix $\tilde{\Lambda}_\mu{}^\nu$ that transforms covariant 4-vectors as $x'_\mu = \tilde{\Lambda}_\mu{}^\nu x_\nu$ which is related to $\Lambda^\mu{}_\nu$ by

$$\tilde{\Lambda}_\mu{}^\nu = \eta_{\mu\tau} \Lambda^\tau{}_\sigma \eta^{\sigma\nu}. \quad (1.27)$$

The *norm* of the 4-vector \vec{x} is defined as

$$s^2 = \vec{x} \cdot \vec{x} = x^\mu x_\mu = -c^2 t^2 + x^2 + y^2 + z^2 = \mathbf{x} \cdot \mathbf{x} - c^2 t^2 \quad (1.28)$$

which we recognize as the invariant *proper separation* of the events. It can be computed as $s^2 = \eta_{\mu\nu} x^\mu x^\nu$ etc. If the norm is positive, negative or zero the separation is said to be ‘space-like’, ‘time-like’ and ‘null’ respectively. Since the norm is invariant, a separation which is space-like in one frame will be space-like in all inertial frames etc.

A 4-component entity \vec{A} is a *four-vector* if it transforms in the same way as \vec{x} under boosts (as well as spatial transformations, rotations etc).

The *scalar product* of two four vectors \vec{A}, \vec{B} is defined as

$$\vec{A} \cdot \vec{B} = A^\mu B_\mu = -A^0 B^0 + \mathbf{A} \cdot \mathbf{B} \quad (1.29)$$

and it is easy to show that the scalar product is invariant under Lorentz transformations.

The *gradient operator* in space-time is a covariant vector since we require that the difference between the values of some scalar quantity f at two neighboring points

$$df = d\vec{x} \cdot \vec{\nabla} f = dx^\mu \frac{\partial}{\partial x^\mu} f \quad (1.30)$$

should transform as a scalar (ie be invariant). We often write the gradient operator as $\vec{\nabla} = \partial_\mu \equiv \partial/\partial x^\mu$. We will also use the notation $\partial_\mu y = y_{,\mu}$ to denote partial derivatives with respect to space time coordinates.

A 4×4 matrix $T^{\mu\nu}$ is a *contravariant rank-2 tensor* if its components transform in the same manner as $A^\mu B^\nu$. Covariant rank-2 tensors $T_{\mu\nu}$, or mixed rank-2 tensors $T_\mu{}^\nu$ are defined similarly, as are higher rank tensors.

- Example rank-2 tensors are $\eta_{\mu\nu}$ and $\delta^\mu{}_\nu$ (the Kronecker δ -symbol). These are both constant; the components have the same numerical value on all inertial frames. Other examples are the *outer product* of a pair of vectors $A^\mu B^\nu$, and the gradient of a vector field $\partial_\mu A^\nu$.
- Tensors can be added, so $A^{\mu\nu} + B^{\mu\nu}$ is a tensor.
- Higher order tensors can be obtained by taking outer products of tensors, such as $T^{\mu\nu\sigma\tau} = A^{\mu\nu} B^{\sigma\tau}$.
- Indices can be raised and lowered with the Minkowski metric.
- Pairs of identical indices can be summed over to construct tensors, vectors of lower rank by *contraction*. For example, one can make a vector by contracting a (mixed) rank-3 tensor $A^\mu = T^{\mu\nu}{}_\nu$. It is important that one contract on one ‘upstairs’ and one ‘downstairs’ index. If necessary, one should raise or lower an index with the Minkowski metric.
- The fundamental *principle of special relativity* is that *all* of the laws of physics can be expressed in terms of 4-vectors and tensors in an invariant manner.

1.5 The 4-velocity

The coordinates of a particle are a 4-vector, as is the difference of the coordinates at two points along its world-line. For two neighboring points or ‘events’, we can divide by the proper-time $d\tau$ between the events, ie the interval between the events as measured by an observer moving with the particle and which is a scalar, to obtain the *4-velocity*

$$\vec{U} = \frac{dx^\mu}{d\tau} \quad (1.31)$$

which is a contravariant 4-vector.

If the particle has 3-velocity \mathbf{u} relative to our inertial frame, then the two events in our frame have temporal coordinate separation $dt = \gamma_{\mathbf{u}} d\tau$, with $\gamma_{\mathbf{u}} \equiv 1/\sqrt{1 - \mathbf{u} \cdot \mathbf{u}/c^2}$ as usual, and hence the particle’s 4-velocity is related to its coordinate velocity by $U^0 = dx^0/d\tau = c dt/d\tau = c\gamma_{\mathbf{u}}$ and $U^i = dx^i/d\tau = \gamma_{\mathbf{u}} dx^i/dt$ or

$$\vec{U} = \gamma_{\mathbf{u}} \begin{bmatrix} c \\ \mathbf{u} \end{bmatrix}. \quad (1.32)$$

If we undergo a boost along the x -axis of velocity $\beta = v/c$ into some new inertial frame then the components of the particle’s 4-velocity transform as

$$\begin{aligned} U'^0 &= \gamma(U^0 - \beta U^1) \\ U'^1 &= \gamma(U^1 - \beta U^0) \\ U'^2 &= U^2 \\ U'^3 &= U^3 \end{aligned} \quad (1.33)$$

These relations can be used to show how speeds and velocities of particles transform under boosts of the observer’s frame of reference as follows: The first of equations (1.33) with $U^0 = \gamma_{\mathbf{u}} c$, $U^1 = \gamma_{\mathbf{u}} u^1$ etc and with x -component of the coordinate velocity in the unprimed frame $u^1 = u \cos \theta$ gives

$$\gamma_{\mathbf{u}'} = \gamma_{\mathbf{u}} \left(1 - \frac{uv}{c^2} \cos \theta \right) \quad (1.34)$$

which allows one to transform the particles *Lorentz factor* $\gamma_{\mathbf{u}}$, and therefore also the particle’s speed $|\mathbf{u}|$, under changes in inertial frame.

The second of equations (1.33) gives $\gamma_{\mathbf{u}'} u'^1 = \gamma(\gamma_{\mathbf{u}} u^1 - \beta c \gamma_{\mathbf{u}}) = \gamma_{\mathbf{u}}(u^1 - v)$ or

$$u'^1 = \frac{u^1 - v}{1 - vu^1/c^2} \quad (1.35)$$

which is the transformation law for the coordinate velocity.

If the particle is moving along the x -axis at the speed of light in the unprimed frame ($u^1 = c$) then the velocity in the unprimed frame is $u'^1 = (c - v)/(1 - v/c) = c$. This is in accord with the constancy of the speed of light in all frames.

Finally, in the rest frame of the particle, $\vec{U} = (c, 0)$, so dotting some vector with a particle's 4-velocity is a useful way to extract the time component of the 4-vector as seen in the particle's frame of reference.

1.6 The 4-acceleration

The *four-acceleration* is

$$\vec{A} = \frac{d\vec{U}}{d\tau}. \quad (1.36)$$

and is another 4-vector.

The scalar product of the 4-acceleration and the 4-velocity is $\vec{A} \cdot \vec{U} = d(\vec{U} \cdot \vec{U})/d\tau$ which vanishes because the squared length of a 4-vector is $\vec{U} \cdot \vec{U} = -c^2$, which is invariant. Thus the four acceleration is always orthogonal to the 4-velocity.

In terms of the coordinate 3-velocity, the 4-acceleration is

$$\vec{A} = \gamma(d(\gamma c)/dt, d(\gamma \mathbf{u})/dt). \quad (1.37)$$

and a little algebra gives the norm of the 4-acceleration in terms of the particle's coordinate acceleration and coordinate velocity as

$$\vec{A} \cdot \vec{A} = \gamma^4(\mathbf{a} \cdot \mathbf{a} + \gamma^2(\mathbf{u} \cdot \mathbf{a}/c)^2). \quad (1.38)$$

In the particle's rest-frame $\vec{U} = (c, 0, 0, 0)$ so $A^0 = 0$, and therefore the norm is just equal to the square of the *proper acceleration*: $\vec{A} \cdot \vec{A} = |\mathbf{a}_0|^2$ so (1.38) gives the acceleration felt by a particle in terms of the coordinate acceleration in the observer's frame of reference.

If we decompose the 3-acceleration into components \mathbf{a}_\perp and \mathbf{a}_\parallel which are perpendicular and parallel to the velocity vector \mathbf{u} it is easy to show that

$$\vec{A} \cdot \vec{A} = |\mathbf{a}_0|^2 = \gamma^4(a_\perp^2 + \gamma^2 a_\parallel^2) \quad (1.39)$$

Of particular interest is the case $\mathbf{a} = \mathbf{a}_\perp$, as is the case for a particle being accelerated by a static magnetic field. In that case, the rest-frame acceleration is larger than in the 'lab-frame' by a factor γ^2 . This is easily understood. Observers in different inertial frames agree on the values of transverse distances as these are not affected by the Lorentz boost matrix. The second time derivative of the transverse position of the particle is larger in the instantaneous rest-frame than in the lab-frame, simply because time runs faster, by a factor γ , in the rest-frame. This will prove useful when we want to calculate relativistic synchrotron radiation.

1.7 The 4-momentum

Multiplying the 4-velocity of a particle by its rest mass m (another invariant) gives the *four-momentum*

$$\vec{P} = m\vec{U} = \gamma m(c, \mathbf{u}). \quad (1.40)$$

The spatial components of the 4-momentum differ from the non-relativistic form by the factor γ . To see why this is necessary consider the situation illustrated in figure 1.5.

Some texts use the notation m_0 for the rest-mass and set $m = \gamma m_0$. The space components of the relativistic 4-momentum are then $\mathbf{P} = m\mathbf{u}$, just as in non-relativistic mechanics. We do not follow that convention.

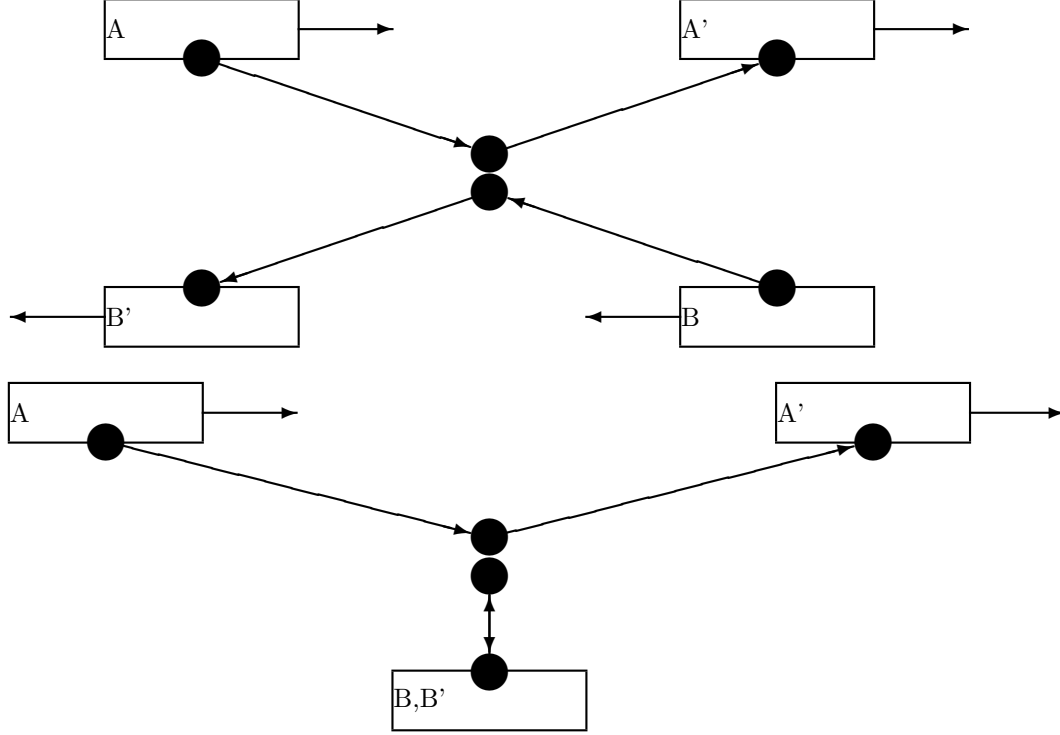


Figure 1.5: Illustration of the relativistic form for the momentum. Two observers A and B pass each other on rapidly moving carriages and as they do so they bounce balls off each other, exchanging momentum. The upper panel shows the symmetric situation in the center of mass frame. The lower panel shows the situation from B's point of view. Now B assigns a longer time interval to the pair of events A and A' than to B, B' while transverse distances are invariant so it follows that the transverse velocity he assigns to A's ball is lower than his own by a factor γ . Thus, in B's frame mu_x is not conserved, but γmu_x is conserved in the collision.

The time component of the 4-momentum is

$$cP^0 = \gamma mc^2 = \frac{mc^2}{\sqrt{1 - v^2/c^2}} = mc^2 + \frac{1}{2}mv^2 + \dots \quad (1.41)$$

which, aside from the constant mc^2 coincides with the kinetic energy for low velocities, and we call $cP^0 = E$ the total energy. The 4-momentum is

$$\vec{P} = (E/c, \mathbf{P}). \quad (1.42)$$

The 4-momentum for a massive particle is a time-like vector and its invariant squared length is

$$E^2/c^2 - \mathbf{P} \cdot \mathbf{P} = m^2 c^2. \quad (1.43)$$

Massive particles are said to 'live on the mass-shell' in 4-momentum space.

All these quantities and relations are well-defined in the limit $m \rightarrow 0$. For massless particles $E^2 = |\mathbf{P}|^2 c^2$, and with $E = \hbar\omega$, $\mathbf{P} = \hbar\mathbf{k}$ the 4-momentum is then

$$\vec{P} = \hbar(\omega/c, \mathbf{k}). \quad (1.44)$$

The total 4-momentum for a composite system is the sum of the 4-momenta for the component parts, and all 4 components are conserved. Note that the mass of a composite system is not the sum of the masses of the components, since the total mass contains, in addition to the rest mass, any energy associated with internal motions etc.

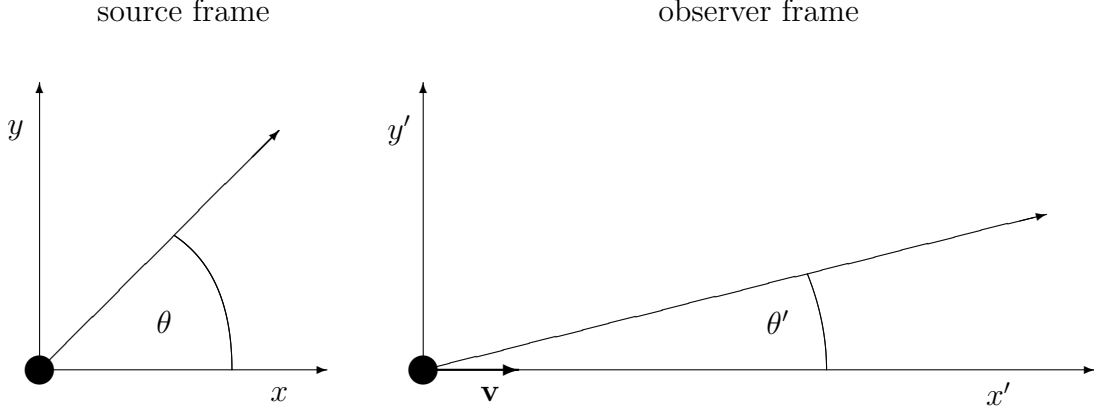


Figure 1.6: Left panel shows a photon emitted from a source as seen in the rest-frame. Right panel shows the situation in the observer frame in which the source has velocity $\mathbf{v} = v\hat{\mathbf{x}}$.

1.8 Doppler Effect

Consider a photon which in the frame of some observer has 4-momentum

$$\vec{P}_0 = \frac{E_0}{c} \begin{bmatrix} 1 \\ \cos \theta_0 \\ \sin \theta_0 \\ 0 \end{bmatrix}. \quad (1.45)$$

If the emitter is moving with velocity $\mathbf{v} = v\hat{\mathbf{x}}$ with respect to the observer then the 4-momentum in the emitter's frame is

$$\vec{P} \equiv \frac{E}{c} \begin{bmatrix} 1 \\ \cos \theta \\ \sin \theta \\ 0 \end{bmatrix} = \Lambda \vec{P}_0 = \frac{E_0}{c} \begin{bmatrix} \gamma(1 - \beta \cos \theta_0) \\ \gamma(\cos \theta_0 - \beta) \\ \sin \theta_0 \\ 0 \end{bmatrix}. \quad (1.46)$$

The observed energy is therefore related to the energy in the emitter's frame by

$$E_0 = \frac{E}{\gamma(1 - \beta \cos \theta_0)} \quad (1.47)$$

which is the *Doppler formula*.

1.9 Relativistic Beaming

Consider a source which emits radiation isotropically in its rest frame. What is the angular distribution of radiation in some other inertial frame?

Let a particular photon have source-frame 4-momentum

$$\vec{P} = \frac{E}{c} \begin{bmatrix} 1 \\ \cos \theta \\ \sin \theta \cos \phi \\ \sin \theta \sin \phi \end{bmatrix} \quad (1.48)$$

and let the source have velocity $\mathbf{v} = v\hat{\mathbf{x}}$ in the observer frame. The observer therefore has velocity $\mathbf{v} = -v\hat{\mathbf{x}}$ in the source-frame, and so the photon 4-momentum in the observer frame (primed frame)

is

$$\vec{P}' = \frac{E'}{c} \begin{bmatrix} 1 \\ \cos \theta' \\ \sin \theta' \cos \phi' \\ \sin \theta' \sin \phi' \end{bmatrix} = \frac{E}{c} \begin{bmatrix} \gamma(1 + \beta \cos \theta) \\ \gamma(\cos \theta + \beta) \\ \sin \theta \cos \phi \\ \sin \theta \sin \phi \end{bmatrix}. \quad (1.49)$$

Comparing the ratio P_y/P_z in the two frames shows that $\phi' = \phi$; the azimuthal angle is the same in both frames. Comparing P_y/P_x reveals that

$$\tan \theta' = \frac{\sin \theta}{\gamma(\cos \theta + \beta)}. \quad (1.50)$$

For large velocities $\beta \rightarrow c$ so $\gamma \gg 1$ and this results in the photon trajectories in the observer frame being confined to a narrow cone of width $\theta \sim 1/\gamma$ around the direction of motion of the source. For example, consider ‘equatorial rays’ in the source-frame for which $\theta = 90^\circ$. In the observer frame these have

$$\tan \theta' = \frac{1}{\gamma\beta} \quad \Rightarrow \quad \theta' \simeq \frac{1}{\gamma} \quad (1.51)$$

so the width of the beam is on the order of $1/\gamma$ for $\gamma \gg 1$. This result will be useful when we consider synchrotron radiation.

It is also interesting to consider the energy flux in the beam. The *Doppler formula* says that the energy of the photons are boosted by a factor

$$\frac{h\nu'}{h\nu} = \frac{1}{\gamma(1 - \beta \cos \theta')}. \quad (1.52)$$

Now $\cos \theta' \simeq (1 - \theta'^2/2 + \dots)$ and $\beta = (1 - \gamma^{-2})^{1/2} \simeq (1 - \gamma^{-2}/2 + \dots)$ so

$$\gamma(1 - \beta \cos \theta') \simeq \gamma(1 - (1 - \gamma^{-2}/2)(1 - \theta'^2/2)) \simeq \frac{\gamma}{2}(\gamma^{-2} + \theta'^2) \sim 1/\gamma, \quad (1.53)$$

where we have used $\theta' \sim 1/\gamma$ for $\gamma \gg 1$. The typical energy boost factor is therefore $h\nu'/h\nu \sim \gamma$. These photons are compressed by a factor $\sim \gamma^2$ in angular width so the energy per unit area is increased by a factor $\sim \gamma^3$. What about the *rate* at which this energy flows? Consider a finite wave train of N waves. This will be emitted in time $\Delta t = N/\nu$ in the rest frame, but will pass our observer in time $\Delta t' = N/\nu' \sim \Delta t/\gamma$, with the net result that the energy flux (ie the energy per unit area per unit time) is increased by a factor $\sim \gamma^4$.

We will consider the transformation of radiation intensity more rigorously below.

1.10 Relativistic Decays

As an example of the use of 4-momentum conservation, consider a massive particle of mass M which spontaneously decays into two lighter decay products of mass m_1 and m_2 with energies (as measured in the rest-frame of the initial particle) E_1 and E_2 (see figure 1.7). We shall set $c = 1$ for clarity in this section.

Conservation of energy and momentum gives

$$\begin{aligned} M &= E_1 + E_2 \\ 0 &= \mathbf{P}_1 + \mathbf{P}_2 \end{aligned} \quad (1.54)$$

where we are using units such that $c = 1$. The latter tells us that $|\mathbf{P}_1|^2 = |\mathbf{P}_2|^2$, but $|\mathbf{P}_1|^2 = E_1^2 - m_1^2$ and so 4-momentum conservation can also be written as

$$\begin{aligned} M &= E_1 + E_2 \\ E_1^2 - m_1^2 &= E_2^2 - m_2^2 \end{aligned} \quad (1.55)$$

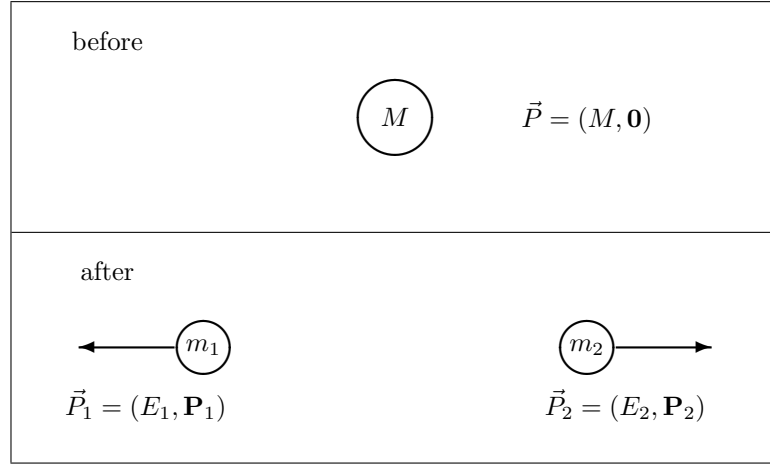


Figure 1.7: Decay of a heavy particle of mass M into two lighter decay products m_1, m_2 . Four momenta in the rest frame of the decaying particle are indicated.

and solving this pair of equations for the two unknowns E_1, E_2 yields

$$E_1 = \frac{M^2 + m_1^2 - m_2^2}{2M} \quad E_2 = \frac{M^2 - m_1^2 + m_2^2}{2M} \quad (1.56)$$

which are the energies of the decay products in the center-of-momentum frame.

Now consider the inverse process where two energetic particles collide and merge to form a heavier particle. The sum of the particle energies in the center of momentum frame then sets a threshold; this is the maximum mass particle that can be created. If we fire two equal mass particles at each other with energy $E_1 = E_2$ then the available energy is $M = E_1 + E_2$. If, on the other hand, we fire a particle of mass E_1 at a stationary target of mass m_2 , then the total energy of the resulting particle is $E = E_1 + m_2$ and the total momentum of the product is $\mathbf{P} = \mathbf{P}_1$, or equivalently $P^2 = P_1^2 = E_1^2 - m_1^2$ and so the mass of the product is

$$M^2 = E^2 - P^2 = (E_1 + m_2)^2 - E_1^2 + m_2^2 = m_1^2 + m_2^2 + 2m_2 E_1. \quad (1.57)$$

In the highly relativistic case where $E_1 \gg m_1, m_2$ the mass threshold is $M \simeq \sqrt{2m_2 E_1}$ which is much less than the mass threshold if one were to collide two particles of energy E_1 in a head-on collision. This is because in the stationary target case most of the energy is carried off in the momentum of the resulting particle, and the mass threshold is reduced by a factor $\sim 1/\sqrt{\gamma}$. This explains why the highest energy collisions are obtained in particle accelerators which collide counter-rotating beams of particles and anti-particles.

1.11 Invariant Volumes and Densities

Boosts of the observer induce changes in the spatio-temporal coordinates of events and thereby modify the 3-volume of a box, for instance. Boosts also cause changes in the energies and momenta of particles, and therefore modify the 3-volume of momentum space occupied by some set of particles, and therefore cause the momentum-space density of particles to vary etc.

There are however certain combinations of volumes, densities etc that remain invariant under Lorentz boosts, and it is highly desirable to write the laws of physics in a manner which makes use of these invariants as much as possible.

1.11.1 Space-Time Volume Element

One example of an invariant volume we have already seen is the space-time volume element:

$$dV dt = dx dy dz dt \quad \text{is Lorentz invariant.} \quad (1.58)$$

1.11.2 Momentum-Space Volume Element

Now consider momentum-space. Consider a set of particles which are nearly at rest, but actually have a small range of 3-momenta $0 \leq p_x \leq \Delta p_x$ etc. Consider the difference in 4-momenta of the particles at the origin and at the maximum of the range of momenta. To first order in Δp_i this is $\Delta \vec{P} = (0, \Delta p_x, \Delta p_y, \Delta p_z)$ since the range in energy is quadratic in the momenta for low momenta so we can ignore it. In a boosted frame, this 4-momentum difference is

$$\Delta \vec{P}' = \begin{bmatrix} \Delta P'_0 \\ \Delta P'_x \\ \Delta P'_y \\ \Delta P'_z \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \Delta P_x \\ \Delta P_y \\ \Delta P_z \end{bmatrix} = \begin{bmatrix} -\gamma\beta\Delta p_x \\ \gamma\Delta p_x \\ \Delta p_y \\ \Delta p_z \end{bmatrix} \quad (1.59)$$

and therefore

$$\Delta P'_x \Delta P'_y \Delta P'_z = \gamma \Delta P_x \Delta P_y \Delta P_z \quad (1.60)$$

but the energy in the rest-frame is $E = m$ while in the lab-frame $E' = \gamma m$ so $\gamma = E'/E$ and hence

$$\frac{d^3p}{E} \text{ is Lorentz invariant.} \quad (1.61)$$

This says that a boost will change both the energy of a bunch of particles and also the 3-momentum volume that they occupy, but the combination above is invariant under boosts. This is very useful in formulating the relativistic Boltzmann equation.

1.11.3 Momentum-Space Density

We define the *momentum space density* or *momentum distribution function* $n(\mathbf{p})$ so that $n(\mathbf{p})d^3p$ is the number of particles in 3-momentum volume element d^3p . Like any number, this is Lorentz invariant, so

$$n(\mathbf{p})d^3p = En(\mathbf{p})\frac{d^3p}{E} \text{ is Lorentz invariant} \quad (1.62)$$

and therefore

$$En(\mathbf{p}) \text{ is Lorentz invariant.} \quad (1.63)$$

1.11.4 Spatial Volume and Density

Consider a set of particles all moving at the same velocity relative to some inertial reference frame, as illustrated in figure 1.8. This figure shows that if a certain number of particles occupy a certain region in the lab-frame then they occupy a region in the rest-frame which is *larger* by a factor γ . This means that spatial volumes in the rest frame d^3r_0 and in the lab-frame are related by $d^3r_0 = \gamma d^3r$, but $E_0 = m$ and $E = \gamma m$, so $\gamma = E/E_0$ and therefore

$$Ed^3r \text{ is Lorentz invariant.} \quad (1.64)$$

Since the number of particles in some volume element $n(\mathbf{r})d^3r$ is clearly Lorentz invariant, this means that the spatial density transforms such that

$$\frac{n(\mathbf{r})}{E} \text{ is Lorentz invariant.} \quad (1.65)$$

This has an interesting consequence. Consider a neutral plasma consisting of streams of electrons and positrons propagating at equal velocities but in opposite directions. The two streams have equal densities by symmetry. Now consider the situation as perceived by an observer moving in the same direction as the electrons. That observer sees the positrons to have a higher energy, and therefore a higher space density, than the electrons. For that observer the plasma is not neutral but has a positive charge density.

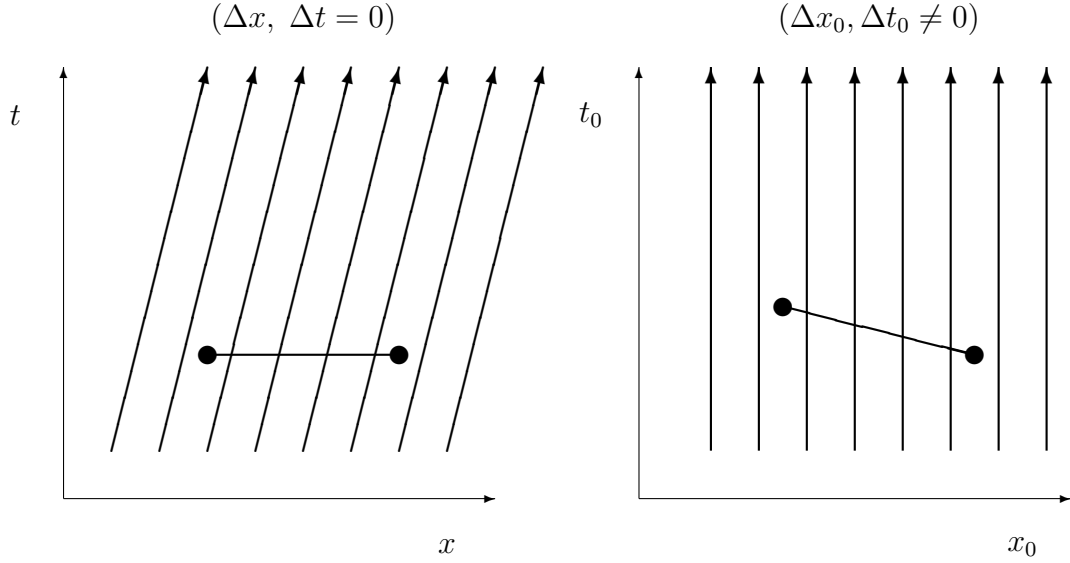


Figure 1.8: Illustration of the transformation of the spatial volume occupied by a set of particles. Left panel shows world-lines for a set of particles all with the same non-zero momentum in the ‘laboratory frame’. The horizontal line spans the range of x -coordinate Δx which contains a certain number of particles (four here) and this interval has $\Delta t = 0$. Right panel shows the same thing from the rest-frame of the particles. The particle world-lines are now vertical, and the transformed interval Δx is now tilted and has $\Delta t_0 \neq 0$. Since $\Delta x^2 - \Delta t^2$ is invariant, this means that $\Delta x_0 > \Delta x$.

1.11.5 Phase-Space Density

We define the *phase-space density* or *phase-space distribution function* $f(\mathbf{r}, \mathbf{p})$ such that $f(\mathbf{r}, \mathbf{p})d^3r d^3p$ is the number of particles in 6-volume $d^3r d^3p$. Since $E d^3r$ and d^3p/E are Lorentz invariant, then so is $d^3r d^3p$ and since $f(\mathbf{r}, \mathbf{p})d^3r d^3p$ is a number of particles this means that

$$f(\mathbf{r}, \mathbf{p}) \text{ is Lorentz invariant.} \quad (1.66)$$

1.11.6 Specific Intensity

We can use the foregoing to compute how the energy density of radiation and the specific intensity transform under a boost of one’s frame of reference.

The spatial energy density for particles occupying a momentum-space volume d^3p is given by the product of the spatial number density and the energy

$$d^3u = f(\mathbf{r}, \mathbf{p})E d^3p = f(\mathbf{r}, \mathbf{p})E p^2 dp d\Omega. \quad (1.67)$$

For photons, or any zero rest mass particle, $p = h\nu/c$ and $E = h\nu$, so $dp = (h/c)d\nu$ so

$$d^3u = u_\nu(\Omega) d\nu d\Omega = \frac{h^4}{c^3} f(\mathbf{r}, \mathbf{p}) \nu^3 d\nu d\Omega \quad (1.68)$$

and, since $f(\mathbf{r}, \mathbf{p})$ is Lorentz invariant, dividing through by $d\nu d\Omega$ shows that the specific energy density transforms such that

$$\frac{u_\nu(\Omega)}{\nu^3} \text{ is Lorentz invariant} \quad (1.69)$$

and therefore that the specific intensity also transforms in such a way that

$$\frac{I_\nu(\Omega)}{\nu^3} \text{ is Lorentz invariant.} \quad (1.70)$$

This is an extremely useful result. If one can compute the change in energy of a photon $\nu \rightarrow \nu'$ induced by a boost of the observer frame then we obtain the transformation of the intensity

$$I'_{\nu'}(\Omega') = \left(\frac{\nu'}{\nu}\right)^3 I_\nu(\Omega). \quad (1.71)$$

If we integrate over frequencies we find that the bolometric intensity transforms as

$$I'(\Omega') = \left(\frac{\nu'_*}{\nu_*}\right)^4 I(\Omega) \quad (1.72)$$

where ν_* is some characteristic frequency (e.g. the median frequency, the energy weighted mean frequency, the frequency of peak intensity or any other fiducial point on the spectrum). For a source emitting black-body radiation, this means that any observer also sees black-body radiation, though with temperature scaled according to the Doppler frequency shift.

1.12 Emission from Relativistic Particles

A useful procedure for computing the total power radiated by an accelerated relativistic particle is to go into the instantaneous rest frame of the particle (primed frame) and compute the power using Larmor's formula and then transform back to the (unprimed) observer frame.

The latter transformation is trivial, since provided the radiation emission is front-back symmetric (as is the case for dipole radiation and for most emission mechanisms covered here) the net momentum of the emitted radiation vanishes, so if an amount of energy dW' is radiated then the net 4-momentum of the radiation is $\vec{P}' = (dW'/c, 0, 0, 0)$ and therefore the energy in the observer frame is just $dW = \gamma dW'$. Similarly, if this energy is emitted in a time dt' in the rest frame, this corresponds to an interval $dt = \gamma dt'$ in the observer frame and therefore

$$P = \frac{dW}{dt} = P' = \frac{dW'}{dt'} \quad (1.73)$$

so the radiated power is Lorentz invariant.

Larmor's formula gives

$$P' = \frac{2q^2}{3c^3} |\mathbf{a}'|^2 \quad (1.74)$$

but, as discussed, the time component of the 4-acceleration vanishes in the rest-frame, and so $|\mathbf{a}'|^2 = \vec{A} \cdot \vec{A}$ and

$$P = \frac{2q^2}{3c^3} \vec{A} \cdot \vec{A} \quad (1.75)$$

which is manifestly covariant.

Using (1.39) this can also be written

$$P = \frac{2q^2}{3c^3} \gamma^4 (a_\perp^2 + \gamma^2 a_\parallel^2) \quad (1.76)$$

which gives the power radiated in terms of the 3-acceleration measured in the observer frame.

1.13 Problems

1.13.1 Speed and Velocity Transformation

Consider a particle moving with coordinate velocity \mathbf{u} in your frame (unprimed).

1. What is the gamma-factor $\gamma_{\mathbf{u}}$ that you assign to the particle?
2. Write down the 4-velocity \vec{U} for this particle in your frame of reference.
3. Apply a boost to obtain the 4-velocity \vec{U}' for an observer (your friend) moving at velocity $\mathbf{v} = (v, 0, 0)$ relative to you.
4. Obtain an expression for the gamma-factor $\gamma_{\mathbf{u}'}$ for the particle in your friend's frame of reference in terms of $\gamma_{\mathbf{u}}, v, c, \cos \theta = \hat{\mathbf{v}} \cdot \hat{\mathbf{u}}$ and c .
5. Obtain an expression for u'^1 , the x -component of the particle's coordinate velocity in your friend's frame in terms of v, u^1 and c .
6. Apply this to the case of a particle with $\mathbf{u} = (c, 0, 0)$. What is \mathbf{u}' ?

1.13.2 Four-Acceleration

1. Write down the 4-velocity \vec{U} for a particle with coordinate 3-velocity \mathbf{u} in terms of \mathbf{u}, c and $\gamma = 1/\sqrt{1 - u^2/c^2}$.
2. The 4-velocity is $\vec{A} = d\vec{U}/d\tau$ where τ is proper time along the particles world line. Rewrite this in terms of the rate of change of the 4-velocity with respect to *coordinate* time t .
3. Obtain an expression for the 4-acceleration in terms of $\gamma, \beta = u/c, \mathbf{a}_{\perp}$ and \mathbf{a}_{\parallel} (these being the components of the coordinate acceleration $\mathbf{a} = d\mathbf{u}/dt$ in the directions perpendicular and parallel to \mathbf{u}). The factor β should appear only once.
4. Use the above to obtain an expression for the norm $\vec{A} \cdot \vec{A}$ in terms of γ, a_{\perp} and a_{\parallel} .
5. What is the time component of the 4-acceleration in the rest-frame of the particle?
6. How is the squared coordinate acceleration $|a^2|$ in the rest-frame related to the invariant $\vec{A} \cdot \vec{A}$.
7. Use the above to obtain the rest-frame $|a|^2$ for a particle which in our frame has coordinate acceleration $\mathbf{a} = \mathbf{a}_{\perp}$.

1.13.3 Geometry of Minkowski space

Consider an explosion which occurs at the origin of Minkowski space coordinates $t = \mathbf{r} = 0$ which results in a cloud of test particles flying radially outward at all velocities $v < c$.

- a. Show that surfaces of constant proper time τ as measured by the test particles is the hyperboloid

$$t^2 = \tau^2 + r^2 \quad (1.77)$$

where $r^2 = x^2 + y^2 + z^2$. Sketch the intersection of this surface with the plane $y = z = 0$ and also show some representative test particle trajectories.

- b. Construct the spatial metric (line element) on this curved hypersurface as follows:

1. Set up polar coords r, θ, φ such that $(x, y, z) = r \sin \theta \cos \varphi, r \sin \theta \sin \varphi, r \cos \theta$
2. For a tangential line element ($r = \text{constant}$) $dt = 0$. Hence show (or argue) that the proper length of a tangential line element is

$$(dl_t)^2 = r^2((d\theta)^2 + \sin^2 \theta (d\varphi)^2) \equiv r^2 d\sigma^2 \quad (1.78)$$

3. For a radially directed line segment connecting points with Minkowski radial coordinate r , $r + dr$ (but at the same τ) there is a non-zero dt given by $d(t^2) = d(\tau^2 + r^2) = d(r^2)$. Hence show that the proper length is

$$(dl_r)^2 = (dr)^2 - (dt)^2 = \frac{(dr)^2}{1 + r^2/\tau^2} \quad (1.79)$$

4. Combine 2,3 to give

$$(dl)^2 = \frac{(dr)^2}{1 + r^2/\tau^2} + r^2 d\sigma^2 \quad (1.80)$$

- c. Now consider a rescaling of the radial coordinate $R = r/\tau$.

1. Show that $R = \gamma v$, where $\gamma \equiv (1 - v^2)^{-1/2}$ and is therefore constant label for each particle; a ‘comoving radial coordinate’.
2. Rewrite the result of b.4. as

$$dl^2 = \tau^2 \left(\frac{dR^2}{1 + R^2} + R^2 d\sigma^2 \right) \quad (1.81)$$

- d. Show that (flat) Minkowski space, when written in τ, R, θ, φ coordinates takes the form

$$ds^2 = -d\tau^2 + \tau^2 \left(\frac{dR^2}{1 + R^2} + R^2 d\sigma^2 \right) \quad (1.82)$$

Rewrite this in terms of an alternative comoving radial coordinate ω where $R = \sinh \omega$. Compare your results with formulae for the space-time geometry of an open Friedmann-Robertson-Walker cosmology from any standard introductory cosmology text.

1.13.4 Relativistic decays

Consider a heavy particle H of mass M which decays into two light particles of equal mass m .

- a. Show that in the rest frame of the heavy particle the energy of a decay product is $E = M/2$ and the modulus of the momentum is $|\mathbf{p}| = \beta_p M/2$ where $\beta_p = (1 - 4m^2/M^2)^{1/2}$, so the 4-momentum is

$$\vec{p} = \frac{M}{2} \begin{bmatrix} 1 \\ \beta_p \mu \\ \beta_p \sqrt{1 - \mu^2} \cos \varphi \\ \beta_p \sqrt{1 - \mu^2} \sin \varphi \end{bmatrix} \quad (1.83)$$

- b. Compute the decay product 4-momentum in a frame (the ‘laboratory frame’) in which the decaying particle has velocity β parallel to the x -axis.
- c. Show that for decays which are isotropic in the rest frame of the decaying particle the decay product energy is uniformly distributed in the range $E_p^- < E_p < E_p^+$ where

$$E_p^\pm = \frac{E_H}{2} \pm \frac{p_H}{2} (1 - 4m^2/M^2)^{1/2} \quad (1.84)$$

Sketch the minimum and maximum decay product energy as a function of the energy of the decaying particle E_H .

- d. Now consider decays from a distribution of heavy particles which all have the same energy in the lab frame but have isotropically distributed momenta. What is the distribution in energy of the decay products? What is the form of the phase-space distribution function $f(\vec{p})$ for the decay products? (Assume that the occupation numbers for the final states are negligible).

Chapter 2

Dynamics

This chapter consists of a review of some useful results from Lagrangian and Hamiltonian dynamics. We first introduce the concepts of generalized coordinates, the Lagrangian and the action. We state the *principle of least action* and then give some examples which show how to construct the Lagrangian to obtain the equations of motion. We show how energy and momentum conservation arise from symmetries of the Lagrangian under shift of time and spatial translations, and we also show how the Lagrangian formalism is useful for generating the equations of motion in transformed coordinates. We then review Hamilton's equations and finally we discuss adiabatic invariance.

2.1 Lagrangian Dynamics

2.1.1 Generalized Coordinates

In Lagrangian dynamics a mechanical system is described by *generalized coordinates* which we denote by q_i with i running from 1 through N with N the number of *degrees of freedom*. We will also use the notation $\mathbf{q} = q_i$.

The values of the coordinates at some instant of time are generally not enough to specify the state of the system, to fully specify the state one needs to give also the values of the *velocities* \dot{q}_i . The future evolution is then determined.

2.1.2 The Lagrangian and the Action

A mechanical system is defined by its *Lagrangian*. This is a scalar function of the coordinates and velocities and optionally time and is denoted by $L(q_i, \dot{q}_i, t)$. It has units of energy.

The *action* S is defined for a bounded path $\mathbf{q}(t)$ in coordinate space and is the time integral of the Lagrangian

$$S \equiv \int_{t_1}^{t_2} dt L(\mathbf{q}(t), \dot{\mathbf{q}}(t), t). \quad (2.1)$$

The action has units of angular momentum.

2.1.3 The Principle of Least Action

The *principle of least action* asserts that the actual evolutionary histories (or more briefly 'paths') $\mathbf{q}(t)$ that a system follows are those which minimize (or more generally extremize) the action:

$$\delta S = \delta \int_{t_1}^{t_2} dt L(\mathbf{q}(t), \dot{\mathbf{q}}(t), t) = 0. \quad (2.2)$$

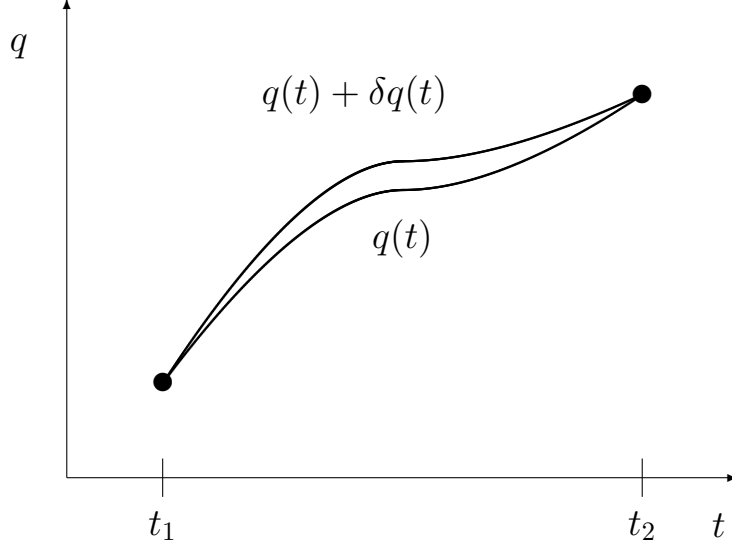


Figure 2.1: The lines show two hypothetical evolutionary histories of a one-dimensional system. The two paths begin and end at the same points.

2.1.4 The Euler-Lagrange Equations

The ‘global’ minimization (2.2) implies a certain condition on the derivatives of the Lagrangian which must be locally satisfied. This condition provides the equations of motion for the system.

Consider a system with one degree of freedom: $L(q, \dot{q}, t)$ and consider two hypothetical neighboring paths $q(t)$ and $q'(t) = q(t) + \delta q(t)$ as illustrated in figure 2.1. The variation of $q(t)$ implies a corresponding variation of the velocity $\delta \dot{q} = d(\delta q)/dt \equiv \dot{\delta q}$.

The variation of the action is

$$\delta S = S' - S = \int dt L(q + \delta q, \dot{q} + \dot{\delta q}, t) - \int dt L(q, \dot{q}, t). \quad (2.3)$$

If we make a Taylor expansion and ignore terms higher than linear in the (assumed infinitesimal) perturbation to the path, we have

$$\delta S = \int_{t_1}^{t_2} dt \left[\delta q \frac{\partial L}{\partial q} - \dot{\delta q} \frac{\partial L}{\partial \dot{q}} \right]. \quad (2.4)$$

Now the second term in brackets here can be written as

$$\dot{\delta q} \frac{\partial L}{\partial \dot{q}} = \frac{d}{dt} \left(\delta q \frac{\partial L}{\partial \dot{q}} \right) - \delta q \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) \quad (2.5)$$

and therefore the action variation can be written as

$$\delta S = \left[\delta q \frac{\partial L}{\partial \dot{q}} \right]_{t_1}^{t_2} + \int_{t_1}^{t_2} dt \delta q \left[\frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} \right] \quad (2.6)$$

but the first term vanishes since $\delta q(t_1) = \delta q(t_2) = 0$ and, since the variation $\delta q(t)$ is arbitrary, the term within the square brackets in the integral must vanish. This gives the *Euler-Lagrange equation*:

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = 0. \quad (2.7)$$

The generalization to multi-dimensional systems is straightforward and we obtain

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0. \quad (2.8)$$

which gives a set of equations, one per degree of freedom i . As we will now see, these equations provide the means to evolve the state of the system; they provide the equations of motion for the system.

2.1.5 Example Lagrangians

So far we have not said how the Lagrangian for a system is determined; we have simply asserted that there is such a function whose minimization provides the equations of motion.

Consider the Lagrangian $L = m|\dot{\mathbf{x}}|^2/2$, where \mathbf{x} is the usual Cartesian spatial coordinate. Note that this is the only function which is at most quadratic in the velocity and satisfies homogeneity (which means L cannot depend on \mathbf{x}) and isotropy (L is independent of the direction of motion). The Euler-Lagrange equations are then

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{x}_i} = 0 \quad \Rightarrow \quad \dot{x}_i = \text{constant} \quad (2.9)$$

which is the *law of inertia*; we have obtained the equations of motion for a free particle, for which the Lagrangian is just the kinetic energy.

We also see an example of what is a general rule: if the Lagrangian is independent of one of the coordinates then the corresponding velocity is constant.

A less trivial example is a set of particles labelled by an index a and with (time independent) Lagrangian

$$L(\mathbf{x}_a, \dot{\mathbf{x}}_a) = \sum_a \frac{1}{2} m_a |\dot{\mathbf{x}}_a|^2 - U(\mathbf{x}_1, \mathbf{x}_2, \dots) \quad (2.10)$$

for which the Euler-Lagrange equations are

$$m \frac{d\dot{\mathbf{x}}_a}{dt} = - \frac{\partial U}{\partial \mathbf{x}_a} \quad (2.11)$$

which we identify with Newton's law for a system with potential energy $U(\mathbf{x}_a)$.

Note that the Lagrangian here is the kinetic energy minus the potential energy: $L = T - U$.

Again, if L is independent of one of the coordinates $\partial U / \partial \mathbf{x}_b = 0$, then the corresponding velocity $\dot{\mathbf{x}}_b$ is constant.

2.2 Conservation Laws

2.2.1 Energy Conservation

If there is no explicit time dependence of the Lagrangian, so $L = L(\mathbf{q}, \dot{\mathbf{q}})$, the total derivative of the Lagrangian is

$$\frac{dL}{dt} = \sum \frac{\partial L}{\partial q_i} \dot{q}_i + \sum \frac{\partial L}{\partial \dot{q}_i} \ddot{q}_i \quad (2.12)$$

which, using the Euler-Lagrange equation, is

$$\frac{dL}{dt} = \sum \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \dot{q}_i + \sum \frac{\partial L}{\partial \dot{q}_i} \ddot{q}_i = \frac{d}{dt} \sum \frac{\partial L}{\partial \dot{q}_i} \dot{q}_i \quad (2.13)$$

so we have

$$\frac{d}{dt} \left(\sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - L \right) = 0. \quad (2.14)$$

which means that the quantity

$$E \equiv \sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - L \quad (2.15)$$

which we call the *energy* is a constant of the motion.

The energy is conserved for any system with $\partial L / \partial t = 0$. Such systems are said to be ‘closed’ or ‘isolated’.

The partial derivative of the Lagrangian $\partial L / \partial \dot{q}_i$ with respect to the coordinate velocity \dot{q}_i is called the *momentum* or, more verbosely, the momentum canonically conjugate to the coordinate q_i . In terms of the momentum, the energy is

$$E \equiv \sum_i \dot{q}_i p_i - L. \quad (2.16)$$

Note that for the system (2.10) the momenta are $\mathbf{p}_a = m_a \dot{\mathbf{x}}_a$ and the energy is

$$E = \sum_a m_a \dot{x}_a^2 / 2 + U(\mathbf{x}_1, \mathbf{x}_2, \dots) \quad (2.17)$$

which is the sum of the kinetic and potential energies as expected.

2.2.2 Momentum Conservation

Consider a system in which the potential energy depends only on the relative values of the Cartesian coordinates: ie the Lagrangian is invariant if we translate the entire system by a distance $\Delta \mathbf{x}$. The change in the Lagrangian is

$$\delta L = \Delta \mathbf{x} \cdot \sum_a \frac{\partial L}{\partial \mathbf{x}_a} = 0 \quad \Rightarrow \quad \sum_a \frac{\partial L}{\partial \mathbf{x}_a} = 0 \quad (2.18)$$

but $\partial L / \partial \mathbf{x}_a = -d(\partial L / \partial \dot{\mathbf{x}}_a) / dt$ from the E-L equations, so

$$\frac{d}{dt} \sum_a \frac{\partial L}{\partial \dot{\mathbf{x}}_a} = 0 \quad (2.19)$$

or equivalently that the quantity

$$\mathbf{p} = \sum_a \mathbf{p}_a = \sum_a \partial L / \partial \dot{\mathbf{x}}_a \quad (2.20)$$

is conserved. This is called the *total momentum* of the system and its conservation law follows directly from spatial homogeneity.

Similar arguments can be used to show that the *angular momentum* is conserved if the Lagrangian is independent of orientation.

These quantities are the *total momenta* of the system. In addition to these conservation laws we also have conservation of individual momenta

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i} = \text{constant} \quad (2.21)$$

if the Lagrangian is independent of the corresponding generalized coordinate: $\partial L / \partial q_i = 0$.

2.3 Coordinate Transformations

The Euler-Lagrange equations are useful for generating the equations of motion in arbitrary coordinate systems. Since L is a scalar quantity it is independent of the representation of the coordinate

system. For example, consider a free particle, for which the Lagrangian is $L = m|\dot{\mathbf{x}}|^2/2$. However, consider what happens if one works in an expanding coordinate system and defines

$$\mathbf{x} = a(t)\mathbf{r} \quad (2.22)$$

with $a(t)$ a scale-factor.

Expressing the Lagrangian in terms of the new \mathbf{r} coordinates we have $\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}}$ which means that the Lagrangian $L = m\dot{x}^2/2$ becomes a function of both \mathbf{r} and $\dot{\mathbf{r}}$:

$$L(\mathbf{r}, \dot{\mathbf{r}}) = \frac{m}{2}(\dot{a}^2 r^2 + 2a\dot{a}\mathbf{r} \cdot \dot{\mathbf{r}} + a^2 \dot{\mathbf{r}}^2) \quad (2.23)$$

To form the Euler-Lagrange equations we need the partial derivatives

$$\begin{aligned} \partial L / \partial \dot{\mathbf{r}} &= m(\dot{a}\mathbf{r} + a\dot{\mathbf{r}}) \\ \partial L / \partial \mathbf{r} &= m(\dot{a}^2 \mathbf{r} + a\dot{a}\dot{\mathbf{r}}) \end{aligned} \quad (2.24)$$

and the Euler-Lagrange equations are then

$$\ddot{\mathbf{r}} + 2\frac{\dot{a}}{a}\dot{\mathbf{r}} + \frac{\ddot{a}}{a}\mathbf{r} = 0. \quad (2.25)$$

2.4 Hamilton's Equations

In Hamilton's formulation of dynamics we define the *Hamiltonian*

$$H \equiv \sum \dot{q}p - L(q, \dot{q}, t). \quad (2.26)$$

This is identical in value to the energy defined in (2.15).

At first sight, the Hamiltonian would seem to be a function of q, \dot{q}, p, t . However, if we write down the differential

$$dH = \dot{q}dp + p d\dot{q} - \frac{\partial L}{\partial \dot{q}} d\dot{q} - \frac{\partial L}{\partial q} dq - \frac{\partial L}{\partial t} dt \quad (2.27)$$

we see that the 2nd and 3rd terms cancel each other (recalling the definition $p \equiv \partial L / \partial \dot{q}$) and so dH contains only terms with dp , dq and dt so evidently the Hamiltonian is only a function of q , p and t :

$$H = H(q, p, t). \quad (2.28)$$

Using the Euler-Lagrange equation to replace $\partial L / \partial q$ in the fourth term in (2.27) with \dot{p} and using the definition of the Hamiltonian (2.26) to replace $\partial L / \partial t$ in the fifth term with $-\partial H / \partial t$ we can write dH as

$$dH = \dot{q}dp - \dot{p}dq + \frac{\partial H}{\partial t} dt \quad (2.29)$$

but we can also write this as

$$dH = \frac{\partial H}{\partial p} dp + \frac{\partial H}{\partial q} dq + \frac{\partial H}{\partial t} dt \quad (2.30)$$

and comparing the coefficients of dp and dq yields *Hamilton's equations*:

$$\begin{aligned} \dot{q} &= \partial H / \partial p \\ \dot{p} &= -\partial H / \partial q \end{aligned} \quad (2.31)$$

The generalization to a multi-dimensional system is straightforward and we then have

$$\begin{aligned} \dot{q}_i &= \partial H / \partial p_i \\ \dot{p}_i &= -\partial H / \partial q_i \end{aligned} \quad (2.32)$$

For a system with N degrees of freedom, the Euler-Lagrange equations provide a set of N second order differential equations. Hamilton's equations, in contrast, are a set of $2N$ coupled first order differential equations. Either set of equations can be integrated to obtain the evolution of the system.

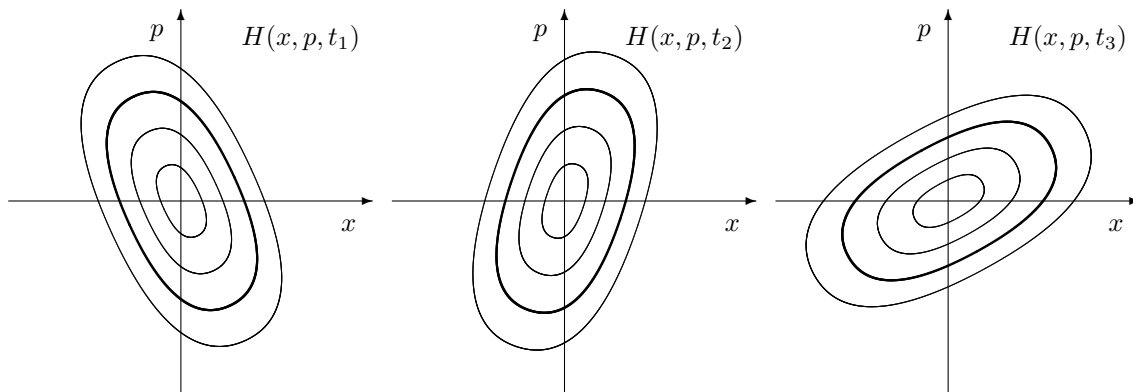


Figure 2.2: Adiabatic invariance concerns systems where the Hamiltonian evolves slowly with time. The curves show contours of the Hamiltonian for such a system at three different times. At each time the system will orbit around a certain contour. Provided the system evolves slowly, the final contour is fully determined by the initial contour, and is independent of the details of how the system changed. The *adiabatic invariant* turns out to be the area within the contour.

2.5 Adiabatic Invariance

Hamilton's equations are convenient for discussing *adiabatic invariance*.

Consider a system with Hamiltonian $H(p, q, t)$. For constant t , the energy is conserved, and the system will move around a contour of the Hamiltonian; ie around some closed loop in q, p space (see figure 2.2). Now ask what happens if there is some slow explicit time dependence of the Hamiltonian (this might be some slow variation of a spring constant or perhaps some external contribution to the potential energy). If the change in the Hamiltonian is sufficiently slow the system will appear to conserve energy over short time-scales (successive orbits will be very similar) but over long time-scales there will be secular evolution and the system will evolve through a series of quasi-periodic orbits and the energy will also change with time.

At time t_1 the Hamiltonian is $H_1(q, p) = H(q, p, t_1)$ and the possible orbits form a 1-parameter family; these are just the contours of H_1 . Similarly at time t_2 we have $H_2(q, p)$ and we have another set of orbits. For a given initial energy E_1 , and for sufficiently slow variation of the potential, the system will end up in a specific energy orbit E_2 , independent of the details of how the variation took place. Equivalently, we can say that something is conserved in the evolution. It turns out that that quantity — which we call an *adiabatic invariant* — is just the area within the orbit in position-momentum space:

$$I = \int \int dp dq = \oint dq p \quad (2.33)$$

is conserved.

The derivation of this powerful and simple result is rather tedious (see L+L Mechanics for the details). The flavor of this result, can however be appreciated with a simple example. Consider a simple system which is a conker on a string of length l and rotating at some initial velocity v . What happens if we slowly shorten the string? Shortening the string does not exert any torque on the conker, so the angular momentum must be conserved and therefore $lv = \text{constant}$ and so $v \propto 1/l$; as the string shortens the velocity of the particle increases. The energy of the conker is *not* conserved; since the string is in tension, the agent shortening the string must do work. The energy of the system is entirely kinetic and increases as $E \propto v^2 \propto 1/l^2$. The angular frequency of the system is $\omega = v/l$ which is also proportional to v^2 and so we find that the energy evolves in proportion to the frequency:

$$E \propto \omega(t). \quad (2.34)$$

Let us derive the equivalent scaling law for a system consisting of a simple harmonic oscillator

with a time varying spring constant using (2.33). The Hamiltonian is

$$H = \frac{1}{2}p^2/m + \frac{1}{2}m\omega^2(t)q^2 \quad (2.35)$$

For $\omega = \text{constant}$ the solution is an elliptical orbit in q, p space with semi-major axes $p_0 = \sqrt{2mE}$ and $q_0 = \sqrt{2E/m\omega^2}$ so the area of the ellipse is

$$I \propto p_0 q_0 \propto E/\omega \quad (2.36)$$

so, if I is conserved we find $E(t) \propto \omega(t)$ as for the conker, provided the frequency changes slowly (the requirement is that the fractional change in the frequency over one orbit should be small).

Another way to get to this result, again for a simple harmonic oscillator with time varying frequency, is to integrate the equations of motion:

$$\ddot{x} + \omega^2(t)x = 0. \quad (2.37)$$

For $\omega = \text{constant}$ the solution is $x = ae^{i\omega t}$, so let's look for a solution with slowly varying amplitude $x = a(t)e^{i\omega t}$. Performing the differentiation with respect to time yields

$$\ddot{x} + \omega^2(t)x = 0 = i(a\dot{\omega} + 2\dot{a}\omega)e^{i\omega t} + \ddot{a}e^{i\omega t} \quad (2.38)$$

but if $a(t)$ is slowly varying we can neglect \ddot{a} as compared to the other terms, and we find that the amplitude must obey the equation

$$a\dot{\omega} + 2\dot{a}\omega = 0 \quad (2.39)$$

which has solution $a(t) \propto \omega(t)^{-1/2}$. The kinetic energy of the system is $E \propto \dot{x}^2 \propto (\omega a)^2 \propto \omega$ so again we obtain the scaling law $E(t) \propto \omega(t)$.

Adiabatic invariance is closely connected to Liouville's theorem. The latter says that particles in 6-dimensional phase space behave like an incompressible fluid. If we populate the orbits inside some given energy contour of $H_1(p, q)$, it makes sense then that the area that these particles end up occupying at the end should be unchanged.

Lastly, let us think about the quantum mechanics of this system. The energy levels for a constant frequency oscillator are

$$E = (n + 1/2)\hbar\omega \quad (2.40)$$

so the classical adiabatic invariant behavior corresponds to conservation of n , which seems reasonable.

For the simple case of SHM with time varying frequency the above approaches yield the same answer, for more complicated systems the most general and powerful approach is to use (2.33).

2.6 Problems

2.6.1 Extremal paths

Extremal paths. Consider photon propagating through a medium with inhomogeneous refractive index $n(\mathbf{x})$. Show that the time of flight is

$$t = \frac{1}{c} \int dl \left(\frac{dx_i}{dl} \frac{dx_i}{dl} \right)^{1/2} n(x_i) \quad (2.41)$$

where $\mathbf{x}(l)$ is the path of the photon and where l is an arbitrary parameterisation along the path. According to Fermat's principle the variation of the time of flight δt vanishes for the actual ray. Show that the Euler-Lagrange equations for this variation problem are 'Snell's law of refraction'

$$\frac{dn\hat{\mathbf{k}}}{dl} = \nabla n \quad (2.42)$$

where $\hat{\mathbf{k}}$ is the photon direction and where the parameterisation has been chosen so that $|d\mathbf{x}/dl| = 1$.

Use Snell's law to estimate the angular deflection of a ray passing through a region of size L with some refractive index fluctuation δn .

If one observes a source at distance D through an inhomogeneous medium with random refractive index fluctuations δn with a coherence scale L , how large does δn need to be to cause multipath propagation?

2.6.2 Schwarzschild Trajectories

In general relativity the metric tensor $g_{\mu\nu}$ is defined such that the proper time interval corresponding to coordinate separation \vec{dx} is $(d\tau)^2 = -g_{\mu\nu}(\vec{x})dx^\mu dx^\nu$. Massive particles move along world lines that minimize the proper time:

$$\delta \int d\lambda \sqrt{-g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda}} = 0, \quad (2.43)$$

where λ is some arbitrary parameterisation of the path.

a) Use this variational principle to show that if $g_{\mu\nu}(\vec{x})$ is independent of one of the coordinates x^α this implies that

$$U_\alpha = g_{\alpha\beta} U^\beta = \text{constant} \quad (2.44)$$

where $\vec{U} = d\vec{x}/d\tau$ is the 4-velocity.

b) The Schwartzschild metric for a mass m is (in units such that $c = G = 1$)

$$-(d\tau)^2 = -\left(1 - \frac{2m}{r}\right) (dt)^2 + \left(1 - \frac{2m}{r}\right)^{-1} (dr)^2 + r^2((d\theta)^2 + \sin^2 \theta (d\phi)^2). \quad (2.45)$$

Use the result from part a) to obtain the 'energy equation' for particles on radial orbits (θ, ϕ constant) as

$$(dr/d\tau)^2 = \dots \quad (2.46)$$

Note: You don't need to know any GR to answer this question!

2.6.3 Lagrangian electrodynamics

From the stationarity of the action

$$S \equiv \int_{t_1}^{t_2} dt L(q_i, \dot{q}_i, t) \quad (2.47)$$

for a system with generalised coordinates q_i , derive the Euler-Lagrange equations

$$\frac{d(\partial L / \partial \dot{q}_i)}{dt} = \partial L / \partial q_i \quad (2.48)$$

The Lagrangian for a particle of mass m and charge q moving in an electromagnetic field is given by

$$L(\mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2} m \dot{\mathbf{x}}^2 + \frac{q}{c} \mathbf{A} \cdot \dot{\mathbf{x}} - q\phi \quad (2.49)$$

where $\mathbf{A}(\mathbf{x}, t)$ and $\phi(\mathbf{x}, t)$ are the vector and scalar potentials.

Show that the momentum conjugate to \mathbf{x} is

$$\mathbf{p} = m\dot{\mathbf{x}} + \frac{q}{c} \mathbf{A} \quad (2.50)$$

and that Euler-Lagrange equation generates the Lorentz force law

$$m\ddot{\mathbf{x}} = q[\mathbf{E} + \frac{1}{c} \dot{\mathbf{x}} \times \mathbf{B}] \quad (2.51)$$

where $\mathbf{B} = \nabla \times \mathbf{A}$ and $\mathbf{E} = -\nabla\phi - (1/c)\partial\mathbf{A}/\partial t$. (You may make use of the identity $\mathbf{v} \times (\nabla \times \mathbf{A}) = \nabla(\mathbf{v} \cdot \mathbf{A}) - (\mathbf{v} \cdot \nabla)\mathbf{A}$).

Chapter 3

Random Fields

Random fields and random, or stochastic, processes are ubiquitous in astronomy. The radiation field entering our instruments, the distribution of stars and galaxies, the distribution of wave-like disturbances in the early universe, in spiral galaxies or in the sun, the distribution of electrons in a CCD image; all of these are random processes. Here we will introduce statistics which are useful for describing such processes and some useful mathematical tools.

3.1 Descriptive Statistics

Let's consider, for concreteness, a random scalar function of position $f(\mathbf{r})$, though we could equally well have chosen a vector function like the electric field, and it might be a function of space and time, or just of time, or of 6-dimensional phase-space etc. Let us also consider *statistically homogeneous* processes, which fluctuate, but in which the *statistical character* of the fluctuations does not vary with location.

The most general description would be some kind of *probability density functional* $P(f(\mathbf{r}))D[f(\mathbf{r})]$ giving the probability to observe a particular configuration of the field $f(\mathbf{r})$. This probability can be thought of as giving the distribution over an ensemble of realizations, or alternatively one might think of it as giving the distribution over samples drawn from a single infinite realization at randomly chosen locations.

3.1.1 N -Point Distribution Functions

A useful description is provided by the hierarchy of *N -point distribution functions*. The 1-point distribution function is

$$p(f)df \tag{3.1}$$

and gives the probability to observe a field value f at some randomly chosen point in space. The 2-point distribution function is

$$p(f_1, f_2)df_1df_2 \tag{3.2}$$

and gives the probability to observe $f(\mathbf{r}_1) = f_1$ and $f(\mathbf{r}_2) = f_2$ at two positions $\mathbf{r}_1, \mathbf{r}_2$. For statistically homogeneous processes this will depend only on the separation $\mathbf{r}_1 - \mathbf{r}_2$, and for a statistically isotropic process it only depends on the modulus of the separation $|\mathbf{r}_1 - \mathbf{r}_2|$.

One can readily generalize this to arbitrary numbers of points, and the whole hierarchy constitutes a full description of the random process. The utility of this approach is that useful physics can often be extracted from a reduced approximate description in terms of a few low order distribution functions.

3.1.2 N -Point Correlation Functions

We obtain the N -point correlation functions by integrating over the distribution functions. For example, the two-point correlation function is

$$\xi(\mathbf{r}_{12}) = \langle f_1 f_2 \rangle = \int df_1 \int df_2 f_1 f_2 p(f_1, f_2) \quad (3.3)$$

which again depends only on the separation of the pair of measurement points.

This can be generalized to give the *three-point correlation function* and so on.

The N -point correlation functions are *moments* of the corresponding distribution functions.

3.2 Two-point Correlation Function

The *two point correlation function* (or *auto-correlation function*) is very useful as it allows one to compute the *variance* of any *linear* function of the random field. For example, the variance of the field itself is $\langle f^2 \rangle$ and is equal to the value of the auto-correlation function at *zero lag*: $\xi(0)$.

For a less trivial example, consider the average of the field over some averaging cell:

$$\bar{f} = \frac{1}{V} \int_V d^3r f(\mathbf{r}) \quad (3.4)$$

with V the volume of the cell. This is also a linear function of f . The variance, or mean square value, of \bar{f} is obtained by writing down two copies of the integral (3.4), with the second having integration variable \mathbf{r}' , and enclosing it within the $\langle \dots \rangle$ ensemble averaging operator:

$$\langle \bar{f}^2 \rangle = \left\langle \frac{1}{V} \int_V d^3r f(\mathbf{r}) \frac{1}{V} \int_V d^3r' f(\mathbf{r}') \right\rangle. \quad (3.5)$$

Rearranging the terms in this double integral and realising that the $\langle \dots \rangle$ operator only acts on the *stochastic* variables $f(\mathbf{r})$, $f(\mathbf{r}')$ gives

$$\langle \bar{f}^2 \rangle = \frac{1}{V^2} \int_V d^3r \int_V d^3r' \langle f(\mathbf{r}) f(\mathbf{r}') \rangle = \frac{1}{V^2} \int_V d^3r \int_V d^3r' \xi(\mathbf{r} - \mathbf{r}'). \quad (3.6)$$

So the variance of the average field may be computed as a double integral over the 2-point function.

It is sometimes the case that the range of correlations of the field is limited, so $\xi(r)$ is appreciable only within some *coherence length* r_c and is negligible for $r \gg r_c$. If the size of the averaging cell is large compared to the coherence length $r_c \ll V^{1/3}$, then

$$\int_V d^3r' \xi(\mathbf{r}' - \mathbf{r}) \simeq \int d^3r' \xi(\mathbf{r}') \quad (3.7)$$

where the range of integration is unrestricted, provided the point \mathbf{r} lies at least a distance r_c from the walls of the cell, and the variance is then

$$\langle \bar{f}^2 \rangle \simeq \frac{1}{V} \int d^3r \xi(r) \sim \xi(0) r_c^3 / V = \langle f^2 \rangle r_c^3 / V \quad (3.8)$$

which says that the rms value of the averaged field \bar{f} will be approximately the rms value of the field f divided by \sqrt{N} , with $N = V/r_c^3$ the number of *coherence volumes* within the cell.

3.3 Power Spectrum

The 2-point function $\xi(r)$ and its generalizations to N -points are *real-space* statistics. The *translational invariance* of statistically homogeneous fields suggests that *Fourier-space* or *spectral* statistics may also be useful.

The Fourier transform of the field f is

$$\tilde{f}(\mathbf{k}) = \int d^3r f(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}. \quad (3.9)$$

and the *power spectrum* is proportional to the expectation value of the squared modulus of $\tilde{f}(\mathbf{k})$

$$P(\mathbf{k}) \propto \langle |\tilde{f}(\mathbf{k})|^2 \rangle. \quad (3.10)$$

The tricky thing here is getting the constant of proportionality since we are dealing with a field of infinite extent. For a random field occupying some large but finite volume *Parseval's theorem* tells us that $(2\pi)^{-3} \int d^3k |\tilde{f}(\mathbf{k})|^2 = \int d^3r f^2(r)$ and the latter integral, and therefore also $|\tilde{f}(\mathbf{k})|^2$, increase in proportion to the volume, so to get a sensible measure of the power we need somehow to divide $\langle |\tilde{f}(\mathbf{k})|^2 \rangle$ by some suitably infinite volume factor.

To make the definition of the power precise, consider the expectation of the product of the transform at two different spatial frequencies \mathbf{k} , \mathbf{k}' , or more specifically, the average of the the product $\tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}')$ with \tilde{f}^* the complex conjugate of \tilde{f} . Writing out two copies of (3.9), wrapping them in the averaging operator $\langle \dots \rangle$, and re-arranging terms yields

$$\langle \tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}') \rangle = \int d^3r \int d^3r' \langle f(\mathbf{r})f(\mathbf{r}') \rangle e^{i\mathbf{k}\cdot\mathbf{r}} e^{-i\mathbf{k}'\cdot\mathbf{r}'}. \quad (3.11)$$

Now $\langle f(\mathbf{r})f(\mathbf{r}') \rangle = \xi(\mathbf{r} - \mathbf{r}')$ so on changing the second integration variable from \mathbf{r}' to $\mathbf{z} = \mathbf{r}' - \mathbf{r}$ the double integral separates into a product of integrals and we have

$$\langle \tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}') \rangle = \int d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} \int d^3z \xi(\mathbf{z}) e^{i\mathbf{z}\cdot\mathbf{k}} \quad (3.12)$$

or

$$\langle \tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}') \rangle = (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') P(\mathbf{k}) \quad (3.13)$$

where we have recognized the first integral in (3.12) as a representation of the Dirac δ -function and where we have now defined the *power spectrum* as

$$P(\mathbf{k}) \equiv \int d^3r \xi(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}. \quad (3.14)$$

Equation (3.13) tells us that different Fourier modes (ie $\mathbf{k}' \neq \mathbf{k}$) are completely uncorrelated. This is a direct consequence of the assumed translational invariance, or statistical homogeneity, of the field. On the other hand, for $\mathbf{k} = \mathbf{k}'$ the infinite volume factor in $\langle \tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}) \rangle = \langle |\tilde{f}(\mathbf{k})|^2 \rangle$ is supplied by the Dirac δ -function.

We have only computed $\langle \tilde{f}(\mathbf{k})\tilde{f}^*(\mathbf{k}') \rangle$ here. Other correlation coefficients such as $\langle \tilde{f}(\mathbf{k})\tilde{f}(\mathbf{k}') \rangle$ can be obtained using the fact that $\tilde{f}(\mathbf{k}) = \tilde{f}^*(-\mathbf{k})$ which follows from the assumed reality of $f(\mathbf{r})$.

It is interesting to contrast the character of the fields in real-space and in Fourier-space. In real space, $f(\mathbf{r})$ will generally have extended correlations, $\langle f(\mathbf{r})f(\mathbf{r}') \rangle \neq 0$ for $\mathbf{r} \neq \mathbf{r}'$, but is statistically homogeneous. In Fourier space $\tilde{f}(\mathbf{k})$ there are no extended correlations — the field $\tilde{f}(\mathbf{k})$ is completely *incoherent* — but the field is *inhomogeneous* since, for example, $\langle |\tilde{f}(\mathbf{k})|^2 \rangle$ varies with position \mathbf{k} .

Properties of the power spectrum:

- The power spectrum and auto-correlation function are Fourier transform pairs of one another. This is the *Wiener-Khinchin theorem*.
- The power spectrum tells us how the variance of the field is distributed over spatial frequency. Taking the inverse transform of (3.14) at $\mathbf{r} = 0$ gives

$$\langle f^2 \rangle = \xi(0) = \int \frac{d^3k}{(2\pi)^3} P(k) \quad (3.15)$$

so the total variance of the field can be obtained by integrating the power spectrum.

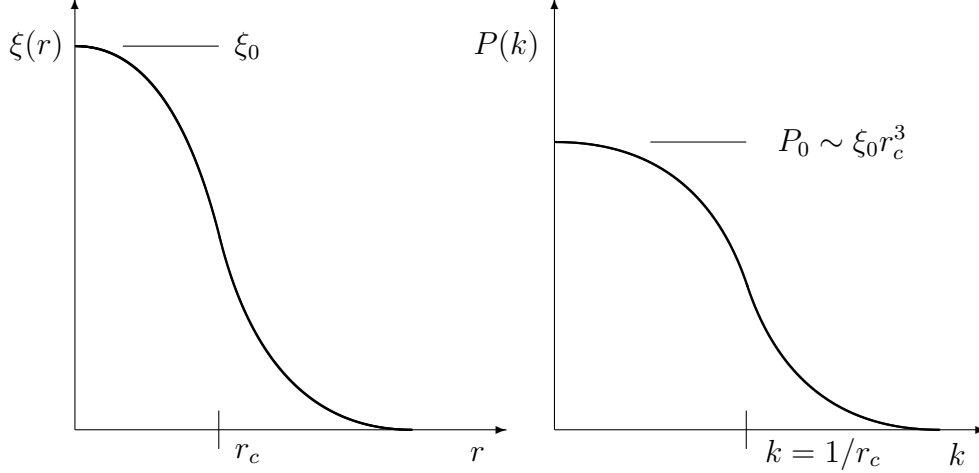


Figure 3.1: Illustration of the relationship between the auto-correlation function $\xi(r)$ and the power spectrum $P(k)$ for the important case where the former has a bell-shaped form with constant asymptote as $\mathbf{r} \rightarrow 0$ and with width r_c .

- If the field is statistically isotropic then the power spectrum depends only on the modulus of the wave vector: $P(\mathbf{k}) = P(k)$.
- If $f(\mathbf{r})$ is dimensionless then so is $\xi(\mathbf{r})$ and therefore from (3.14) $P(\mathbf{k})$ has units of volume. This is for fields in three spatial dimensions. Similarly, the power spectrum of some temporal process $f(t)$ has units of time etc.
- One sometimes sees the power expressed as $\Delta^2(k) = k^3 P(k)/2\pi^2$, in terms of which the field variance is $\langle f^2 \rangle = \int d \ln k \Delta^2(k)$, so $\Delta^2(k)$ gives the contribution to the field variance per log-interval of wave number, and we have $\Delta^2(k) \simeq \xi(r \sim 1/k)$.
- If the field $f(\mathbf{r})$ is *incoherent*, by which we mean that the values of the field at different points are uncorrelated, so $\xi(\mathbf{r}) \propto \delta(\mathbf{r})$, then the power spectrum is constant. Such fields are referred to as *white-noise*.
- If the auto-correlation function is a bell-shaped function with *coherence length* r_c , then from (3.14) the power spectrum will be flat for $k \ll 1/r_c$, since then $e^{i\mathbf{k} \cdot \mathbf{r}} \simeq 1$ where ξ is non-negligible. The value of the power at these low frequencies is then $P(k \ll 1/r_c) \simeq \int d^3r \xi(r) \simeq \xi(0)r_c^3$. For $k \gg 1/r_c$ the power will be small since we then have many oscillations of $e^{i\mathbf{k} \cdot \mathbf{r}}$ within $r \sim r_c$ which will tend to cancel. These results are illustrated in figure 3.1.

3.4 Measuring the Power Spectrum

It is illuminating to consider estimating the power spectrum from a finite sample of the infinite random field. We can write such a sample as $f_s(\mathbf{r}) = W(\mathbf{r})f(\mathbf{r})$ where the function $W(\mathbf{r})$ describes the sample volume geometry. For concreteness, imagine W to be unity within a cubical sample volume of side L and $W = 0$ otherwise.

The Fourier transform of the finite sample is, from the convolution theorem, the convolution of the transforms \tilde{f} and \tilde{W} :

$$\tilde{f}_s(\mathbf{k}) = \int \frac{d^3k'}{(2\pi)^3} \tilde{f}(\mathbf{k}') \tilde{W}(\mathbf{k} - \mathbf{k}'). \quad (3.16)$$

This says that the transform of the sample is a somewhat smoothed version of the intrinsically incoherent $\tilde{f}(\mathbf{k})$. The width of the smoothing function \tilde{W} is $\Delta k \sim 1/L$, so the transform $\tilde{f}_s(\mathbf{k})$

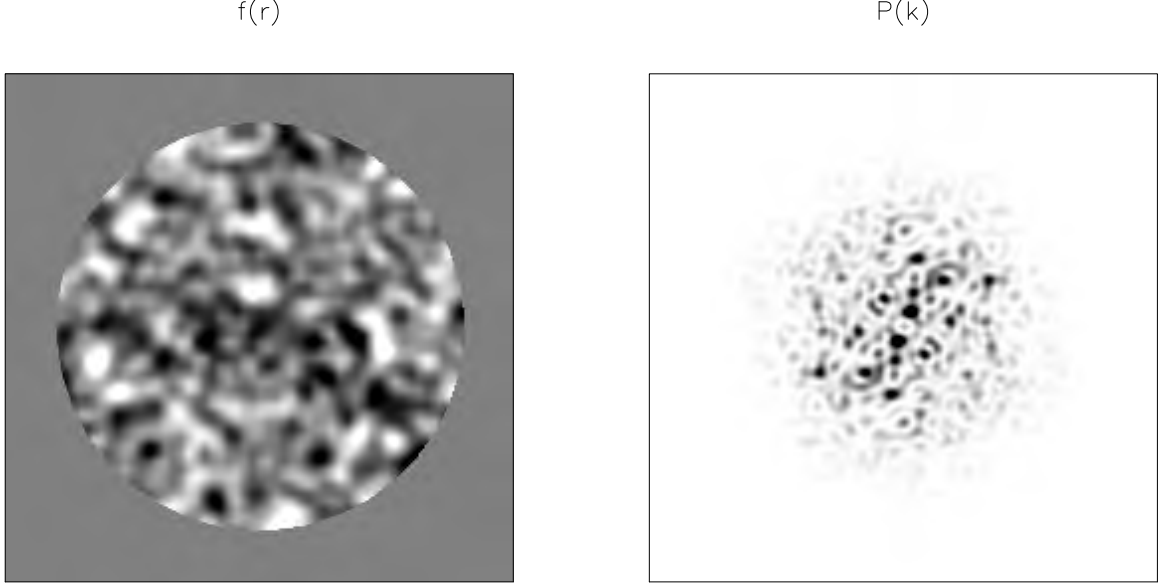


Figure 3.2: The left panel shows a realization of 2-dimensional random field $f(\mathbf{r})$ ‘windowed’ by the function $W(\mathbf{r})$ which is a disk of radius R . The random field is Gaussian white noise filtered by smoothing with a Gaussian kernel with scale length $\sigma \ll R$. The right-hand panel shows the power spectrum. The ‘speckly’ nature of the power spectrum is readily apparent. The overall extent of the power spectrum is $k_{\max} \sim 1/\sigma$ while the size of the individual speckles is on the order $\Delta k \sim 1/R$. The speckly nature of the power spectrum is not specific to the choice of Gaussian random fields. As one makes the window function — or survey size — larger, the size of the speckles decreases. The fractional precision with which one can measure the power averaged over some region of k -space is $\Delta P/P \simeq 1/\sqrt{N}$ where N is the number of speckles. Estimating the statistical uncertainty in power spectra is essentially a counting exercise.

will be coherent over scales $\delta k \ll 1/L$ but will be incoherent on larger scales. The act of sampling introduces finite range correlations in the transform.

If we square the sample transform and take the expectation value we find

$$\langle |\tilde{f}_s(\mathbf{k})|^2 \rangle = \int \frac{d^3 k'}{(2\pi)^3} \int \frac{d^3 k''}{(2\pi)^3} \tilde{W}(\mathbf{k} - \mathbf{k}') \tilde{W}^*(\mathbf{k} - \mathbf{k}'') \langle \tilde{f}(\mathbf{k}') \tilde{f}^*(\mathbf{k}'') \rangle \quad (3.17)$$

and using (3.13) allows one to evaluate one of the integrals to give

$$\langle |\tilde{f}_s(\mathbf{k})|^2 \rangle = \int \frac{d^3 k'}{(2\pi)^3} |\tilde{W}(\mathbf{k} - \mathbf{k}')|^2 P(\mathbf{k}') \quad (3.18)$$

ie a convolution of the power spectrum with $|\tilde{W}(\mathbf{k})|^2$.

For naturally occurring random fields it is often the case that the power spectrum $P(\mathbf{k})$ is a smoothly varying function, with fractional change in the power at two points $\mathbf{k}, \mathbf{k}' = \mathbf{k} + \delta\mathbf{k}$ being small, provided that $\delta k \ll k$. For spatial frequencies $k \gg 1/L$ then $P(\mathbf{k})$ will be effectively constant over the range of frequencies $|\mathbf{k} - \mathbf{k}'| \lesssim 1/L$ that $|\tilde{W}(\mathbf{k} - \mathbf{k}')|^2$ is non negligible, and therefore

$$\langle |\tilde{f}_s(\mathbf{k})|^2 \rangle \simeq P(\mathbf{k}) \int \frac{d^3 k'}{(2\pi)^3} |\tilde{W}(\mathbf{k}')|^2 = P(\mathbf{k}) \int d^3 r W^2(\mathbf{r}) = L^3 P(\mathbf{k}) \quad (3.19)$$

where we have used Parseval’s theorem. This means that a fair approximate *estimator* of the power is

$$\hat{P}(\mathbf{k}) \simeq L^{-3} |\tilde{f}_s(\mathbf{k})|^2. \quad (3.20)$$

It is interesting to note that this estimator $\hat{P}(\mathbf{k})$ does *not* really converge to the true spectrum as the sample volume tends to infinity. This is because of the ‘grainy’ or ‘blotchy’ nature of \tilde{f}_s already commented on (see figure 3.2). What happens is that the number of grains within some *finite* volume of frequency space becomes larger as the volume increases, so the *average* of $\hat{P}(\mathbf{k})$ over this finite region will tend to the true value, but at the microscopic scale $\hat{P}(\mathbf{k})$ still fluctuates from point to point with $\langle (\hat{P}(\mathbf{k}) - P(\mathbf{k}))^2 \rangle^{1/2} \simeq P(\mathbf{k})$. This means that one can simply estimate the fractional uncertainty in the estimated power simply by counting the number of independent cells.

The above results largely apply for arbitrary sample volume $W(\mathbf{r})$, and have a bearing on the design of experiments to measure the power spectrum for e.g. galaxy clustering. Say you can afford to sample a certain finite volume of space, but can choose how to lay out a set of survey fields. A contiguous cubic survey would give $|\tilde{W}(\mathbf{k})|^2$ as a 3-dimensional sinc function with a ‘central lobe’ of width $\delta k \sim 1/L$. Laying out the fields in a broader grid may be advantageous as this will decrease the width of the central lobe and increase the number of effectively independent samples of the power considerably. However, it will also tend to result in extended side-lobes in the ‘window function’ $|\tilde{W}(\mathbf{k})|^2$ and, if one is trying to measure power at low frequencies, the signal will be contaminated by power aliased from high frequencies. The mean value of the aliased power can be estimated and subtracted, but the fluctuations in the aliased power increase the noise in one’s measurement. The optimal sampling strategy depends on the actual power spectrum.

3.5 Moments of the Power Spectrum

Consider a 1-dimensional field $f(r)$. The variance of the field is

$$\langle f^2 \rangle = \int \frac{dk}{2\pi} P(k) = \xi(0) \quad (3.21)$$

so the variance is the *zeroth moment* of the power, and is also the auto-correlation function at zero lag.

Higher moments of the power spectrum are also of considerable physical significance. Consider a random field $F(r)$ which is the derivative of $f(r)$:

$$F(r) = f'(r) \equiv \frac{df}{dr}. \quad (3.22)$$

Taking the derivative in real space corresponds to multiplying by ik in Fourier space, so

$$\tilde{F}(k) = ik\tilde{f}(k) \quad (3.23)$$

and the power spectrum of F is

$$P_F(k) = k^2 P_f(k). \quad (3.24)$$

This means that the variance of the gradient f' is

$$\langle f'^2 \rangle = \langle F^2 \rangle = \int \frac{dk}{2\pi} k^2 P(k) \quad (3.25)$$

which is the second moment of the power spectrum. We can also write this in terms of the auto-correlation function since if $P(k)$ is the transform of $\xi(r)$ then $k^2 P(k) = -(ik)^2 P(k)$ is minus the transform of the second derivative of $\xi(r)$ so

$$\langle f'^2 \rangle = \int \frac{dk}{2\pi} k^2 P(k) = - \left(\frac{\partial^2 \xi}{\partial r^2} \right)_{r=0}. \quad (3.26)$$

We can also compute such quantities as $\langle f f' \rangle$ which is proportional to $\int dk k P(k)$, but for isotropic fields $P(k)$ is an even function while k is odd, so $\langle f f' \rangle = 0$. For such fields the covariance matrix $\langle f^{(n)} f^{(m)} \rangle$ for derivatives of order m and n is non-zero only if $n - m$ is even. For example, the co-variance of the field and its second derivative is

$$\langle f f'' \rangle = - \int \frac{dk}{2\pi} k^2 P(k) = \left(\frac{\partial^2 \xi}{\partial r^2} \right)_{r=0}. \quad (3.27)$$

which, aside from the sign, is identical to $\langle f'^2 \rangle$.

3.6 Variance of Smoothed Fields

The power spectrum is very useful for computing the variance of fields which have been smoothed or filtered with a *smoothing kernel* $W(\mathbf{r})$, a common example of which being the point spread function of an instrument.

If the smoothed field is

$$f_S(\mathbf{r}) \equiv \int d^3r' f(\mathbf{r} - \mathbf{r}')W(\mathbf{r}') \quad (3.28)$$

then, by the convolution theorem, its transform is

$$\tilde{f}_S(\mathbf{k}) = \tilde{f}(\mathbf{k})\tilde{W}(\mathbf{k}). \quad (3.29)$$

The power spectrum of f_S is then

$$P_{f_S}(\mathbf{k}) = |\tilde{W}(\mathbf{k})|^2 P_f(\mathbf{k}) \quad (3.30)$$

and therefore the variance of the smoothed field is

$$\langle f_S^2 \rangle = \int \frac{d^3k}{(2\pi)^3} |\tilde{W}(\mathbf{k})|^2 P_f(\mathbf{k}). \quad (3.31)$$

This is equivalent to (3.6) but is a single rather than double integration.

3.7 Power Law Power Spectra

Many physical processes give rise to fields with power spectra which can be approximated as power laws in temporal or spatial frequencies,

$$P(\omega) \propto \omega^n \quad \text{or} \quad P(k) \propto k^n \quad (3.32)$$

where n is the *spectral index*. As we have seen, an incoherent process in which the field values at different places or times are uncorrelated gives a flat or ‘white’ spectrum with $n = 0$. Spectra with indices $n > 0$ (with more power at high frequencies than is) are sometimes said to be blue while those with $n < 0$ are called red. Some examples are shown in figure 3.3.

An example of a red spectrum is ‘Brownian noise’ obtained by integrating an incoherent process. Physical realizations include the ‘drunkards walk’, and the displacement of a molecule being buffeted by collisions in a gas. The displacement as a function of time has a spectrum $P(\omega) \propto \omega^n$ with $n = -2$.

Another example of a very red spectrum is the phase fluctuation in the wavefront from a distant source introduced by atmospheric turbulence. This is a two-dimensional field with spectral index $n = -11/3$.

Note that for a process with spectral index more negative than minus the number of dimensions N , as is the case in the two above examples, the auto-correlation function ξ is ill-defined since $\xi(0) \sim \int d^N k k^n$ and this integral does not converge at low frequencies. This is not usually a serious problem, since the power-law may only be obeyed over some range of frequencies, and there may be some physical cut-off (such as the ‘outer-scale’ in atmospheric turbulence) which renders the variance finite. In such cases it is more useful to define a *structure function*

$$S(\mathbf{r}) \equiv \langle (f(\mathbf{r}) - f(0))^2 \rangle = 2(\xi(0) - \xi(\mathbf{r})). \quad (3.33)$$

The infra-red divergence renders both $\xi(0)$ and $\xi(\mathbf{r})$ formally infinite, but the difference is well defined provided $n > -(N + 2)$ and has a power-law form $S(r) \propto r^{-(N+n)}$. For atmospheric turbulence, for instance, $N = 2$, $n = -11/3$ and the structure function for phase fluctuations is $S_\varphi(r) \propto r^{5/3}$.

One particularly interesting class of processes are so-called ‘flicker-noise’ processes with $n = -N$. For a temporal process, for instance, this would be $n = -1$, and such fields are often referred to as

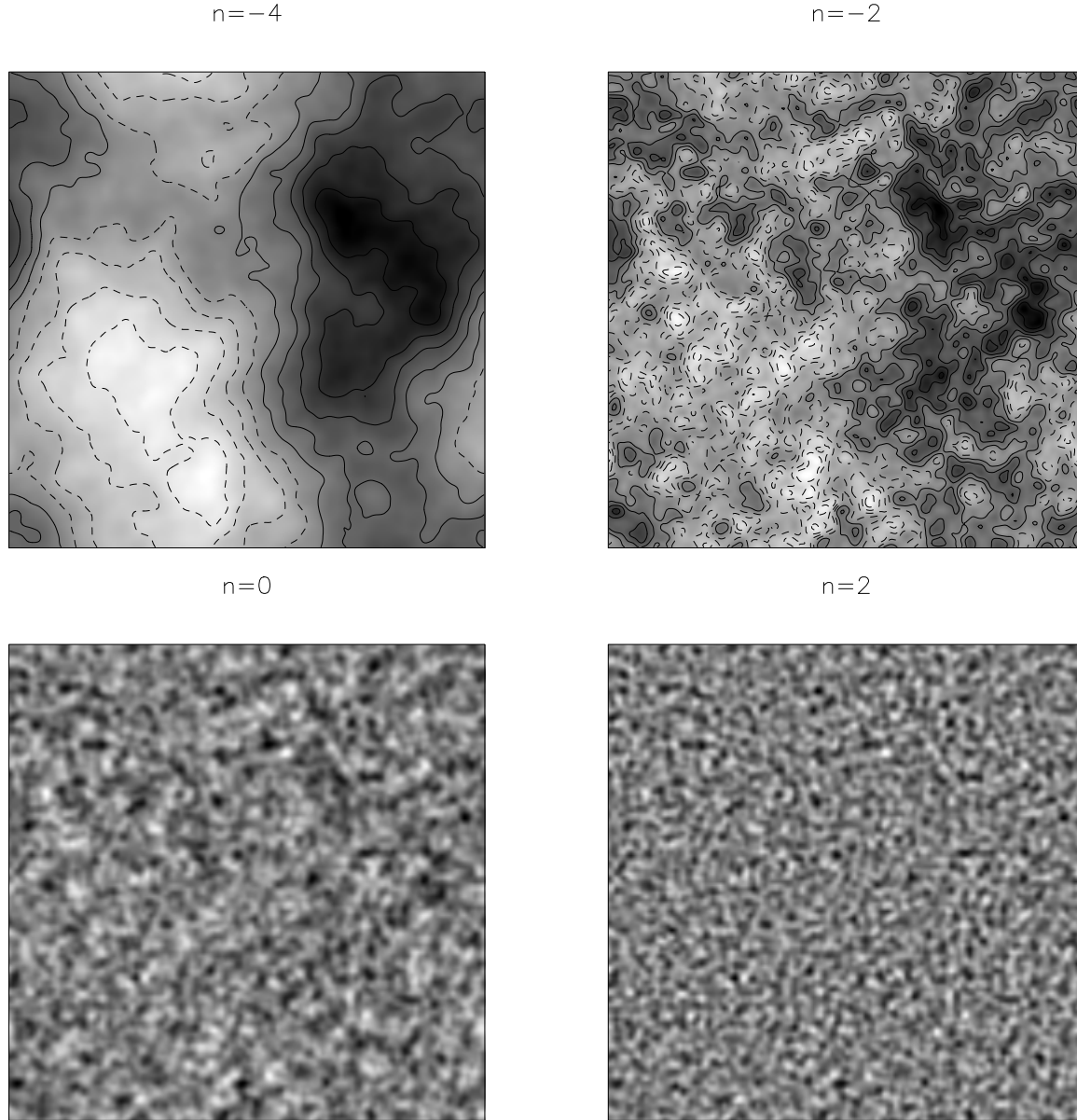


Figure 3.3: Examples of 2-dimensional random fields with power-law like power spectra. These were generated by first generating a white noise Gaussian random field — see below — and then smoothing it with a small Gaussian kernel of scale-length σ . This smoothing introduces a coherence in the field. The Fourier transform of the white noise field was then multiplied by $k^{n/2}$ and the result was inverse transformed. The upper right panel shows 2-dimensional ‘flicker-noise’; such fields have the same variance on all scales. The lower left panel is ‘white-noise’. The lower right panel has $n = 2$ and is more uniform on large scales than white noise. Such fields, with $n > 0$ that is, have $P(k) \rightarrow 0$ as $k \rightarrow 0$. Since $P(k)$ is the transform of $\xi(r)$ this means that such fields have $\int d^2r \xi(r) = 0$. The auto-correlation function $\xi(\mathbf{r})$ for our $n = 2$ examples has a positive peak of width $\sim \sigma$ around $\mathbf{r} = 0$, but is surrounded by a compensating ‘moat’ where $\xi(\mathbf{r})$ is negative. The field values at pairs of points separated by a few times the coherence length are anti-correlated.

having a ‘ $1/f$ ’ spectrum, since the power spectrum scales inversely with frequency ($f \rightarrow \omega$ in our notation). In this case the contribution to the variance of the field is

$$\langle f^2 \rangle = \int \frac{d\omega}{2\pi} P(\omega) \sim [\ln \omega]_{\omega_1}^{\omega_2}. \quad (3.34)$$

Such fields have the characteristic that there is equal contribution to the variance per logarithmic interval of frequency, and are often said to be ‘scale-invariant’. One can generate something approximating a flicker noise process by drawing a wiggly curve with some long coherence length and some rms amplitude, and then adding to this another wiggly curve with the same amplitude but with half the coherence length and so on.

Physical processes that have been claimed to approximate flicker noise include:

- The brightness of quasars as a function of time.
- The intensity of classical music (when averaged over a time-scale much longer than the period of acoustic waves). This presumably reflects the interesting ‘hierarchical’ structure of such music, with notes of varying strength, phrases of varying strength, movements of varying strength. It is not the case for some other types of music.
- The fluctuation in the resistance of carbon resistors as a function of time.
- The large-angle fluctuations in the microwave background as revealed by the COBE satellite.
- Seismic noise.
- Deflection of light by gravity waves.
- Telescope mirror roughness.
- Variation in height of high tides over long periods of time.

Not all of these are well understood (see Press article).

3.8 Projections of Random Fields

It is often the case that we observe a projection of some random field onto a lower dimensional space. Examples include the 3-dimensional galaxy density field projected onto the 2-dimensional sky, and the 3-dimensional atmospheric refractive index fluctuations projected onto the 2-dimensional wavefront. It is useful to be able to transform the power spectrum and auto-correlation function in the different spaces, either to predict the observed power, or, though this is generally more difficult, to *de-project* the observed power to reconstruct the power in the higher dimensional case.

Consider a planar projection

$$F(x, y) = \int dz W(z) f(x, y, z) \quad (3.35)$$

where $W(z)$ is a normalized ‘box-car’ function of width Δz and height $1/\Delta z$, so (3.35) gives the average of $f(x, y, z)$ through a slab.

If the field $f(\mathbf{r})$ has a well defined coherence length r_c , and variance $\langle f^2 \rangle$, then one can crudely picture the field as a set of contiguous domains or cells of size $\sim r_c$ within each of which the field is assigned a constant, but randomly chosen, value with amplitude $f \sim \langle f^2 \rangle^{1/2}$. In this model, and if the coherence length is small compared to the slab thickness, the projected field will be the average over $N = \Delta z/r_c$ domains so we expect

$$\langle F^2 \rangle \simeq \frac{1}{N} \langle f^2 \rangle \simeq \frac{r_c}{\Delta z} \langle f^2 \rangle. \quad (3.36)$$

The coherence scale of the projected field will be similar to that of the unprojected field, so we expect

$$\xi_{2D}(0) \simeq \frac{r_c}{\Delta z} \xi_{3D}(0) \quad (3.37)$$

In the same spirit, one might model a more general field with some 3-dimensional correlation function $\xi_{3D}(r)$ as the superposition of components with coherence length r and $\langle f^2 \rangle \simeq \xi_{3D}(r)$, for which we should have

$$\xi_{2D}(r) \simeq \frac{r}{\Delta z} \xi_{3D}(r). \quad (3.38)$$

For a power law $\xi_{3D}(r) \propto r^{-\gamma_{3D}}$ then the 2-dimensional correlation function will also be a power law with $\gamma_{2D} = \gamma_{3D} - 1$. This is all assuming the slab is thick. If the slab is small compared to the coherence length one expects $\xi_{2D}(r) = \xi_{3D}(r)$, so for a finite slab, one would expect the 2-dimensional auto-correlation function to have a locally power behavior for both $r \ll \Delta z$ and $r \gg \Delta z$, with slopes $\gamma_{2D} = \gamma_{3D} - 1$ and $\gamma_{2D} = \gamma_{3D}$ respectively.

This order of magnitude result can be made more precise. The 2-dimensional auto-correlation is

$$\begin{aligned} \xi_{2D}(x, y) &= \langle F(x' + x, y' + y) F(x', y') \rangle \\ &= \int dz \int dz' W(z) W(z') \langle f(x' + x, y' + y, z) f(x', y', z') \rangle \\ &= \int dz \int dz' W(z) W(z') \xi_{3D}(\sqrt{x^2 + y^2 + (z - z')^2}) \end{aligned} \quad (3.39)$$

and if $r = \sqrt{x^2 + y^2} \ll \Delta z$ we find

$$\xi_{2D}(r) \simeq \int dz' W^2(z') \int dz \xi_{3D}(\sqrt{r^2 + z^2}) = \frac{1}{\Delta z} \int dz \xi_{3D}(\sqrt{r^2 + z^2}). \quad (3.40)$$

This allows one to transform from 3-D to 2-D. For a power law, this integral gives

$$\xi_{2D}(r) \simeq \frac{r}{\Delta z} \xi_{3D}(r) \quad (3.41)$$

provided $\gamma > 1$, which is just the condition that the slope should be such that $\xi_{2D}(r)$ be a decreasing function of r . This agrees with the random-walk argument above.

The generalization of this approach to 3-dimensional spherical geometry (with the observer at the origin) leads to what is called *Limber's equation*.

One can also obtain a similar transformation law for the power spectrum. Imagine one generates a 3-D random field as a sum of sinusoidal waves with appropriately chosen wavelengths $\lambda \sim 2\pi/k^*$ and amplitudes $\tilde{f}_{\mathbf{k}} \propto \sqrt{P(k^*)}$. In projection, and assuming $k^* \Delta z \gg 1$, most of these waves will suffer strong attenuation as positive and negative half cycles cancel one another. The only modes which are not attenuated are those such that the phase along each line of sight varies by less than a radian or so, or equivalently, those modes with $k_z \Delta z \lesssim 1$. These modes have wave-vector nearly perpendicular to the line of sight through the slab. Thus projecting through a thick slab of thickness Δz in real-space has the effect of selecting modes in a narrow slice $\delta k_z \lesssim 1/\Delta z$ in Fourier space. One therefore has the simple result that

$$P_{2D}(k) \simeq P_{3D}(k)/\Delta z \quad (3.42)$$

where the constant of proportionality $1/\Delta z$ is consistent with the requirement that P_{ND} should have dimensions of $(\text{length})^N$.

For a power law spectrum $P \propto k^n$ the slope is the same in both three and two dimensions. This is in accord with the reasonable requirement that if the field is incoherent in 3-D, so $n = 0$, then the projected field should also be incoherent. It is also in accord with the results for transforming the auto-correlation function above since if $P(k) = P_*(k/k_*)^n$ then

$$\xi_{ND} = \int \frac{d^N k}{(2\pi)^N} P(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{r}} \sim P_* k_*^{-n} r^{-(n+N)} \int d^N y y^n e^{iy}. \quad (3.43)$$

The integral here is dimensionless and generally of order unity, so this tells us that a spectral index n corresponds to a correlation function slope $\gamma = -(n + N)$ so, for instance, $\gamma_{2D} = \gamma_{3D} + 1$ as before.

These results are true for a wide range of spectral indices, though certain values such as $n = 0$ are special cases ($n = 0$ does *not* correspond to $\xi_{3D} \propto r^{-3}$ for instance). What happens in this case is that the value of the dimensionless integral vanishes. For the integral to have a finite low-frequency contribution requires $n > -N$. As discussed, for redder (ie more negative) indices, ξ is ill defined, but the structure function

$$S(r) = 2(\xi(0) - \xi(r)) = 2 \int \frac{d^N k}{(2\pi)^N} P(\mathbf{k})(1 - e^{i\mathbf{k}\cdot\mathbf{r}}) \quad (3.44)$$

is well-defined providing $n > -(2 + N)$.

3.9 Gaussian Random Fields

A significant special class of random fields are *Gaussian random fields*, examples of which are legion. Gaussian statistics arise whenever one has a random process which is the sum of a large number of independent disturbances. Gaussian random fields also arise from the quantum-mechanics of the early universe, and such fields play a major role in cosmology.

3.9.1 Central Limit Theorem

Consider a random variate Y which is the sum of a large number of random components

$$Y = \sum_{i=1}^N X \quad (3.45)$$

where the X values are drawn randomly and independently from some probability distribution function $p(X)$. The central limit theorem states that for large N , and provided $p(X)$ satisfies certain reasonable conditions, the probability distribution for Y tends to a universal form

$$p_N(Y)dY = \frac{dY}{\sqrt{2\pi}\sigma_Y} \exp(-Y^2/2\sigma_Y^2) \quad (3.46)$$

where

$$\sigma_Y^2 = N\sigma_X^2 \quad \text{and} \quad \sigma_X^2 \equiv \int dX X^2 p(X). \quad (3.47)$$

We can prove this by induction. Let Y denote the sum of N random X values, and Y' the partial sum of the first $N - 1$ values, so $Y = Y' + X_N$. The probability of some Y is the sum over all Y' of $p_{N-1}(Y')$ times the probability that $X_N = Y - Y'$, or

$$p_N(Y) = \int dY' p_{N-1}(Y') p(Y - Y'). \quad (3.48)$$

This is a convolution, so the Fourier transform of the probability distribution, also called the *generating function*, is

$$\tilde{p}_N(\omega) = \tilde{p}_{N-1}(\omega) \tilde{p}(\omega) \quad (3.49)$$

and since for $N = 1$ we have $p_1(\omega) = p(\omega)$,

$$\tilde{p}_N(\omega) = \tilde{p}(\omega)^N. \quad (3.50)$$

This means that $p_N(\omega)$ is maximized where $p(\omega)$ is maximized, ie at $\omega = 0$, but will tend to be much more tightly peaked, and so will only depend on the form of $\tilde{p}(\omega)$ very close to the origin. Expanding the complex exponential factor in the Fourier transform $\tilde{p}(\omega)$ for $\omega \ll 1/\sigma_X$ gives

$$\tilde{p}(\omega) = \int dX p(X) e^{i\omega X} = \int dX p(X) (1 + i\omega X - \omega^2 X^2/2 + \dots). \quad (3.51)$$

The first term here is unity. The second term is $i\omega\langle X \rangle = 0$. The third term is $-\omega^2\langle X^2 \rangle/2 = -\omega^2\sigma_X^2/2$ and therefore

$$\tilde{p}(\omega) \simeq 1 - \omega^2\sigma_X^2/2 \quad \text{for } \omega\sigma_X \ll 1 \quad (3.52)$$

Raising this to the N th power gives

$$\tilde{p}_N(\omega) = \tilde{p}(\omega)^N = (1 - \omega^2\sigma_X^2/2)^N \rightarrow \exp(-\omega^2 N\sigma_X^2/2) \quad (3.53)$$

provided N is large. Performing the inverse transform $\tilde{p}_N \rightarrow p_N$ we obtain

$$p_N(Y) = \frac{1}{\sqrt{2\pi N\sigma_X^2}} \exp(-Y^2/2N\sigma_X^2) \quad (3.54)$$

QED.

3.9.2 Multi-Variate Central Limit Theorem

Consider a random vector $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_m\}$ which is the sum of N random vectors \mathbf{X}

$$\mathbf{Y} = \sum_N \mathbf{X}. \quad (3.55)$$

The generating function is now

$$\tilde{p}(\boldsymbol{\omega}) = \int d^m X p(\mathbf{X}) e^{i\boldsymbol{\omega} \cdot \mathbf{X}}. \quad (3.56)$$

Expanding the exponential gives

$$e^{i\boldsymbol{\omega} \cdot \mathbf{X}} = 1 + i\omega_i X_i - \omega_i X_i \omega_j X_j / 2 \dots \quad (3.57)$$

so assuming we have, as before, chosen the origin such that $\langle \mathbf{X} \rangle = 0$, we have

$$\tilde{p}(\boldsymbol{\omega}) = 1 - m_{ij} \omega_i \omega_j / 2 + \dots \quad (3.58)$$

with

$$m_{ij} \equiv \int d^m X X_i X_j p(\mathbf{X}) = \langle X_i X_j \rangle \quad (3.59)$$

the *covariance matrix* for the \mathbf{X} vectors. The same argument as above gives the generating function for \mathbf{Y} as

$$\tilde{p}_N(\boldsymbol{\omega}) \simeq (1 - m_{ij} \omega_i \omega_j / 2)^N = \exp(-M_{ij} \omega_i \omega_j / 2) \quad (3.60)$$

with

$$M_{ij} = N m_{ij} \quad (3.61)$$

and inverse transforming $\tilde{p}_N(\boldsymbol{\omega})$ gives

$$p_N(\mathbf{Y}) d^m Y = \frac{d^m Y}{\sqrt{(2\pi)^m |M|}} \exp(-Y_i M_{ij}^{-1} Y_j / 2). \quad (3.62)$$

This is most easily verified by working in a ‘rotated frame’ such that M_{ij} is diagonal, but is true in general since the determinant $|M|$ and the quadratic form $Y_i M_{ij}^{-1} Y_j$ are invariants.

This calculation shows that the probability distribution function for a Gaussian random vector \mathbf{Y} is fully specified by the matrix of covariance elements $M_{ij} \equiv \langle Y_i Y_j \rangle$.

3.9.3 Gaussian Fields

A 1-dimensional Gaussian random field $f(r)$ can be thought of as a very large vector $\mathbf{f} = \{f_1, f_2, f_3, \dots\}$ giving the values of the field f_m at a set of very finely spaced set of points $r_m = m\Delta r$.

If the field is statistically homogeneous then the correspondingly huge covariance matrix is easily generated since $\langle f_n f_m \rangle = \xi(|n - m|\Delta r)$.

A convenient prescription for generating a Gaussian random field on the computer is to make a *Fourier synthesis*

$$f(r) = \sum_k \tilde{f}_k e^{ikr} \quad (3.63)$$

with randomly chosen Fourier amplitudes \tilde{f}_k with $|\tilde{f}_k| \propto \sqrt{P(k)}$. These are complex, but must satisfy $\tilde{f}_{-k} = \tilde{f}_k^*$ to ensure reality of $f(r)$. This prescription clearly satisfies the conditions for the CLT that $f(r)$ be a sum of independent random variables, and it is not particularly critical precisely what distribution function is used to generate the \tilde{f}_k values.

A key feature of a *Gaussian* random field is that *all* of its properties are completely specified by the two point function $\xi(r)$, and therefore by the power spectrum $P(k)$. This is in contrast to a general random field where only all *variance* statistics of the field are thus specified. Two general random fields may have the same two-point function but their higher order correlation functions may differ. For Gaussian random fields, all higher order statistics are uniquely specified by the two-point function $\xi(r)$.

The above results are readily generalized to fields in multi-dimensional spaces.

3.10 Gaussian N -point Distribution Functions

Given the power spectrum or auto-correlation function one can simply write down the N -point distribution function for a GRF:

$$p(f_1, f_2 \dots f_N) d^N f \quad (3.64)$$

or indeed for any set of *linear* functions of the field (including derivatives of arbitrary order etc). All that one needs to do is compute the $N \times N$ covariance matrix elements, each of which can be written as an integral over the power spectrum.

For example, the 2-point distribution function $p(f_1, f_2)$ involves the 2×2 covariance matrix

$$M_{ij} = \begin{bmatrix} \langle f_1^2 \rangle & \langle f_1 f_2 \rangle \\ \langle f_2 f_1 \rangle & \langle f_2^2 \rangle \end{bmatrix} = \begin{bmatrix} \xi(0) & \xi(r) \\ \xi(r) & \xi(0) \end{bmatrix}. \quad (3.65)$$

The determinant is $|M| = \xi_0^2 - \xi_r^2$ and the inverse is

$$M_{ij}^{-1} = \frac{1}{\xi_0^2 - \xi_r^2} \begin{bmatrix} \xi_0 & \xi_r \\ \xi_r & \xi_0 \end{bmatrix} \quad (3.66)$$

so the bi-variate probability distribution is

$$p(f_1, f_2) df_1 df_2 = \frac{df_1 df_2}{2\pi \sqrt{\xi_0^2 - \xi_r^2}} \exp \left(-\frac{\xi_0 f_1^2 - 2\xi_r f_1 f_2 + \xi_0 f_2^2}{2(\xi_0^2 - \xi_r^2)} \right) \quad (3.67)$$

Note that if ξ_r tends to zero for large r then the bivariate distribution function factorizes $p(f_1, f_2) \rightarrow p(f_1)p(f_2)$.

3.11 Gaussian Conditional Probabilities

The conditional probability for the value of the field f_2 at r_2 given a measurement of the field f_1 at r_1 is, by *Bayes' theorem*,

$$p(f_2|f_1) = p(f_1, f_2)/p(f_1) \quad (3.68)$$

or

$$p(f_2|f_1) = \frac{1}{\sqrt{2\pi\xi_0(1-c^2)}} \exp\left(-\frac{(f_2 - cf_1)^2}{2\xi_0(1-c^2)}\right) \quad (3.69)$$

where $c \equiv \xi_r/\xi_0$.

This is a shifted Gaussian distribution with non-zero conditional mean

$$\langle f_2|f_1 \rangle = cf_1 \quad (3.70)$$

so the mean value is close to the conditional value for $r_{12} \ll r_c$ and relaxes to zero for $r_{12} \gg r_c$. The conditional variance is

$$\langle (f_2 - cf_1)^2|f_1 \rangle = \xi_0(1-c^2) \quad (3.71)$$

which is much smaller than the unconstrained variance for $r_{12} \ll r_c$.

3.12 Ricean Calculations

Imagine one is monitoring a noisy radio receiver for pulses and one would like to estimate the frequency of spurious pulse detections simply due to the noise in the output. This problem first arose in telegraphy, and the solution to such problems was first clearly set out by Rice.

A classical ‘Ricean’ calculation is to ask, given a Gaussian random time series $f(t)$, with some specified auto-correlation function $\xi(\tau) \equiv \langle f(t)f(t+\tau) \rangle$, what is the frequency of up-crossings of the level $f = F$?

To answer this, consider the bi-variate probability element

$$dp = p(f = F, f') df df' \quad (3.72)$$

with $f' = df/dt$, the time derivative of the field. This quantity tells us the fraction of time that the following conditions are satisfied

$$\begin{aligned} F &< f < F + df \\ f' &< \frac{df}{dt} < f' + df' \end{aligned} \quad (3.73)$$

These conditions individually specify a set of intervals of time; the former of length $\Delta t \sim df/\sqrt{\langle f'^2 \rangle}$ and the latter of length $\Delta t' \sim df'/\sqrt{\langle f''^2 \rangle}$, and the combined conditions are satisfied within the intersection of these interval sets. Let us choose the infinitesimals so that $\Delta t \ll \Delta t'$ (ie df is an infinitesimal of higher order than df'), so that the length of the interval is determined by the first condition, and

$$\Delta t = df/f'. \quad (3.74)$$

Define $n_{\text{up}}(F, f')df'$ to be the frequency (number per unit time) of up-crossings with $f' < df/dt < f' + df'$. Multiplying this by the length of the intervals must equal the probability element above:

$$n_{\text{up}}(F, f')df'\Delta t = p(f = F, f')dfdf' \quad (3.75)$$

and therefore

$$n_{\text{up}}(F, f') = |f'| P(f = F, f') \quad (3.76)$$

and the total rate of up-crossings is obtained by integrating over all $f' > 0$:

$$n_{\text{up}}(F) = \int_0^\infty df' |f'| P(F, f'). \quad (3.77)$$

This integration is straightforward. As we have seen, $\langle ff' \rangle = 0$, so the joint distribution factorizes into two independent Gaussians $p(f, f') = p(f)p(f')$ with

$$\begin{aligned} p(f) &= (2\pi\xi_0)^{-1/2} \exp(-f^2/2\xi_0) \\ p(f') &= (2\pi(-\xi_0''))^{-1/2} \exp(-f'^2/2(-\xi_0'')) \end{aligned} \quad (3.78)$$

and hence

$$n_{\text{up}}(F) = \frac{1}{2\pi} \sqrt{\frac{-\xi''(0)}{\xi(0)}} \exp(-F^2/2\xi(0)). \quad (3.79)$$

The curvature of $\xi(\tau)$ at $\tau = 0$ therefore determines the characteristic time-scale $\tau^* = \sqrt{\xi_0/\xi''_0}$ and the rate of up-crossings is essentially this rate times the probability that the field has value $f = F$.

This type of calculation can be generalized to fields in multi-dimensional space, and can also be generalized to give the frequency of extrema, or of peaks, the latter playing a big role in cosmological structure formation theory.

As a second example, consider the distribution of heights of extrema. At first sight this is trivial, since this is surely just

$$p_{\text{ext}}(F) = p(f = F | f' = 0) = p(f = F, f' = 0) / p(f' = 0) \quad (3.80)$$

right? And since the field and its first derivative are uncorrelated $\langle ff' \rangle = 0$, so $p(f = F, f' = 0) = p(f = F)p(f' = 0)$ and therefore

$$p_{\text{ext}}(F) = p(f = F). \quad (3.81)$$

This is a very simple result, but also very puzzling, as it says that the distribution of field values at extrema is the same as the unconstrained distribution of field values, whereas common sense seems to indicate that the distribution of extremal values should be, well, more extreme. To bolster one's confidence in this intuitive feeling consider the case of band limited noise with power spectrum $P(\omega)$ which vanishes outside of some narrow range of frequency of width $\delta\omega$ around some central frequency ω_0 . A realization of such a process takes the form of a locally sinusoidal wave of frequency ω_0 , and with slowly varying envelope. Over some interval of length $\delta t \ll 1/\delta\omega$, where the peak amplitude has some nearly constant value f_{max} , the mean square value of the field is $f_{\text{max}}^2/2$ whereas the mean square value of the field at extrema is just f_{max}^2 which is surely inconsistent with $p_{\text{ext}}(F) = p(f = F)$.

To see what is wrong with the above analysis one needs to examine more carefully what is meant by $p(f = F, f' = 0)$. By itself, this is quite abstract, but multiplied by infinitesimals df , df' the meaning is clear:

$$p(f = F, f' = 0) df df' \quad (3.82)$$

gives the fraction of time that the field and its derivative lie in the prescribed ranges, or the fraction of time occupied by a set of intervals. Now in the vicinity of an extremum at t_0 say the value of the derivative is $f' = 0 + (t - t_0)f''$, so the the length of the interval will be inversely proportional to the second derivative $\delta t = df'/f''$, which in turn is (anti)-correlated with the field f . The simple conditional probability $p(f = F | f' = 0)$ gives the distribution of field values near extrema, but it in a way which gives more weight to those extrema with small f'' .

3.13 Variance of the Median

Consider the common situation: One has obtained a set of CCD images of some object, which one would like to average in order to beat down the noise. Unfortunately, the images contain not just an approximately Gaussian noise component, arising from photon counting statistics, but also a highly non-Gaussian noise component coming from cosmic rays, so one is tempted to take a median of the images rather than a straight average (which, in the absence of cosmic rays, and assuming homogeneous data, would be optimal). What is the penalty in terms of final variance for taking the median rather than the average?

We wish to compute the variance of the median of N independent random variates x . Denote the parent probability distribution by $p(x)$ and the cumulative distribution by $P(x) \equiv \int_{-\infty}^x dx' p(x')$. Without loss of generality we can take the origin in x -space to lie at the median of the parent distribution so that $P(0) = 1/2$.

Consider, for simplicity, the case of even N . The probability that the first $N/2$ samples lie below some value x while the last $N/2$ lie above x is just

$$p_{\text{median}}(x) = (1 - P(x))^{N/2} P(x)^{N/2} \quad (3.83)$$

To get the total probability (that $N/2$ samples lie below x) we sum over all combinations. This introduces factorials, but these are independent of x so the final result is proportional to the above factor.

For large N we expect the median of N samples to lie very close to the median of the parent population, so we make a Taylor series of $p(x)$ around $x = 0$:

$$p(x) \simeq p_0 + p'_0 x + \frac{1}{2} p''_0 x^2 + \dots \quad (3.84)$$

with corresponding expansion for the cumulative distribution $P(x) = 1/2 + \int_0^x dx p(x)$ or

$$P(x) \simeq \frac{1}{2} + p_0 x + \frac{1}{2} p'_0 x^2 + \dots \quad (3.85)$$

Keeping the leading order terms gives

$$p_{\text{median}}(x) \propto (1 + 2p_0 x)^{N/2} (1 - 2p_0 x)^{N/2} = (1 - 4p_0^2 x^2)^{N/2} \simeq \exp(-2Np_0^2 x^2) \quad (3.86)$$

where in the last step we have assumed N is large.

In this limit the distribution is a Gaussian: $p_{\text{median}}(x) \propto \exp(-x^2/2\sigma_{\text{median}}^2)$ with median variance

$$\sigma_{\text{median}}^2 = \frac{1}{4Np_0^2} \quad (3.87)$$

which depends only on $p_0 \equiv p(x=0)$. This is an example of a Gaussian distribution arising in a case where the central limit theorem does not apply.

For the special case of a Gaussian *parent* distribution with variance σ_0^2 we have $p_0 = 1/\sqrt{2\pi}\sigma_0$ so the median variance is

$$\sigma_{\text{median}}^2 = \frac{\pi}{2} \times \frac{\sigma_0^2}{N}. \quad (3.88)$$

which states that the median variance is $\pi/2$ times larger than the variance of the mean.

3.14 Problems

3.14.1 Radiation autocorrelation function

Radiation from a thermal source at temperature T is passed through a band pass filter with (energy) transmission function $\mathcal{T}(\omega) = 1/\omega^2$ for $\omega_0 - \delta\omega < \omega < \omega_0 + \delta\omega$ and falls on a sensitive detector which measure one component the electric field $E(t)$ as a function of time. Assuming that $\omega_0 + \delta\omega \ll kT/h$, what is the normalised autocorrelation function of the measured electric field $c_E(\tau) \equiv \langle E(t)E(t+\tau) \rangle / \langle E(t)^2 \rangle$.

Part II

Radiation

Chapter 4

Properties of Electromagnetic Radiation

In this chapter we introduce the specific intensity and other quantities that describe the energy flux associated with electromagnetic radiation. We show the relation between the intensity and the energy flux, momentum flux, radiation pressure and energy density. We discuss the constancy of the intensity along light rays.

4.1 Electromagnetic Spectrum

Astronomical observations mostly deal with *electromagnetic radiation*. Refraction, diffraction and interference phenomena indicate that this radiation behaves as waves with wavelength λ and frequency ν related by

$$\lambda = c/\nu \quad (4.1)$$

with c the *speed of light*

$$c = 3 \times 10^{10} \text{ cm/s}. \quad (4.2)$$

The *photo-electric effect* shows that energy is given to, or taken from, the radiation field in discrete quanta — photons — with energy

$$E = h\nu \quad (4.3)$$

where h is *Planck's constant*

$$h = 6.6 \times 10^{-27} \text{ erg s}. \quad (4.4)$$

Of great importance is *thermal radiation* as emitted by matter in thermodynamic equilibrium, for which the characteristic photon energy is related to the temperature of the emitting material by

$$T = E/k \quad (4.5)$$

where k is *Boltzmann's constant*

$$k = 1.38 \times 10^{-16} \text{ erg/K}. \quad (4.6)$$

Astronomers deal with radiation covering a wide range of frequencies, conventionally delineated into various regimes:

- Gamma rays: $T \gtrsim 10^{10} \text{ K}$
- X-rays: $10^9 \text{ K} \lesssim T \lesssim 10^6 \text{ K}$
- UV: $10^6 \text{ K} \lesssim T \lesssim 10^5 \text{ K}$
- Optical: $T \sim 3 \times 10^4 \text{ K}$

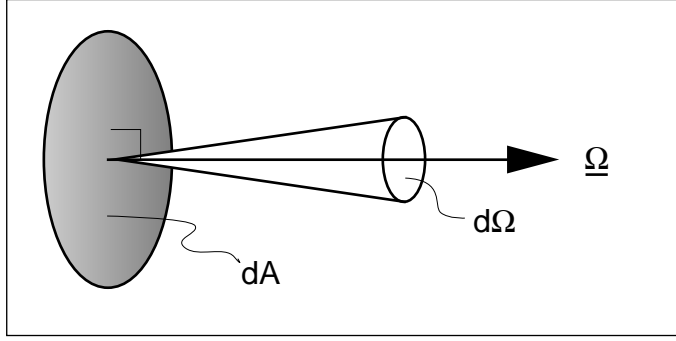


Figure 4.1: Illustration of the geometry for defining the specific intensity

- IR: $10^4\text{K} \lesssim T \lesssim 10^2\text{K}$
- Radio: $10^1\text{K} \lesssim T \lesssim 10^{-1}\text{K}$

4.2 Macroscopic Description of Radiation

4.2.1 The Specific Intensity

The macroscopic properties of freely propagating radiation are well described by *geometric optics* in which radiative energy travels along ‘rays’. One can also picture this as streams of photons carrying energy moving ballistically. The flow of energy along these rays is the *radiative flux*.

In general the radiative energy flux depends on direction Ω and on frequency ν (or wavelength λ). We define the *monochromatic specific intensity* $I_\nu(\Omega)$ (also known as the *brightness* or *surface brightness*) such that if we erect a small surface dA perpendicular to the rays propagating in some small range of directions $d\Omega$ (as illustrated in figure 4.1) then in time dt , those rays in a range of frequencies $d\nu$ around ν will transport through the surface an amount of energy

$$dE = I_\nu(\Omega) dA dt d\Omega d\nu. \quad (4.7)$$

The units of the intensity are therefore $\text{erg}/\text{cm}^2/\text{s}/\text{steradian}/\text{Hz}$.

The intensity provides a fairly complete description of the transport of energy *via* radiation. It can be generalised to describe how the energy is distributed between the different polarization states (see §7.5.3).

4.2.2 Energy Flux

Consider a surface element with some arbitrary orientation, and rays in some small cone $d\Omega$ around direction Ω as illustrated in figure 4.2. The area of the surface projected onto a plane perpendicular to the rays is $dA_\perp = dA \cos \theta$, where θ is the angle between the rays and the normal to the surface, so the *differential flux* (energy per unit area per unit time) for this range of directions is

$$dF_\nu = I_\nu \cos(\theta) d\Omega. \quad (4.8)$$

The *net flux* is obtained by integrating over direction

$$F_\nu = \int d\Omega I_\nu(\Omega) \cos \theta. \quad (4.9)$$

Multiplying some function of direction by the n th power of $\cos \theta$ and integrating is often termed ‘taking the n th moment’ of the function, so we can say that the net flux is the first moment of the intensity. Note that the net flux is zero for isotropic radiation.

Note also that the energy flux is not truly intrinsic to the radiation field, since it depends on the orientation of the surface element. Energy propagating ‘downwards’ through the surface counts as negative energy flux.

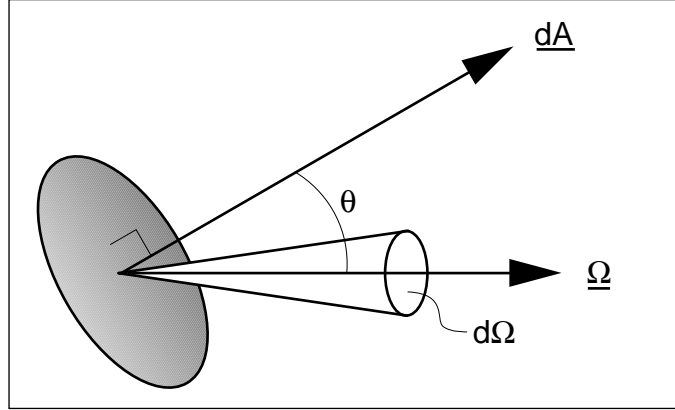


Figure 4.2: Illustration of the geometry of the area and bundle of rays for calculating the contribution to the net flux.

4.2.3 Momentum Flux

Photons carry momentum $\mathbf{p} = \hat{\mathbf{n}}E/c$, where $\hat{\mathbf{n}}$ is a unit vector in the direction of the photon motion, so the component of the momentum in the direction $d\mathbf{A}$ normal to a surface element is $p_{\perp} = |p| \cos \theta$ and so the *differential momentum flux* is $dF_{\nu} \cos \theta$ and integrating over all directions gives the *momentum flux*

$$P_{\nu} \equiv \frac{1}{c} \int d\Omega I_{\nu}(\Omega) \cos^2 \theta \quad (4.10)$$

so the momentum flux is the second moment of the intensity. If the surface element is perpendicular to the x -axis, for example, this quantity gives the rate at which the x component of momentum is being transported in the positive x direction through the surface.

Perhaps surprisingly, the momentum flux does *not* vanish for isotropic radiation; rays propagating up (down) are considered a positive (negative) flux of particles, but carry positive (negative) momentum.

All the above quantities are *monochromatic* and refer to a single frequency. One can obtain the corresponding *total* quantities by integrating over frequency

$$I(\Omega) \equiv \int d\nu I_{\nu}(\Omega) \quad \text{etc...} \quad (4.11)$$

We have used subscript frequency above. One can also define analogous quantities with subscript λ such that e.g. F_{λ} is the energy flux per unit range of wavelength around λ . These are related by e.g.

$$I_{\nu} d\nu = I_{\lambda} d\lambda \quad (4.12)$$

with $\lambda = c/\nu \rightarrow d\lambda = -c/\nu^2 d\nu$ so

$$I_{\lambda} = \nu^2 I_{\nu} / c. \quad (4.13)$$

4.2.4 Inverse Square Law for Energy Flux

Consider an isotropically emitting source. The net rate at which energy crosses an enclosing sphere is independent of the radius of the sphere, and is equal to the product of F and the area, hence

$$F \propto 1/r^2. \quad (4.14)$$

4.2.5 Specific Energy Density

We define the specific energy density as

$$u_{\nu}(\Omega) = \text{energy/volume/solid angle/frequency}. \quad (4.15)$$

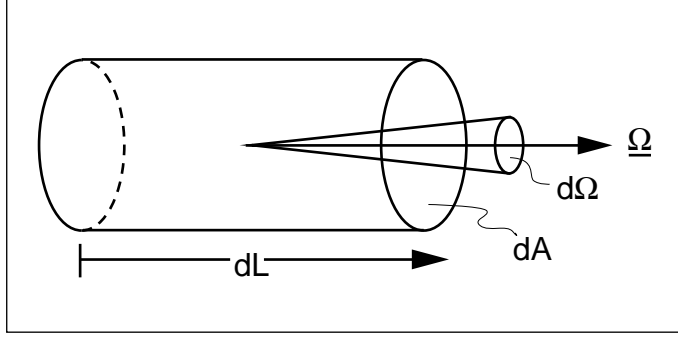


Figure 4.3: Illustration of the relation between the specific intensity and specific energy density.

This is proportional to the intensity. To determine the constant of proportionality, consider a cylinder of length dl and cross-section area dA as illustrated in figure 4.3. The energy enclosed (and in range of direction $d\Omega$ around the axis and in $d\nu$ around ν) is

$$dE = u_\nu(\Omega) d\Omega dl dA d\nu \quad (4.16)$$

and this energy will pass out of the cylinder in a time $dt = dl/c$. However, we also have $I_\nu = dE/(dtdAd\Omega d\nu)$ and hence

$$u_\nu(\Omega) = I_\nu(\Omega)/c. \quad (4.17)$$

This is reasonable; since all photons move at the same velocity c , the rate at which photons in a certain range of directions cross a perpendicular surface is just equal to the photon number density times the speed of light. Similarly, the energy flux for rays in this range of directions — i.e. the brightness — is equal to the energy density times c .

Integrating over direction gives

$$u_\nu = \int d\Omega u_\nu(\Omega) = \frac{1}{c} \int d\Omega I_\nu(\Omega) \quad (4.18)$$

so the energy density is proportional to the zeroth moment of the intensity. Equivalently

$$u_\nu = \frac{4\pi}{c} J_\nu \quad (4.19)$$

where

$$J_\nu \equiv \frac{1}{4\pi} \int d\Omega I_\nu(\Omega) \quad (4.20)$$

is the mean intensity.

4.2.6 Radiation Pressure

For isotropic radiation the radiation pressure is

$$P = u/3. \quad (4.21)$$

The simplest way to see this is to consider a single photon bouncing off the walls of a perfectly reflecting box, compute the rate of transfer of momentum to one of the walls, and then average over all possible directions for the photon (see figure 4.4).

More formally, though also more generally, the rate at which photons of a certain direction bounce off an element of the wall is given by dividing the energy flux by the energy per photon

$$\frac{dn}{dt} = \frac{I_\nu \cos \theta dA d\Omega d\nu}{h\nu}. \quad (4.22)$$

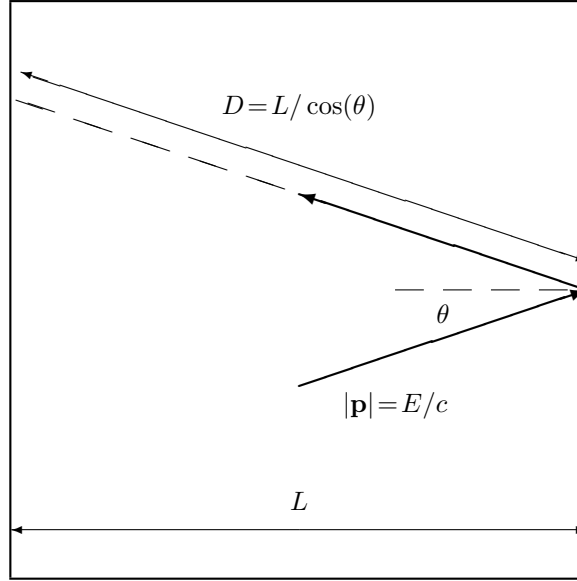


Figure 4.4: Imagine a box with perfectly reflecting walls of side L containing a single photon of energy E . When the photon reflects off the wall at the right the transfer of momentum is $\Delta P = 2(E/c) \cos \theta$. The time between reflections off this wall is $\Delta t = 2L/(c \cos \theta)$. The rate of transfer of momentum to the wall per unit area — i.e. the pressure exerted by the radiation — is $\Delta P/(L^2 \Delta t) = (E/L^3) \cos^2 \theta = u \cos^2 \theta$. For isotropic radiation, $\langle \cos^2 \theta \rangle = 1/3$, so $P = u/3$.

Here θ is the angle between the photon direction and the normal to the wall. In each such reflection the x -component of the momentum is reversed, giving a momentum transfer to the wall $\Delta p = 2(E/c) \cos \theta$, so the rate at which momentum is transferred to the wall per unit area per unit frequency is

$$P_\nu = \frac{\text{force}}{dA d\nu} = \frac{2}{c} \int_{\cos \theta > 0} d\Omega I_\nu \cos^2 \theta. \quad (4.23)$$

The radiation pressure is therefore proportional to the second moment of the intensity. Integrating this over frequency gives the total radiation pressure. For isotropic radiation this is

$$P \equiv \int d\nu P_\nu = \frac{2}{c} \int d\nu J_\nu 2\pi \int d\mu \mu^2 = \frac{u}{3} \quad (4.24)$$

so the radiation pressure is one third of the energy density.

Another way to think about radiation pressure is to consider the change in energy of a box containing standing waves of radiation if we slowly change the volume of the box. If we make a fractional decrease ϵ in the linear size — so $L \rightarrow L' = L(1 - \epsilon)$ then the wavelength scales in the same way, and so the energy of each quantum increases as $h\nu \rightarrow h\nu' = h\nu/(1 - \epsilon) \simeq h\nu(1 + \epsilon)$ (ignoring terms of order ϵ^2 and higher). Thus, if the number of quanta are conserved then the energy also increases by this factor, and the change in energy is $\Delta E = \epsilon E$. This increase of energy had to come from work done in compressing the box against the radiation pressure. Since there are 6 walls, and each moves a distance $\epsilon L/2$, we have $dW = \epsilon E = 6PL^2 \times \epsilon L/2$ and hence $P = E/(3L^3) = u/3$.

4.3 Constancy of Specific Intensity

Consider two circular disks of areas dA_1 and dA_2 with separation R and let both disks be normal to their separation (see figure 4.5). Consider those rays which pass through both disks. The rate at

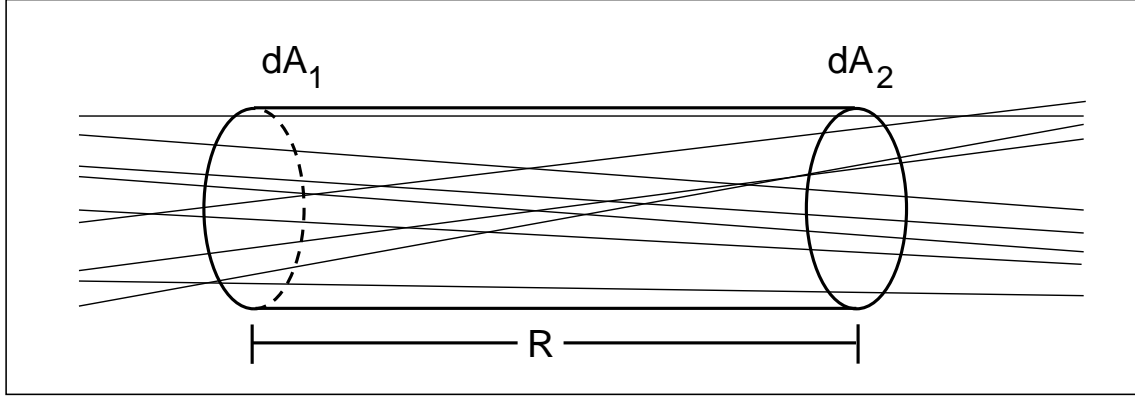


Figure 4.5: Illustration of a sample of rays which pass through both of two small area elements.

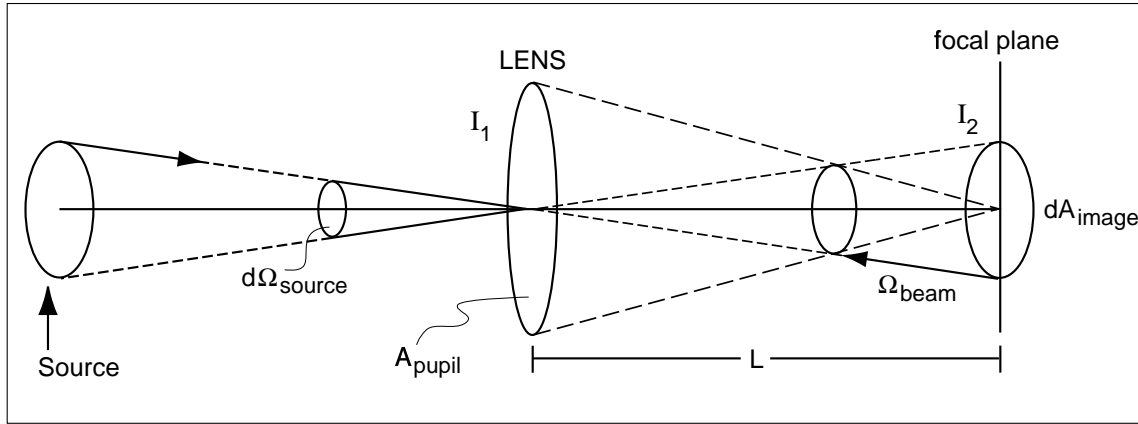


Figure 4.6: An illustration of how intensity is conserved even under extreme focusing of rays. The intensity of light coming into the telescope is I_1 . The energy per unit time entering the telescope pupil is equal to $I_1 A_{\text{pupil}} d\Omega$ where $d\Omega$ is the solid angle of the source. In front of the image (of area $A_{\text{image}} = L^2 d\Omega$) the intensity is I_2 and the radiation is extended over a solid angle $\Omega_{\text{beam}} = A_{\text{pupil}}/L^2$. Equating the rate at which energy is deposited on the focal plane with the rate energy enters the pupil tells us that $I_2 = I_1$.

which energy passes through the disks can be written as

$$\frac{dE}{dt} = I_{\nu 1} dA_1 d\Omega_1 d\nu = I_{\nu 2} dA_2 d\Omega_2 d\nu \quad (4.25)$$

but the solid angle of the cone of rays is $d\Omega_1 = dA_2/R^2$ and $d\Omega_2 = dA_1/R^2$ so

$$I_{\nu 1} = I_{\nu 2} \quad (4.26)$$

so the specific intensity is conserved.

This is consistent with the inverse square law for the energy flux from a isotropically emitting source, since the energy flux is the product of the intensity and the solid angle of the source, and the latter varies inversely with radius squared.

This is for radiation propagating in free space. It also applies to radiation propagating through (static) lenses. A telescope cannot change the intensity; what it does is increase the apparent solid angle of the object and this increases the energy flux falling on the detector (see figure 4.6).

One consequence of the constancy of specific intensity is that we can readily detect even very distant objects such as galaxies, clusters of galaxies, provided they are big enough to be resolved.

Note that for cosmologically distant objects the surface brightness is *not* conserved; rather they suffer from $(1+z)^4$ dimming due to the red-shifting of the radiation. This is not in conflict with the law of constancy of intensity for a static system since an expanding Universe is not static.

Constancy of intensity is also invoked in *Olber's paradox* which says that in an infinite Universe any line of sight should end on a star, so the sky should be bright, not dark. The 'resolution' of this non-paradox involves the finite age of the Universe — there is generally a 'horizon' beyond which we cannot see — and also the $(1+z)^4$ dimming mentioned above.

We will see that the surface brightness of *black body radiation* is proportional to T^4 . If a static optical system *could* change the intensity, this would be equivalent to changing the temperature of the radiation passing through the system. This difference in temperature would allow one to extract useful work from a thermal system, which is forbidden; it would allow one to make a perpetual motion machine.

Chapter 5

Thermal Radiation

We now consider *thermal radiation*, by which we mean radiation emitted by matter which is in thermal equilibrium, and *black body radiation*, which is radiation which is itself in thermal equilibrium (e.g. the radiation inside a kiln, or deep inside a star).

The standard *gedanken* apparatus for generating black-body radiation is a cavity with perfectly reflecting walls and containing some specks of matter which can absorb and re-radiate the photons and thereby allow the radiation to reach equilibrium. (Since photons are massless and do not carry any conserved quantum numbers they can be created or destroyed in the interaction with matter). Our task here is to find the equilibrium distribution of photon energies.

The resulting intensity will, of course, depend on how much energy we put in the cavity. The equilibrium intensity must be isotropic and homogeneous in space, and so should therefore depend only on the energy density. The intensity, which we will denote by $I_\nu = B_\nu(T)$, must be a universal function of frequency parameterized only by the temperature T . Were this not the case — i.e. if the equilibrium intensity were to depend on some other parameter (such as a magnetic field say) then it would be possible to extract useful work from an equilibrated system (see RL), but this is forbidden by the second law of thermodynamics.

Several important properties of black-body radiation — the Stefan-Boltzmann law, the thermodynamic entropy, and the adiabatic expansion laws — can be deduced from purely thermodynamical considerations. We review these results first in §5.1. We then apply statistical mechanical considerations to derive the detailed form of the *Planck spectrum* in §5.2. All of the results derived thermodynamically can, of course, be obtained from the Planck spectrum.

5.1 Thermodynamics of Black Body Radiation

5.1.1 Stefan-Boltzmann Law

Consider a cylinder containing radiation (and a speck of matter in order to allow the radiation to equilibrate), with the possibility of heat input from some external source, and with a piston to connect the radiation mechanically to the outside world and allow the radiation to do work on its environment or *vice versa*. See figure 5.1.

The first law of thermodynamics relates changes dU in the total energy U , heat input dQ and mechanical work done by the system on the outside world PdV :

$$dU = dQ - PdV. \quad (5.1)$$

The change in the *thermodynamic entropy* of the system is

$$dS = \frac{dQ}{T} = \frac{dU}{T} + \frac{PdV}{T} \quad (5.2)$$

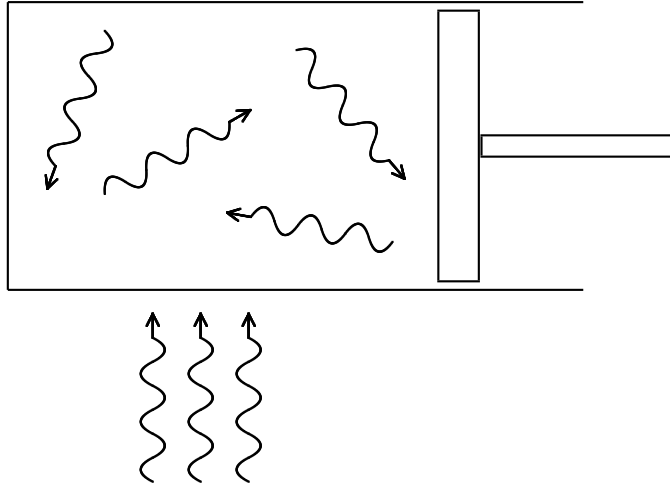


Figure 5.1: A cylinder with perfectly reflecting walls (and perhaps a speck of dust to allow the radiation to thermalize) contains black-body radiation. The piston allows the radiation to interact mechanically with the outside world, and there is also an external source of heat.

with $U = uV$, $P = u/3$ and energy density $u = u(T)$ we have $dU = d(uV) = u dV + V du$ and therefore find

$$dS = \left(\frac{V}{T} \frac{du}{dT} \right) dT + \left(\frac{4}{3} \frac{u}{T} \right) dV. \quad (5.3)$$

However, since the total entropy is a function only of the temperature and volume, we also have

$$dS = \frac{\partial S}{\partial T} dT + \frac{\partial S}{\partial V} dV. \quad (5.4)$$

Equating the coefficients of dT and dV in (5.3) and (5.4) shows that the partial derivatives of the entropy are $\partial S/\partial T = (V/T)(du/dT)$ and $\partial S/\partial V = (4/3)(u/T)$ and equating $\partial^2 S/\partial V \partial T$ and $\partial^2 S/\partial T \partial V$ yields

$$\frac{du}{u} = 4 \frac{dT}{T} \quad (5.5)$$

which has solution $u \propto T^4$ (we are assuming that $u \rightarrow 0$ as $T \rightarrow 0$). Introducing a constant of proportionality a gives us the *Stefan-Boltzmann law*

$$u = aT^4. \quad (5.6)$$

The energy density is related to the brightness by $u = 4\pi B/c$ so we have

$$B_\nu(T) = \frac{acT^4}{4\pi} \quad (5.7)$$

as an alternative statement of the Stefan-Boltzmann law.

Finally, the emergent flux density from a black-body radiator is

$$F = \int_{\cos \theta > 0} d\Omega B \cos \theta = \pi B \quad (5.8)$$

and so

$$F = \sigma T^4 \quad (5.9)$$

where the constants of proportionality in these various versions of the Stefan-Boltzmann law are

$$a = \frac{4\sigma}{c} = 7.56 \times 10^{-15} \text{erg cm}^{-3} \text{K}^{-4} \quad (5.10)$$

and

$$\sigma = \frac{ac}{4} = 5.67 \times 10^{-5} \text{erg cm}^{-2} \text{K}^{-4} \text{s}^{-1}. \quad (5.11)$$

These were originally determined empirically. Below we shall derive them in terms of the fundamental constants h , c .

5.1.2 Entropy of Black-Body Radiation

Start with a cold cavity, with negligible energy, and apply heat at constant volume. The entropy can be obtained by integrating $dS = dQ/T$:

$$S = \int \frac{dQ}{T} = V \int \frac{du}{T} = aV \int \frac{dT^4}{T} = 4aV \int dT T^2 \quad (5.12)$$

or

$$S = \frac{4}{3} aVT^3. \quad (5.13)$$

The entropy scales with the volume, as one could easily have concluded from the fact that Q is an extensive quantity, and it scales as the cube of the temperature.

The radiation energy is $U = aVT^4$. If we assert that the typical photon energy is $E \sim kT$ then the number of photons in a black-body cavity is

$$N \simeq \frac{aVT^4}{kT} = (a/k)VT^3 \quad (5.14)$$

so, for black-body radiation, the entropy is just proportional to the number of photons.

5.1.3 Adiabatic Expansion Laws

We can now compute how the temperature and radiation pressure vary if we change the volume while keeping the system thermally isolated from the external world ($dQ = 0$). At constant S we have

$$T \propto V^{-1/3} \quad \text{and} \quad P \propto T^4 \propto V^{-4/3} \quad (5.15)$$

The *adiabatic equation of state* for black-body radiation is therefore

$$PV^\gamma = \text{constant} \quad (5.16)$$

with *adiabatic index* $\gamma = 4/3$.

The above results are consistent with the picture of the radiation as a conserved number of photons, with wavelength scaling as the linear dimension of the cavity $\lambda \propto L \propto V^{1/3}$, so the energy of each photon scales as $E \propto h\nu \propto 1/L$, and since the characteristic energy is $E \sim kT$ this means $T \propto E \propto V^{-1/3}$.

This tells us that black-body radiation remains black-body under adiabatic expansion — it does not require any matter to maintain this form.

5.2 Planck Spectrum

We now derive the *Planck spectrum* $B_\nu(T)$. This involves two steps: first we compute the density of states as a function of frequency, and then we compute the mean energy per state using the Boltzmann formula. After discussing what this means in terms of occupation numbers, we combine these to obtain the energy density $u_\nu(\Omega)$ and $B_\nu(T)$. Finally, we discuss some general properties of the Planck spectrum and the various ‘characteristic temperatures’ that are observationally useful.

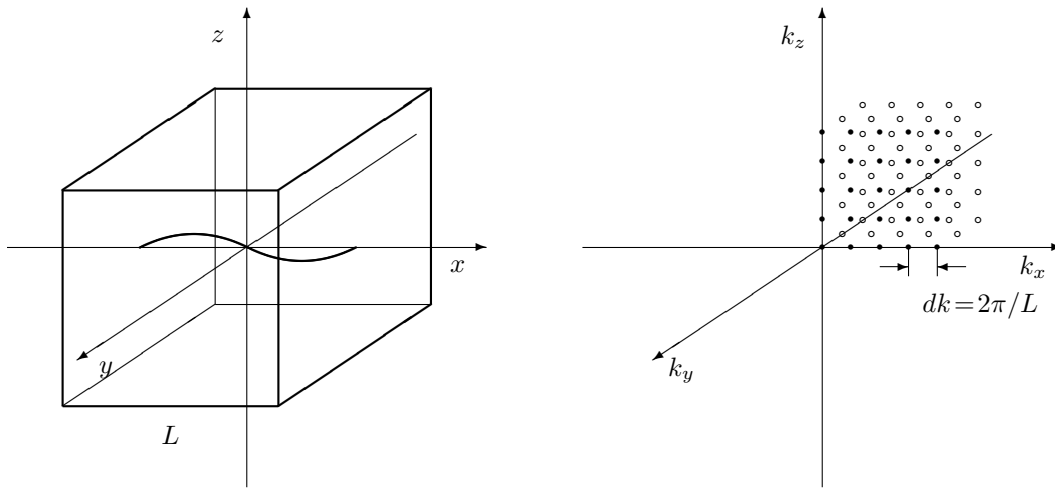


Figure 5.2: Left panel shows an idealised cubical box of side L . Also indicated is a standing wave (actually the fundamental mode). On the right is the corresponding lattice of states in wave-number space. As the box becomes very large, the spacing of the states becomes very fine. The number of states in some volume Δk^3 of k -space is $2 \times L^3 \Delta k^3 / (2\pi)^3$.

5.2.1 Density of States

Consider a large cubical box of side L with perfectly reflecting walls, as illustrated in figure 5.2. The states of the electromagnetic field consistent with these boundary conditions are a set of standing waves, labelled by a set of three numbers (n_x, n_y, n_z) giving the number of periods in the x , y , and z dimensions respectively. Equivalently, since L is considered to be constant here, the states can be labelled by the wave-number vector $\mathbf{k} = (2\pi/L)\mathbf{n}$ which has amplitude $|\mathbf{k}| = 2\pi/\lambda = 2\pi\nu/c$. The momentum of a photon with this wavelength is $p = \hbar k/2\pi = \hbar k$.

These standing wave states form a regular cubical lattice in \mathbf{k} -space with spacing $dk = 2\pi/L$. As $L \rightarrow \infty$, the spacing dk shrinks to zero.

Consider a finite volume in wave-number space $(\Delta k)^3$. The number of states in this volume is

$$N_{\text{states}} = 2 \times (\Delta k)^3 / dk^3 = 2L^3 (\Delta k)^3 / (2\pi)^3 \quad (5.17)$$

where the factor 2 arises because there are two polarization states for each allowed oscillation mode.

The number of states N_{states} is proportional to the product of the real space volume element $V = L^3$ and the k -space volume element $(\Delta k)^3$. There is therefore a well defined and constant

density of states in 6-dimensional space:

$$\frac{N_{\text{states}}}{\text{volume} \times k - \text{volume}} = \frac{2}{(2\pi)^3}. \quad (5.18)$$

Consider the modes with wave normal $\hat{\mathbf{k}}$ in some element $d\Omega$ of solid angle and with $|\mathbf{k}|$ in the range k to $k + \Delta k$. The k -space volume is $k^2 dk d\Omega = (2\pi)^3 \nu^2 d\nu d\Omega / c^3$, so the number of states in this volume element is $N_{\text{states}} = 2L^3 \nu^2 d\nu d\Omega / c^3$ and therefore the density of states (in volume-frequency space) is

$$\rho \equiv N_{\text{states}} / \text{volume} / \text{solid angle} / \text{frequency} = 2\nu^2 / c^3 \quad (5.19)$$

5.2.2 Mean Energy per State

The next step is to compute the mean energy for an oscillator of a given spatial frequency \mathbf{k} (and polarization). According to elementary quantum mechanics, a simple harmonic oscillator has quantized energy levels

$$E = (n + 1/2)h\nu. \quad (5.20)$$

Henceforth we shall ignore the constant $h\nu/2$ ground state energy.

These oscillator states can be thought of as distinguishable particles, and we can therefore apply the Boltzmann formula to say that the probability to find an oscillator with energy E is proportional to $\exp(-\beta E) = \exp(-E/kT)$. The mean energy is

$$\bar{E} = \frac{\sum_{n=0}^{\infty} E e^{-\beta E}}{\sum_{n=0}^{\infty} e^{-\beta E}} = -\frac{\partial}{\partial \beta} \ln \sum_{n=0}^{\infty} e^{-\beta E}. \quad (5.21)$$

but the sum here is a simple geometric series

$$\sum_{n=0}^{\infty} e^{-\beta E} = \sum_{n=0}^{\infty} e^{-n h \nu \beta} = (1 - e^{-\beta h \nu})^{-1} \quad (5.22)$$

from which we obtain

$$\bar{E} = \frac{h\nu e^{-\beta h \nu}}{1 - e^{-\beta h \nu}} = \frac{h\nu}{e^{\beta h \nu} - 1}. \quad (5.23)$$

5.2.3 Occupation Number in the Planck Spectrum

If we divide the mean energy \bar{E} by the energy per photon $h\nu$ we obtain the *mean occupation number*

$$\bar{n} = \frac{1}{e^{\beta h \nu} - 1} \quad (5.24)$$

(see figure 5.3). This has the following asymptotic dependence on frequency:

- $h\nu \ll kT$: expanding the exponential as a Taylor series the leading order term is $n \simeq kT/h\nu \gg 1$. In this ‘Raleigh-Jeans’ region the occupation number is large — each state contains a large number of photons, and photon discreteness effects are negligible. The occupation number diverges as $\nu \rightarrow 0$, but the mean energy remains finite $\bar{E} \simeq kT$.
- $h\nu \simeq kT$: this is the characteristic photon energy, and the occupation number is of order unity.
- $h\nu \gg kT$: the +1 in the denominator is negligible and $\bar{n} \simeq \exp(-h\nu/kT)$ and becomes exponentially small.
- The occupation number \bar{n} derived here is the equilibrium distribution for massless bosonic particles.

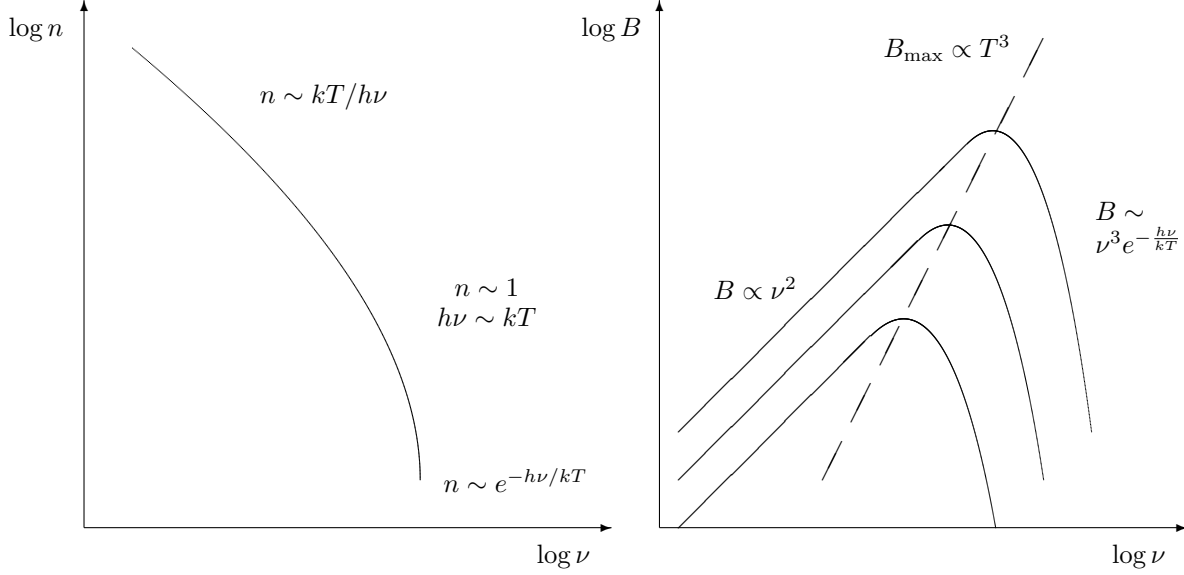


Figure 5.3: Left panel shows the occupations number for the Planck function. Right panel shows the brightness as a function of frequency for various temperature black-bodies.

5.2.4 Specific Energy Density and Brightness

Multiplying the number of states $dN = (2\nu^2/c^3)dV d\nu d\Omega$ by the mean energy for a state (5.23) yields

$$dE = u_\nu(\Omega)dV d\nu d\Omega = (2\nu^2/c^3) \frac{h\nu}{e^{\beta h\nu} - 1} dV d\nu d\Omega \quad (5.25)$$

so the specific energy density is

$$u_\nu(\Omega) = \frac{2h\nu^3/c^3}{e^{\beta h\nu} - 1} \quad (5.26)$$

and the brightness $B_\nu = I_\nu = cu_\nu$ is

$$B_\nu = \frac{2h\nu^3/c^2}{e^{\beta h\nu} - 1}. \quad (5.27)$$

The Planck function grows as $B \propto \nu^2$ for $h\nu \ll kT$ and then falls rather abruptly due to the exponential for $h\nu \gtrsim kT$ (see figure 5.3).

This is consistent with the $u \propto T^4$ behavior deduced from thermodynamic considerations. The total energy density is $u = 4\pi \int d\nu B_\nu/c$, this integral is dominated by modes around the characteristic frequency $\nu^* = kT/h$, and the value of the integrand at the peak is on the order of $B_{\max} \sim h\nu^3/c^2 \sim (kT)^3/(h^2c^2)$ so the integral is on the order of $u \sim \nu^* B_{\max}/c \sim (kT)^4/(h^3c^3)$, which is proportional to T^4 .

Note that the energy density can also be written as $u \sim kT/\lambda^3$. Equivalently, the number density of photons with energy $h\nu \sim kT$ is on the order of $n \sim 1/\lambda^3$.

5.3 Properties of the Planck Spectrum

5.3.1 Raleigh-Jeans Law

As discussed, for $h\nu \ll kT$ the argument of the exponential is small, and expanding gives mean energy per state $\bar{E} \simeq kT$, so we can say that these modes are ‘in equipartition’. The intensity is

$$B_\nu^{\text{RJ}} \simeq 2\nu^2 kT/c^2 \quad (5.28)$$

5.3.2 Wien Law

In the other extreme $h\nu \gg kT$ we have

$$B_\nu^{\text{Wien}} \simeq \frac{2h\nu^3}{c^2} e^{-h\nu/kT} \quad (5.29)$$

which falls sharply with increasing frequency.

5.3.3 Monotonicity with Temperature

From (5.27) we find $\partial B_\nu(T)/\partial T > 0$ for all ν . Thus $B_\nu(T' > T)$ lies everywhere above $B_\nu(T)$.

5.3.4 Wien Displacement Law

The brightness B_ν peaks at $h\nu_{\text{max}} = 2.82kT$ so $\nu_{\text{max}} = (5.88 \times 10^{10} \text{Hz})kT/h$. In wavelength, B_λ peaks at $\lambda_{\text{max}}T = 0.29 \text{cm K}$.

5.3.5 Radiation Constants

The constants occurring in the various versions of the Stefan-Boltzmann law can be computed in terms of the ‘fundamental’ constants h , k , c . The total brightness is

$$B = \int d\nu B_\nu(T) = \left(\frac{2h}{c^3}\right) \left(\frac{kT}{h}\right)^4 \int \frac{x^3 dx}{e^x - 1}. \quad (5.30)$$

The value of the dimensionless integral here is $\pi^4/15$ so

$$B = \frac{2\pi^4 k^4}{15c^2 h^3} T^4 \quad (5.31)$$

and hence

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3} \quad \text{and} \quad a = \frac{8\pi^5 k^4}{15c^3 h^3}. \quad (5.32)$$

5.4 Characteristic Temperatures

5.4.1 Brightness Temperature

If we observe a brightness I_ν at some frequency ν then we can define a ‘brightness temperature’ T_b such that $I_\nu = B_\nu(T_b)$.

This is simply computed in the RJ regime:

$$T_b = \frac{c^2}{2\nu^2 k} I_\nu \quad (5.33)$$

This requires that the source be resolved in order for I_ν to be defined, and will give a reliable estimate of the temperature if the source is optically thick.

5.4.2 Color Temperature

If we can deduce the characteristic frequency ν_* , say from broad band observations at a range of frequencies, then we can define the ‘color temperature’

$$T_{\text{color}} = h\nu_*/k. \quad (5.34)$$

The color temperature does not require the source to be resolved or optically thick.

5.4.3 Effective Temperature

If a bolometer provides the total flux density F but does not provide any detailed frequency distribution information one can deduce the temperature if the size of the source $d\Omega$ is known by equating

$$B = \frac{F}{d\Omega} = \frac{ac}{4\pi} T_{\text{eff}}^4 \quad (5.35)$$

This requires that the source be resolved.

5.5 Bose-Einstein Distribution

The Planck spectrum derived above is the fully equilibrated distribution function for photons, and in the derivation we placed no restrictions on the occupation numbers, implicitly assuming that reactions which create or destroy photons are efficient. Under certain circumstances (e.g. electron scattering) such reactions may be inefficient, though scattering may be able to efficiently redistribute energy among a fixed number of photons. In this case, the photon energy distribution will deviate from the Planck spectrum, and one has the more general *Bose-Einstein* distribution, with mean occupation number

$$\bar{n} = \frac{1}{e^{(h\nu+\mu)/kT} - 1} \quad (5.36)$$

The constant μ here is known as the *chemical potential*.

If the chemical potential is positive then this leads to a finite occupation number at zero energy. Such a situation arises if there are too few photons for the given energy (i.e. fewer than for a Planck spectrum of a given energy).

If the chemical potential is negative then the Bose-Einstein occupation number becomes infinite at a finite frequency $h\nu = \mu$, which is unphysical. Now one can vary the chemical potential by supplying or removing heat (without allowing photon number changing reactions). If one were to start with a photon gas with positive μ then one can decrease it by extracting heat from the gas. Once μ reaches zero, any further extraction of heat will result in a *Bose condensation* with a Planckian distribution for $E > 0$ and any excess photons in the zero energy state.

We will see how the Bose-Einstein distribution and its analog for fermions - the Fermi-Dirac distribution arise when we discuss kinetic theory in chapter 19.

5.6 Problems

5.6.1 Thermodynamics of Black-Body Radiation

The entropy S of a system is defined by $dQ = TdS$ where dQ is the heat input and T is the temperature.

- Derive an expression for the entropy of black body radiation in an enclosure in terms of the temperature T , the volume V , and the radiation constant a .
- If one expands the enclosure adiabatically (i.e. with no heat input or output) how does the temperature scale with the volume.

5.6.2 Black body radiation and adiabatic invariance

a) Consider a perfectly reflecting cavity containing black body radiation. By invoking ‘adiabatic invariance’ (i.e. constancy of occupation number for each fundamental mode of oscillation) show that if the cavity (a balloon perhaps) expands isotropically, then the radiation maintains a black body spectrum. How does the temperature scale with the linear size of the cavity?

b) Assume that the 3K background radiation has an energy density smaller than the current matter density by a factor 10^3 . What was the temperature of the background radiation at the epoch when the matter and radiation had equal energy density?

c) Now consider a perfectly reflecting cavity which expands only in one direction (a cylinder capped by a piston for instance) and which initially contains black-body radiation. Again applying adiabatic invariance, compute by what factor the energy in the radiation decreases in the limit of a very large expansion factor. You should *not* assume that the cavity contains any matter to scatter the radiation and maintain isotropy and/or thermal equilibrium.

d) Apply the same principle of constancy of occupation number to another simple harmonic oscillator: a simple pendulum undergoing small oscillations but with a slowly varying pendulum length L . How does the energy of the oscillator depend on frequency? How does the amplitude of the oscillations depend on frequency?

5.6.3 Black-body radiation

Derive the Planck spectrum for a thermal black-body radiation field. You should proceed in two steps:

First compute the density of states in frequency space for a cubicle cavity of side L (don't forget to allow for the two independent polarisation states). Obtain an expression for ρ , the number of states per unit volume per unit frequency per unit solid angle.

Next calculate the mean energy per state as follows. In thermal equilibrium at temperature T , the probability distribution for the energy $E = nh\nu$ follows the Boltzmann law $p(E) \propto \exp(-\beta E)$ where $\beta = 1/kT$ and k is Boltzmann's constant. Thereby show that the mean energy of a mode is given by $\bar{E} = -\partial/\partial\beta \ln(\sum_{n=0}^{\infty} e^{-\beta nh\nu})$. Evaluate the sum to obtain a closed expression for \bar{E} and combine with your expression for the density of states to obtain an expression for the specific energy density $u_{\nu}(\Omega)$.

Integrate $u_{\nu}(\Omega)$ to obtain the Stefan-Boltzmann law for the total energy density: $u = aT^4$ and express a in terms of fundamental constants (you may use the result for the dimensionless integral $\int dx x^3/(e^x - 1) = \pi^4/15$).

5.6.4 Planck Spectrum

A cubical cavity of side 1m contains radiation at temperature $T = 300\text{K}$. Give a rough estimate of the characteristic wavelength of the radiation and estimate of the number of photons in the enclosure.

Chapter 6

Radiative Transfer

We now consider radiation passing through matter, which may absorb, emit and/or scatter radiation out of or into the beam. We will derive the equation governing the evolution of the intensity. Real scattering processes (e.g. electron scattering) are generally anisotropic and introduce polarization. Here we will consider only isotropic scatterers and unpolarized radiation. We will also limit attention to ‘elastic’ scattering in which there is negligible change in the photon energy in scattering (this is a reasonable approximation if the photon energy is much less than the rest mass of the scatterer and the latter has thermal or random velocity much less than the speed of light).

6.1 Emission

We define the (spontaneous) emission coefficient such that matter in a volume element dV adds to the radiation field an amount of energy

$$dE = j_\nu(\Omega) dV d\Omega dt d\nu. \quad (6.1)$$

For isotropically emitting particles

$$j_\nu = \frac{P_\nu}{4\pi} \quad (6.2)$$

where P_ν here is the radiated power per unit volume per unit frequency.

The (angle averaged) *emissivity* ϵ_ν is the energy input per unit mass per unit time per unit frequency and is defined such that

$$dE = \epsilon_\nu \rho dV dt d\nu (d\Omega/4\pi) \quad (6.3)$$

with ρ the mass density, from which follows the relation

$$j_\nu = \frac{\epsilon_\nu \rho}{4\pi}. \quad (6.4)$$

Consider a cylindrical tube of cross-section area dA and length dl . The radiation energy in a co-axial cone of direction $d\Omega$ is

$$E = u_\nu dA dl d\Omega d\nu \quad (6.5)$$

in a time $dt = dl/c$ this tube will have moved through its length, and the matter in the volume will have injected an extra amount of energy

$$\delta E = j_\nu dA dl d\Omega d\nu dt \quad (6.6)$$

so $u'_\nu \equiv (E + \delta E)/dA dl d\Omega d\nu = u_\nu + j_\nu dt$, or $du_\nu = j_\nu dt$ or, since $I_\nu = cu_\nu$

$$dI_\nu = j_\nu ds \quad (6.7)$$

where ds is an infinitesimal path length element.

6.2 Absorption

Absorption will remove from a beam an amount of intensity proportional to the incident intensity and proportional to the path length:

$$dI_\nu = -\alpha_\nu I_\nu ds \quad (6.8)$$

where the *absorption coefficient* α has units of $(\text{length})^{-1}$.

For a simple model of randomly placed absorbing spheres with cross-section σ and number density n the mean covering fraction for objects in a tube of area A and length ds is $\delta A/A = n\sigma dV/A = n\sigma dl$ so we would expect an attenuation in intensity

$$dI = -n\sigma I ds \quad (6.9)$$

so, for this model, $\alpha = n\sigma$.

One can also define the *opacity* κ_ν

$$\alpha_\nu = \rho\kappa_\nu \quad (6.10)$$

which is more convenient if one wishes to calculate the attenuation for propagation through a given column density of material.

6.3 The Equation of Radiative Transfer

Combining emission and absorption terms give the *equation of radiative transfer*:

$$\frac{dI_\nu}{ds} = -\alpha_\nu I_\nu + j_\nu \quad (6.11)$$

This is easy to solve if j_ν is given, but tricky in general since j_ν contains scattered radiation which is proportion to the angle averaged intensity, so we obtain an integro-differential equation.

Solutions can easily be found for two special cases:

- Emission only:

$$\frac{dI_\nu}{ds} = j_\nu \longrightarrow I_\nu(s) = I_\nu(0) + \int ds j_\nu \quad (6.12)$$

- Absorption only:

$$\frac{dI_\nu}{ds} = -\alpha_\nu I_\nu \longrightarrow I_\nu(s) = I_\nu(0)e^{-\int ds \alpha_\nu} \quad (6.13)$$

It is useful to define an alternative dimensionless parameterization of path length called the *optical depth* τ such that

$$d\tau_\nu \equiv \alpha_\nu ds \quad (6.14)$$

and in terms of which the pure absorption solution is $I_\nu(s) = I_\nu(0)e^{-\tau}$.

With both emission and absorption it is useful to define the *source function*

$$S_\nu = j_\nu/\alpha_\nu \quad (6.15)$$

in terms of which the RTE becomes

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu + S_\nu \quad (6.16)$$

and the formal solution of the RTE is

$$I_\nu(\tau) = I_\nu(0)e^{-\tau} + \int_0^\tau e^{-(\tau-\tau')} S_\nu(\tau') d\tau' \quad (6.17)$$

which can be verified by direct differentiation. The utility of this solution is limited by the fact that the source function is not given, but must usually be determined as an integral over the intensity.

6.4 Kirchhoff's Law

Consider a kiln containing some matter and radiation in thermal equilibrium at temperature T . The intensity obeys the RTE:

$$\frac{dI_\nu}{d\tau} = S_\nu - I_\nu \quad (6.18)$$

but in equilibrium the intensity must be spatially constant, so $dI/d\tau = 0$, and hence the $S_\nu = I_\nu = B_\nu(T)$, ie the source function and intensity are equal. Since the source function is defined as $S \equiv j/\alpha$, this means that

$$j_\nu = \alpha_\nu B_\nu(T) \quad (6.19)$$

which is *Kirchoff's law*. This tells us, for instance, that if some material absorbs well at some particular wavelength (perhaps because of a resonance) then it will also radiate well at the same frequency.

6.5 Mean Free Path

We can compute the probability distribution for photon paths and the mean free path as follows. The probability that a photon is absorbed within an infinitesimal distance ds is

$$P_{\text{abs}} = \alpha ds \quad (6.20)$$

so the probability it survives is

$$P_{\text{survive}}(ds) = 1 - \alpha ds \quad (6.21)$$

and the probability it survives at least a finite distance s (i.e. it survives a succession of $N = s/ds$ steps of length ds) is

$$P_{\text{survive}}(> s) = (1 - \alpha ds)^N = (1 - \alpha s/N)^N = \exp(-\alpha s) \quad (6.22)$$

This is the cumulative distribution function. The differential distribution function for path lengths is

$$P(s) = -\frac{dP_{\text{survive}}(> s)}{ds} = \alpha \exp(-\alpha s) \quad (6.23)$$

so the distribution of path lengths is exponential with typical path length $\sim 1/\alpha$ as might have been expected.

The mean path length is

$$\langle s \rangle = \int ds s P(s) = 1/\alpha. \quad (6.24)$$

6.6 Radiation Force

Radiation incident on scattering/absorbing material will result in transfer of momentum to the matter. We now calculate this assuming that the scattered or re-radiated energy is emitted isotropically in the rest frame of the matter. Consider the radiation in a narrow cone of direction $d\Omega$ and in a co-axial cylinder. The energy in the cylinder is

$$E = u dl dA d\Omega. \quad (6.25)$$

In a time $dt = dl/c$ this tube will travel its length, and absorbers and/or scatterers will have reduced the energy content by an amount

$$\delta E = (\alpha dl) E \quad (6.26)$$

with a corresponding transfer of momentum

$$\delta \mathbf{p} = \hat{\mathbf{n}} \alpha dl E / c = \hat{\mathbf{n}} \alpha u dt dV d\Omega = \hat{\mathbf{n}} \alpha \frac{I}{c} dt dV d\Omega \quad (6.27)$$

so

$$\frac{\text{momentum}}{\text{volume} \times \text{time}} = \frac{\text{force}}{\text{volume}} = \frac{1}{c} \int d\nu \alpha_\nu \mathbf{F}_\nu \quad (6.28)$$

where $\mathbf{F}_\nu \equiv \int d\Omega \hat{\mathbf{n}} I_\nu$ is the *radiation flux vector*.

The acceleration of the material is

$$\frac{\text{force}}{\text{mass}} = \text{acceleration} = \frac{1}{c} \int d\nu \kappa_\nu \mathbf{F}_\nu \quad (6.29)$$

Note that these results assume that the scattering process is front-back symmetric.

6.7 Random Walks

As discussed above, the RTE with scattering is an integro-differential equation and it is therefore quite hard to find exact solutions for most problems of interest.

However, order-of-magnitude estimates of some important features can be obtained using random walk arguments. The basic idea here is that if the typical distance between scatterings is l say (ie the mean free path) and if each scattering gives a random change in direction then, after N scatterings, a photon will have traversed a net distance on the order of $L_\star = \sqrt{N}l$.

A simple way to derive this famous ‘drunkards walk’ law is to consider a simple 1-dimensional walk, where at each step the particle can move forward and backward one unit. Let a particle have reached position x_n after n steps. After the next step it will be at $x_{n+1} = x_n + 1$ or $x_{n+1} = x_n - 1$ with equal probability of $1/2$. The mean square displacement after $n+1$ steps *given* that the particle is at x_n after n steps is

$$\langle x_{n+1}^2 | x_n \rangle = \frac{1}{2} [(x_n + 1)^2 + (x_n - 1)^2] = x_n^2 + 1 \quad (6.30)$$

so the average increase of $\langle x^2 \rangle$ in one step is unity, regardless of the value of x_n . More formally, the unconstrained mean square of x_{n+1} is given by integrating over all possible values for x_n :

$$\langle x_{n+1}^2 \rangle = \int dx_n p(x_n) \langle x_{n+1}^2 | x_n \rangle = \int dx_n p(x_n) (x_n^2 + 1) = \langle x_n^2 \rangle + 1 \quad (6.31)$$

so the mean square displacement increases by unity for each step and since the particle starts at $x_0 = 0$ we have $\langle x_n^2 \rangle = n$.

Alternatively, and in 3-dimensions, one can write the net vector displacement \mathbf{R} as

$$\mathbf{R} = \mathbf{r}_1 + \mathbf{r}_2 + \dots + \mathbf{r}_N \quad (6.32)$$

The mean vector displacement vanishes,

$$\langle \mathbf{R} \rangle = N \langle \mathbf{r}_i \rangle = 0 \quad (6.33)$$

but the mean square displacement is

$$\langle R^2 \rangle = \langle r_1^2 \rangle + 2 \langle \mathbf{r}_1 \cdot \mathbf{r}_2 \rangle + \dots + \langle r_2^2 \rangle + \dots \quad (6.34)$$

now all the cross terms like $\langle \mathbf{r}_1 \cdot \mathbf{r}_2 \rangle$ vanish (since the directions of different path segments are assumed to be uncorrelated), and we have

$$\langle R^2 \rangle = N \langle r^2 \rangle. \quad (6.35)$$

As an example, consider the escape of a photon from a cloud of size L . If the mean free path is $l \ll L$ then the optical depth of the cloud is $\tau \sim L/l$. In N steps, the photon will travel a distance $l_\star \sim \sqrt{N}l$. Equating this with the size of the cloud L yields the required number of steps for escape $N \sim \tau^2$.

6.8 Combined Scattering and Absorption

Of some interest are media in which there is both absorption, characterized by the absorption coefficient α_ν , and scattering, which we shall characterize by the scattering coefficient σ_ν , such that the mean free path for scattering alone is $1/\sigma_\nu$. The distinction is that scattering can isotropize radiation, but cannot thermalize it, as this requires true absorption and re-emission.

The mean free path is

$$l_\nu = (\alpha_\nu + \sigma_\nu)^{-1}. \quad (6.36)$$

The probability that a path ends with an absorption event is called the *single scattering albedo*

$$\epsilon_\nu = \frac{\alpha_\nu}{\alpha_\nu + \sigma_\nu}. \quad (6.37)$$

The probability that a photon gets absorbed after N paths is then $P(N) \sim \epsilon N$ which approaches unity after $N \sim \epsilon_\nu^{-1}$ paths, or equivalently after traveling a net distance

$$l_\star \sim l/\sqrt{\epsilon} \sim \frac{1}{\sqrt{\alpha_\nu(\alpha_\nu + \sigma_\nu)}}. \quad (6.38)$$

This is called the *thermalization length* or the *effective mean free path*. One can define the *effective optical depth* for a cloud of size R as

$$\tau_\star = \frac{R}{l_\star} = \sqrt{\tau_\alpha(\tau_\alpha + \tau_\sigma)} \quad (6.39)$$

with $\tau_\alpha \equiv \alpha R$ and $\tau_\sigma \equiv \sigma R$.

- If $\tau_\star \ll 1$ then the cloud is ‘effectively thin’ and thermal photons emitted by the absorbing particles will typically escape from the cloud without being re-absorbed. The emissivity (due exclusively to the absorbing particles) is $j = \alpha B$ and the luminosity is $L \sim jV \sim \alpha BV$. The brightness is $I \sim L/A \sim \alpha BL \sim \epsilon \tau_\star B$.
- If $\tau \gg 1$ the cloud is ‘effectively thick’ and one expects the radiation within the cloud to be thermalized. In this case the photons which escape will typically have been emitted by absorbers within a distance $\sim l_\star$ of the surface, so the luminosity of the cloud will be $L \sim A j l_\star \sim \alpha B A l_\star \sim \sqrt{\epsilon} B A$. This luminosity will be much less than that of a black-body radiator of the same area if $\epsilon \ll 1$.

6.9 Rosseland Approximation

The *Rosseland approximation* allows one to compute the transport of energy in a cloud with scattering and absorption.

We can write the RTE as

$$\frac{dI_\nu}{ds} = -\alpha_\nu(I_\nu - B_\nu) - \sigma_\nu(I_\nu - J_\nu) \quad (6.40)$$

since the absorbing particles emit with emissivity $j_\nu = \alpha_\nu B_\nu$ whereas the radiation scattered is proportional to the mean intensity J_ν .

This can also be written as

$$\frac{dI_\nu}{ds} = -(\alpha_\nu + \sigma_\nu)(I_\nu - S_\nu) \quad (6.41)$$

where the source function is here defined as

$$S_\nu \equiv \frac{\alpha_\nu B_\nu + \sigma_\nu J_\nu}{\alpha_\nu + \sigma_\nu} \quad (6.42)$$

which, in the isotropic scattering model, is independent of direction.

If we assume a plane-parallel or stratified system $dz = \mu ds$ with $\mu = \cos \theta$ and we have

$$\mu \frac{\partial I_\nu}{\partial z} = -(\alpha_\nu + \sigma_\nu)(I_\nu - S_\nu) \quad (6.43)$$

The Rosseland approximation scheme exploits the fact that deep within a cloud, the radiation will be close to isotropic, and the fractional change in the intensity over one mean free path will be small, or equivalently, that $\partial I / \partial z \ll (\alpha + \sigma)I$.

To zeroth order we ignore $\partial I_\nu / \partial z$ entirely and readily obtain the zeroth order solution

$$I_\nu^{(0)} = B_\nu. \quad (6.44)$$

At next higher order

$$I_\nu^{(1)} = S_\nu - \frac{\mu}{\alpha_\nu + \sigma_\nu} \frac{\partial B_\nu}{\partial z} \quad (6.45)$$

Taking the first moment of this, the source function (which is isotropic) drops out, and we obtain the first order flux

$$F_\nu(z) = 2\pi \int d\mu \mu I_\nu^{(1)}(z, \mu) = -\frac{4\pi}{3(\alpha_\nu + \sigma_\nu)} \frac{\partial B_\nu(T)}{\partial T} \frac{dT}{dz} \quad (6.46)$$

The integrated flux is

$$F(z) = -\frac{4\pi}{3} \frac{dT}{dz} \int d\nu (\alpha_\nu + \sigma_\nu)^{-1} \frac{\partial B_\nu(T)}{\partial T} \quad (6.47)$$

or

$$F(z) = -\frac{16\sigma_{\text{SB}}T^3}{3\alpha_R} \frac{dT}{dz} \quad (6.48)$$

where the Rosseland absorption coefficient α_R is defined as

$$\frac{1}{\alpha_R} \equiv \int d\nu (\alpha_\nu + \sigma_\nu)^{-1} \frac{\partial B_\nu(T)}{\partial T} / \int d\nu \frac{\partial B_\nu(T)}{\partial T} \quad (6.49)$$

and σ_{SB} is the Stefan-Boltzmann constant, and where we have made use of $\int d\nu \partial B_\nu / \partial T = \partial B(T) / \partial T = 4\sigma T^3 / \pi$.

- Equation (6.48) shows the heat flux to be proportional to the temperature gradient, as might have been expected, or more generally $\mathbf{F} \propto \nabla T$.
- The thermal conductivity is $16\sigma_{\text{SB}}T^3/3\alpha_R$
- α_R^{-1} is largest at frequencies where the combined absorption + scattering coefficient $\alpha_\nu + \sigma_\nu$ is small.
- The Rosseland approximation requires both nearly isotropic conditions and the stronger condition that the medium should be nearly thermalized.
- It tells us how energy seeps upwards through a scattering and absorbing medium, but it does not describe the manner in which the intensity $I_\nu \rightarrow B_\nu$ within such a medium.

One can understand the general form of (6.48) from a simple order-of-magnitude argument. Consider two adjacent slabs of material of thickness $\Delta z \sim l$, the mean-free-path, and with the upper and lower slabs having temperatures T and $T + \Delta T$ respectively. The energy within the upper slab is $E = aA\Delta zT^4$, while that in the lower is $E' = aA\Delta z(T + \Delta T)^4 \simeq E + 4aA\Delta zT^3\Delta T$. In a time $dt \sim dl/c \sim \Delta z/c$, a substantial fraction of the photons originally in the lower slab will now be in the upper slab and *vice versa*, with a corresponding transfer of energy per unit area per time of

$$F \sim \frac{E' - E}{Adt} \sim alT^3c \frac{\Delta T}{\Delta z} \sim l\sigma_{\text{SB}}T^3 \frac{\Delta T}{\Delta z} \sim \frac{\sigma_{\text{SB}}T^3}{\alpha} \frac{\partial T}{\partial z} \quad (6.50)$$

in agreement with (6.48).

6.10 The Eddington Approximation

The *Eddington approximation* also assumes that conditions are close to isotropic, but not necessarily that the radiation is locally nearly thermalized.

Eddington expands the intensity as a function of angle $\mu = \cos \theta$:

$$I_\nu(\tau, \mu) = a_\nu(\tau) + b_\nu(\tau)\mu + \dots \quad (6.51)$$

and ignores all but the first two terms.

Taking the first three moments of the intensity gives

$$\begin{aligned} J_\nu &= \frac{1}{2} \int d\mu I_\nu = a_\nu \\ H_\nu &= \frac{1}{2} \int d\mu \mu I_\nu = b_\nu/3 \\ K_\nu &= \frac{1}{2} \int d\mu \mu^2 I_\nu = a_\nu/3 \end{aligned} \quad (6.52)$$

These are proportional to the mean intensity, the energy flux, and the radiation pressure respectively. The first and last together imply the relation $K_\nu = J_\nu/3$.

Write the RTE as

$$\mu \frac{\partial I_\nu}{\partial \tau_\nu} = -(I_\nu - S_\nu) \quad (6.53)$$

with

$$d\tau_\nu = -(\alpha_\nu + \sigma_\nu)dz \quad (6.54)$$

and

$$S_\nu \equiv \frac{\alpha_\nu B_\nu + \sigma_\nu J_\nu}{\alpha_\nu + \sigma_\nu} = \epsilon_\nu B_\nu + (1 - \epsilon_\nu)J_\nu \quad (6.55)$$

and take the first few moments.

The zeroth moment yields

$$\frac{\partial H_\nu}{\partial \tau} = J_\nu - S_\nu \quad (6.56)$$

and the first moment is

$$\frac{\partial K_\nu}{\partial \tau} = \frac{1}{3} \frac{\partial J_\nu}{\partial \tau} = H_\nu \quad (6.57)$$

and combining these and eliminating H_ν yields

$$\frac{1}{3} \frac{\partial^2 J_\nu}{\partial \tau^2} = J_\nu - S_\nu = J_\nu - \epsilon_\nu B_\nu + (1 - \epsilon_\nu)J_\nu \quad (6.58)$$

or, more simply,

$$\frac{1}{3} \frac{\partial^2 J_\nu}{\partial \tau^2} = \epsilon_\nu (J_\nu - B_\nu). \quad (6.59)$$

Given some temperature profile $T(z)$, and hence $B_\nu(z)$, equation (6.59) allows one to solve for J_ν and hence the source function S_ν and from this one can obtain I_ν from the formal solution of the RTE.

Equation (6.59) has the form of a diffusion equation and describes how the intensity relaxes towards the black-body from within a cloud. For example, if we assume the temperature, and therefore B , to be nearly independent of position then the general solution of (6.59) contains exponentially growing and decaying terms $J - B \sim \exp(\pm \tau_\star)$ with $\tau_\star \equiv \sqrt{3\epsilon}\tau$ the effective optical depth. If we require that $J \rightarrow B$ deep within the cloud, this boundary condition kills the exponentially growing term, and we find that the radiation relaxes towards black-body exponentially with a scale length equal to the effective mean free path, in agreement with the order-of-magnitude result above.

See R+L for further discussion of this and the ‘two-stream’ approximation.

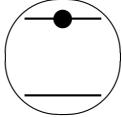
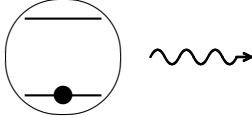
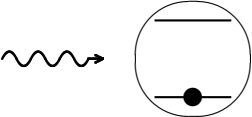
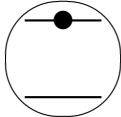
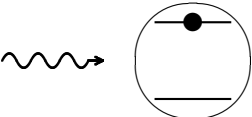
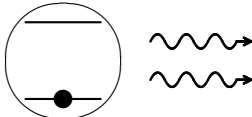
before	rate	after
	A_{21}	
	JB_{12}	
	JB_{21}	

Figure 6.1: Processes considered in the Einstein ‘two-state atom’ model. The filled circle indicates the internal energy level of the atom. Lower/upper level is the ground/excited state. Top row is spontaneous emission. Middle panel is absorption. Bottom panel is stimulated emission, where the atom is ‘encouraged’ to de-excite as a result of the ambient radiation.

6.11 Einstein A , B Coefficients

Einstein considered a idealized two-state system with energy levels E_1 , E_2 in equilibrium with a thermal radiation bath with which it can exchange photons of energy $h\nu = E_2 - E_1$. He deduced two important facts:

- In addition to the processes of absorption to excite the system and spontaneous emission to de-excite it there must be another process ‘stimulated emission’ in which the system de-excites stimulated by the ambient radiation.
- The ratio of the rates for these processes are universal and given in terms of fundamental constants.

6.11.1 Einstein Relations

The reactions we consider are tabulated in figure 6.1. These are

- Let the rate for spontaneous emission be A_{21} , which is the probability that the system drops from E_2 to E_1 per unit time.
- Let $B_{12}J_\nu$ be the probability per unit time that the system be excited from E_1 to E_2 , in the ambient mean intensity J_ν .
- Let $B_{21}J_\nu$ be the probability per unit time that the system be de-excited from E_2 to E_1 , as a result of the ambient mean intensity J_ν . This is referred to as *stimulated emission*.

For an ensemble of such systems, there will in equilibrium be number densities n_1 , n_2 which must satisfy

$$n_1 B_{12} J = n_2 A_{21} + n_2 B_{21} J \quad (6.60)$$

from which we can solve for J

$$J = \frac{A_{21}/B_{21}}{(n_1/n_2)(B_{12}/B_{21}) - 1} \quad (6.61)$$

However, in thermal equilibrium, the ratio of occupation numbers must satisfy the Boltzmann law

$$\frac{n_1}{n_2} = e^{\Delta E/kT} = e^{h\nu/kT} \quad (6.62)$$

and therefore we have

$$J = \frac{A_{21}/B_{21}}{(B_{12}/B_{21})e^{h\nu/kT} - 1} \quad (6.63)$$

Comparing this with the Planck function we find that these are consistent, provided the microscopic rates satisfy the ‘Einstein relations’

$$B_{21} = B_{12} \quad (6.64)$$

and

$$A_{21} = 2h\nu^3 c^2 B_{21}. \quad (6.65)$$

- These relations are somewhat reminiscent of Kirchoff’s law, which relates emission and absorption for matter in thermal equilibrium. This is superficial. The Einstein relations (also known as ‘detailed balance’ relations) are more profound as they relate the microscopic absorption and emission rates for any system, regardless of whether it happens to be in thermal equilibrium.
- The discussion here is somewhat oversimplified. See RL for inclusion of ‘statistical weights’.

6.11.2 Emission and Absorption Coefficients

We can derive the absorption and emission coefficients in terms of the A, B coefficients.

The spontaneous emission is

$$dE = n_2 A_{21} h\nu \frac{d\Omega}{4\pi} dV \phi(\nu) d\nu dt = j_\nu d\Omega d\nu dV dt \quad (6.66)$$

where we have introduced a narrow, but finite, ‘line profile’ $\phi(\nu)$. Thus

$$j_\nu = \frac{h\nu}{4\pi} n_2 A_{21} \phi(\nu). \quad (6.67)$$

The (uncorrected) absorption is

$$\alpha_\nu = \frac{h\nu}{4\pi} n_1 B_{12} \phi(\nu) \quad (6.68)$$

and allowing for stimulated emission (which is most naturally considered negative absorption)

$$\alpha_\nu = \frac{h\nu}{4\pi} B_{12} \phi(\nu) (n_1 - n_2) \quad (6.69)$$

The source function is then

$$S_\nu \equiv j_\nu / \alpha_\nu = \frac{n_2 A_{21} / B_{12}}{n_1 - n_2}. \quad (6.70)$$

- If the matter is in thermal equilibrium, this results gives the absorption coefficient as

$$\alpha_\nu = \frac{h\nu}{4\pi} n_1 B_{12} (1 - e^{-h\nu/kT}) \phi(\nu) \quad (6.71)$$

and source function $S_\nu = B_\nu$, in accord with Kirchoff’s law.

- For general (ie non-thermal) occupation numbers, this formula gives emission for given n_1, n_2 .
- If the occupation numbers are *inverted*, so $n_2 > n_1$ then the absorption becomes negative. This leads to an instability with the matter dumping it’s energy into the radiation field giving rise to huge photon occupation numbers. This is the process underlying lasers and masers.

6.12 Problems

6.12.1 Main sequence

Consider the optically thick interior of a star.

a) Using ‘random walk’ arguments (or otherwise) show that the radiative flux (energy/area/time) is on the order of

$$F \sim acT^4\lambda/R \quad (6.72)$$

where λ is the mean free path (MFP), R is the radius and a is the Stefan-Boltzmann constant.

b) How is the MFP related to the density of particles n and their scattering cross section σ ?

c) Now use the equation of hydrostatic equilibrium to show that

$$\frac{kT}{m} \sim \frac{GM}{R} \quad (6.73)$$

where M is the mass interior to R and m is the mean molecular weight (assume that the gas kinetic pressure \gg radiation pressure).

d) Combine the above results to show that the luminosity net luminosity scales as the cube of the mass, and more specifically,

$$L \sim \frac{ac}{\sigma} \left(\frac{Gm}{k} \right)^4 mM^3. \quad (6.74)$$

e) Assuming electron scattering dominates, so the appropriate cross-section is the Thomson cross section $\sigma \sim 6 \times 10^{-25} \text{cm}^2$, estimate the luminosity of a star of mass $2 \times 10^{33} \text{gm}$. How does this compare to the luminosity of the sun?

6.12.2 Radiative Transfer.

A large homogeneous sphere of material of uniform density and temperature T has a scattering coefficient $\sigma = 1 \times 10^{-2} \text{cm}^{-1}$ and an absorption coefficient $\alpha = 1 \times 10^{-4} \text{cm}^{-1}$ (both assumed to be independent of frequency and depth).

a. Use random walk arguments to estimate the depth below the surface at which the radiation approaches a thermal ‘black-body’ spectrum.

b. How does the luminosity of the sphere compare to that of a black body of the same size and temperature.

6.12.3 Eddington luminosity

a) Show that the condition that an optically thin cloud can be ejected by radiation pressure from a nearby luminous object is that the mass-to-light ratio should be

$$\frac{M}{L} < \frac{\kappa}{4\pi Gc} \quad (6.75)$$

where κ is the mass absorption coefficient for the cloud (assumed independent of frequency).

b) Show that the terminal velocity of the cloud, if it starts from rest at a distance R is

$$v = \sqrt{\frac{2GM}{R} \left(\frac{\kappa L}{4\pi GMc} - 1 \right)} \quad (6.76)$$

c) Taking the minimum value of κ to be that due to Thomson scattering when the cloud is fully ionised show that the maximum luminosity the object can have and not eject hydrogen by radiation pressure is

$$L_{\text{EDD}} = \frac{4\pi Gcm_H M}{\sigma_T} = 1.25 \times 10^{38} \text{ergs}^{-1} (M/M_\odot) \quad (6.77)$$

6.12.4 Poissonian statistics

3. Consider a “Poissonian” or “shot noise” bus service with buses arriving at your stop randomly at the rate of n per unit time (i.e. the probability that a bus shall arrive in a short time interval δt is $n\delta t$).

a) Assuming you have just arrived at the bus stop, what is the probability distribution for the time you will have to wait until the next bus arrives?

b) What is the probability distribution for intervals between bus arrivals?

Chapter 7

Radiation Fields

7.1 Lorentz Force Law

In the non-relativistic limit the force on a particle of charge q is

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}/c) \quad (7.1)$$

so the rate of work is $\mathbf{v} \cdot \mathbf{F} = q\mathbf{v} \cdot \mathbf{E}$ (the magnetic field forces particles in a direction perpendicular to their motion so therefore does no work), and therefore $d(mv^2/2)/dt = q\mathbf{v} \cdot \mathbf{E}$.

For a distribution of charges with charge density $\rho \equiv \sum q/\Delta V$ and current density $\mathbf{j} \equiv \sum q\mathbf{v}/\Delta V$ the force per unit volume is

$$\mathbf{f} = \rho\mathbf{E} + \frac{1}{c}\mathbf{j} \times \mathbf{B} \quad (7.2)$$

and the rate of work per unit volume is $\mathbf{j} \cdot \mathbf{E}$ and is equal to the rate of increase of the mechanical energy density.

7.2 Field Energy Density

The electromagnetic field has energy which is quadratic in the fields \mathbf{E} and \mathbf{B} . A simple way to compute the electric field energy is to consider a sphere of radius R carrying total charge Q uniformly distributed in a very thin shell. The electric field ramps up from zero to $E_0 = Q/R^2$ as one passes from the inner to the outer edge of the shell, so the mean electric field felt by the charges is $E_0/2$. If we force the sphere to contract from R to $R - \Delta R$ then we must do work against the repulsive electrostatic field

$$\Delta W = Q(E_0/2)\Delta R = \frac{1}{2}E_0^2 R^2 \Delta R = \frac{1}{8\pi}E_0^2 \Delta V \quad (7.3)$$

with $\Delta V = 4\pi R^2 \Delta R$ the volume swept out. Since in the process we have created field of strength E_0 in this volume, we can identify $E^2/(8\pi)$ with the energy density of the field.

Similar arguments can be applied to show that the magnetic field energy density is $B^2/(8\pi)$.

7.3 Maxwell's Equations

Maxwell's equations are

$$\begin{aligned} M1: \quad \nabla \cdot \mathbf{E} &= 4\pi\rho & M2: \quad \nabla \cdot \mathbf{B} &= 0 \\ M3: \quad \nabla \times \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} & M4: \quad \nabla \times \mathbf{B} &= \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \end{aligned} \quad (7.4)$$

which are respectively the differential statements of M1: Gauss' law; M2: the absence of free magnetic charges; M3: Faraday's or Lenz's law of inductions; and M4: Ampere's law (plus the *displacement current* term) (see figure 7.1).

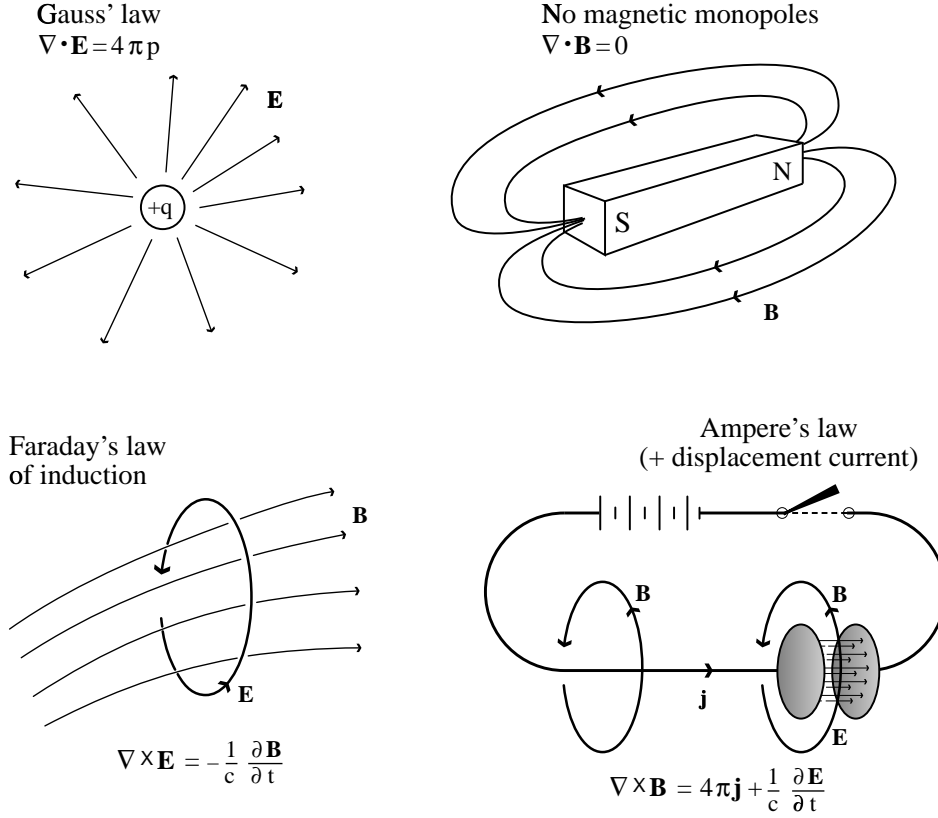


Figure 7.1: Maxwell's equations are, for the most part, the differential statement of the laws of electro- and magneto-statics, plus the law of induction. Maxwell's major contribution was to see that Ampere's law for steady currents $\nabla \times \mathbf{B} = 4\pi\mathbf{j}$ needed to be augmented by the 'displacement current' term. In the lower right panel, for the loop on the left it is clear that we can identify the loop integral of \mathbf{B} with the current piercing the loop. For the right hand loop we can either close the surface so that it has a current through it, or, by taking the surface to fall between the capacitor plates it will have no current, but it will have a net $\partial E/\partial t$. Maxwell's displacement current takes care of the latter possibility, and guarantees consistent results.

Taking the divergence of $M4$ and using $M1$ yields

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0 \quad (7.5)$$

which expresses the conservation of charge (to see this integrate over some volume and invoke the divergence theorem).

Dotting Ampere's law with the electric field \mathbf{E} gives

$$\mathbf{j} \cdot \mathbf{E} = \frac{1}{4\pi} \left[c(\nabla \times \mathbf{B}) \cdot \mathbf{E} - \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \right] \quad (7.6)$$

using the result for vector calculus that $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B})$ we can replace $(\nabla \times \mathbf{B}) \cdot \mathbf{E}$ by $\mathbf{B} \cdot (\nabla \times \mathbf{E}) - \nabla \cdot (\mathbf{E} \times \mathbf{B})$ and finally replacing $\nabla \times \mathbf{E}$ with $-\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t}$ by Lenz's law yields

$$\mathbf{j} \cdot \mathbf{E} + \frac{1}{8\pi} \frac{\partial (E^2 + B^2)}{\partial t} = -\frac{c}{4\pi} \nabla \cdot (\mathbf{E} \times \mathbf{B}) \quad (7.7)$$

Now the first term on the left hand side is the rate of change of the mechanical energy of the charges. The second term is the change of the field energy density, so the left hand side terms together are

the rate of change of the total energy density. This equation expresses conservation of energy, with the right hand side being the divergence of a field energy flux vector, the *Poynting vector*:

$$\mathbf{S} = \frac{c}{4\pi} \mathbf{E} \times \mathbf{B} \quad (7.8)$$

The Poynting flux has some peculiar features. For example, it appears to say that a charged bar magnet has an energy flux that circulates around the bar in a toroidal sense. This is not very meaningful. However, if one integrates the terms in the above equation over some finite volume, the terms on the left give the rate of change of the total energy within the volume, and the right hand side can be converted to an integral over the surface the volume element of the Poynting flux, and this is well defined and free from peculiarities. Note that static fields tend to fall off as $1/r^2$ so $S \sim 1/r^4$ for such fields and the integral over the surface for a static system is $\int d\mathbf{A} \cdot \mathbf{S} \sim 1/r^2$ which converges to zero as it should.

7.4 Electromagnetic Waves

Specializing to empty space, with $\rho = 0$, $\mathbf{j} = 0$, Maxwell's equations become

$$\begin{aligned} M1 : \quad \nabla \cdot \mathbf{E} &= 0 & M2 : \quad \nabla \cdot \mathbf{B} &= 0 \\ M3 : \quad \nabla \times \mathbf{E} &= -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} & M4 : \quad \nabla \times \mathbf{B} &= \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} \end{aligned} \quad (7.9)$$

Taking the curl of $M3$ and using $M4$ gives $\nabla \times (\nabla \times \mathbf{E}) = -c^{-2} \partial^2 \mathbf{E} / \partial t^2$ and invoking the vector identity $\nabla \times (\nabla \times \mathbf{E}) = \nabla \cdot (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$ and Gauss' law $\nabla \cdot \mathbf{E} = 0$ yields

$$\nabla^2 \mathbf{E} - \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \quad (7.10)$$

which has the form of a wave equation. Performing the same sequence of operations with $M1 \leftrightarrow M2$, $M3 \leftrightarrow M4$ yields an identical wave equation for \mathbf{B} .

Let us look for traveling wave solutions of the form

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= \mathbf{e} E_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \\ \mathbf{B}(\mathbf{r}, t) &= \mathbf{b} B_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \end{aligned} \quad (7.11)$$

with unit vectors $|\mathbf{e}| = |\mathbf{b}| = 1$ and with complex amplitudes E_0 , B_0 , though with the understanding that the physical field is the real part. Inserting these trial solutions in Maxwell's equations yields the following inter-relations:

$$\begin{aligned} M1 : \quad \mathbf{k} \cdot \mathbf{e} &= 0 \\ M2 : \quad \mathbf{k} \cdot \mathbf{b} &= 0 \\ M3 : \quad \mathbf{k} \times \mathbf{e} E_0 &= (\omega/c) \mathbf{b} B_0 \\ M4 : \quad \mathbf{k} \times \mathbf{b} B_0 &= -(\omega/c) \mathbf{e} E_0 \end{aligned} \quad (7.12)$$

The first two of these tell us that the waves are transverse: both \mathbf{b} and \mathbf{e} must be perpendicular to the wave vector \mathbf{k} . The second pair tell us that \mathbf{e} and \mathbf{b} must be orthogonal to each other, so \mathbf{k} , \mathbf{e} , and \mathbf{b} form an orthonormal triad. Inspecting the magnitude of the latter pair gives $E_0 = B_0 \omega / kc$ and $B_0 = E_0 \omega / kc$ which imply

$$E_0 = B_0 \quad (7.13)$$

and also yield the *dispersion relation*

$$\omega = ck \quad (7.14)$$

from which we can infer that

$$\text{phase velocity} = \frac{\omega}{k} = \text{group velocity} = \frac{d\omega}{dk} = c \quad (7.15)$$

so these waves are non-dispersive.

The energy density is $U_{\text{field}} = (E^2 + B^2)/8\pi$. Writing $E_0 = |E_0|e^{i\phi}$ we have $\text{Re}(E)^2 = \text{Re}(B)^2 = |E_0|^2 \cos^2(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi)$. The cosine-squared term averages to 1/2 so we have

$$\langle U \rangle = \frac{1}{8\pi} |E_0|^2 = \frac{1}{8\pi} |B_0|^2. \quad (7.16)$$

The Poynting vector is

$$\mathbf{S} = \frac{c}{4\pi} E_0 B_0 \mathbf{e} \times \mathbf{b} = \frac{c}{4\pi} \hat{\mathbf{k}} |E_0|^2 \cos^2(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi) \quad (7.17)$$

which points along the wave vector and has expectation value

$$\langle S \rangle = \frac{c}{8\pi} |E_0|^2 = c \langle U_{\text{field}} \rangle. \quad (7.18)$$

7.5 Radiation Power Spectrum

If we sit in a collimated beam of radiation, the Poynting vector tells us the energy flux per unit time per unit area

$$\frac{dW}{dAdt} = \frac{c}{4\pi} E^2(t) \quad (7.19)$$

so the average energy flux per unit area is

$$\left\langle \frac{dW}{dAdt} \right\rangle = \frac{c}{4\pi} \frac{1}{T} \int_0^T dt E^2(t) \quad (7.20)$$

or equivalently, by Parseval's theorem,

$$\left\langle \frac{dW}{dAdt} \right\rangle = \frac{c}{4\pi} \int \frac{d\omega}{2\pi} \frac{\langle |\tilde{E}(\omega)|^2 \rangle}{T} \quad (7.21)$$

with $\tilde{E}(\omega)$ the Fourier transform of $E(t)$. The squared transform of the electric field is called the *power spectrum*, and is proportional to the intensity I_ν , with $\nu = \omega/2\pi$.

For some purposes this description is too simplistic, as it hides the vector nature of the field, which allows the possibility of polarized radiation, and for eg radiative transfer one needs a more detailed description which gives the radiation flux in the different polarization modes.

7.5.1 Polarization of Planar Waves

A *linearly polarized* plane wave may be written as

$$\mathbf{E}(\mathbf{r}, t) = \hat{\mathbf{a}} \text{Re} \left[E_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \right] \quad (7.22)$$

where E_0 is a complex amplitude, and $\hat{\mathbf{a}}$ is perpendicular to \mathbf{k} . The most general plane wave (with $\mathbf{k} \propto \hat{\mathbf{z}}$) is

$$\mathbf{E}(t) = \text{Re}[(\hat{\mathbf{x}} E_x + \hat{\mathbf{y}} E_y) e^{-i\omega t}] = \text{Re}[\mathbf{E}_0 e^{-i\omega t}] \quad (7.23)$$

where, for simplicity, we give the value of the field at $z = 0$. Both E_x and E_y are complex amplitudes, and \mathbf{E}_0 is a complex vector. Writing $E_x = \mathcal{E}_x e^{i\varphi_x}$ and $E_y = \mathcal{E}_y e^{i\varphi_y}$, with $\mathcal{E}_x, \mathcal{E}_y$ real, the physical field components are

$$\mathbf{E}(t) = \begin{bmatrix} E_x(t) \\ E_y(t) \end{bmatrix} = \begin{bmatrix} \mathcal{E}_x \cos(\omega t - \varphi_x) \\ \mathcal{E}_y \cos(\omega t - \varphi_y) \end{bmatrix}. \quad (7.24)$$

The 2-vector $\mathbf{E}(t)$ moves along a closed periodic curve in E_x, E_y space with period $T = 2\pi/\omega$. In fact, this curve is an ellipse. To see this, write down the parametric formula for an ellipse with

principle axes aligned with the coordinate axes in some rotated frame $\{E'_x, E'_y\} = \{a \cos(\omega(t - t_0)), b \sin(\omega(t - t_0))\}$. Then apply the rotation operator to obtain the formula in an arbitrary frame, rotated by some angle χ :

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} \cos \chi & -\sin \chi \\ \sin \chi & \cos \chi \end{bmatrix} \begin{bmatrix} E'_x \\ E'_y \end{bmatrix}. \quad (7.25)$$

Expanding the cosine terms in (7.24) and comparing terms containing $\cos \omega t$ and $\sin \omega t$ in each of E_x, E_y yields 4 equations which one can solve to obtain a, b, t_0 and χ for given $\mathcal{E}_x, \mathcal{E}_y, \varphi_x$ and φ_y .

Special cases are $a = 0$ or $b = 0$ which result in a *linearly polarized* wave, in which the point $\mathbf{E}(t)$ oscillates back and forth along a line through the origin. For $a = b$ the ellipse becomes a circle, and the wave is said to be *circularly polarized* with *helicity* which is negative or positive if, for a wave propagating towards the observer, the field value rotates clockwise or counter-clockwise respectively.

7.5.2 Polarization of Quasi-Monochromatic Waves

For a nearly monochromatic wave of angular frequency ω_0 (eg a general wave which has been passed through a narrow band filter) we know that each component of the electric field will be locally sinusoidal, but that the amplitude, or *envelope* of the wave will evolve slowly with time. Over any limited time $1/\omega_0 \ll \delta t \ll 1/\delta\omega$, with $\delta\omega$ the width of the filter, the wave will behave much like a pure planar wave discussed above, but over longer periods, the wave will evolve through a succession of different planar wave states. Now it may be that this succession of states has some persistent features; it may be that there is a general tendency for the wave to be found with $E_x > E_y$, in which case we would say the wave has some degree of linearly polarization, or it may be that the wave tends to be found rotating counter-clockwise, in which case we would say that it has some degree of circular polarization. On the other hand, it might be that there are no persistent correlations; the field might be found to be rotating clockwise at one moment and counter-clockwise at another time, for instance. If there is no systematic sense of rotation, and if there is no systematic tendency for the field values to prefer any particular direction in field space then we describe the wave as unpolarized, or ‘natural’.

7.5.3 Power Spectrum Tensor and Stokes Parameters

We can formalize this using the language of random fields; since, in general, the electric field for a collimated beam is a 2-vector valued random function of time, with transform

$$\tilde{E}_i(\omega) = \int dt E_i(t) e^{i\omega t} \quad (7.26)$$

so we can write down the expectation for the product of components of transforms much as for a scalar function:

$$\begin{aligned} \langle \tilde{E}_i(\omega) \tilde{E}_j^*(\omega') \rangle &= \int dt \int dt' \langle E_i(t) E_j(t') \rangle e^{i(\omega t - \omega' t')} \\ &= \int dt e^{i(\omega - \omega')t} \int d\tau \xi_{ij}(\tau) e^{i\omega' \tau} \end{aligned} \quad (7.27)$$

where we have defined the auto-correlation function tensor

$$\xi_{ij}(\tau) \equiv \langle E_i(t + \tau) E_j(t) \rangle. \quad (7.28)$$

Recognizing the first integral in (7.27) as a representation of the Dirac δ -function we have

$$\langle \tilde{E}_i(\omega) \tilde{E}_j^*(\omega') \rangle = 2\pi \delta(\omega - \omega') P_{ij}(\omega') \quad (7.29)$$

where the power spectrum tensor is

$$P_{ij}(\omega) \equiv \int d\tau \xi_{ij}(\tau) e^{i\omega \tau} \quad (7.30)$$

This is very similar to the result for a scalar random field, but with one important distinction: the power here is generally complex. From the definition of $\xi_{ij}(\tau)$ (7.28) we have

$$\xi_{ij}(\tau) = \xi_{ji}(-\tau) \quad (7.31)$$

so $\xi_{ij}(\tau)$ is symmetric under the combination of reflection in time and exchanging $i \leftrightarrow j$. For $i = j$ this implies that both $P_{xx}(\omega)$ and $P_{yy}(\omega)$ are real. For $i \neq j$ though this implies $P_{xy}(\omega) = P_{yx}^*(\omega)$. It also follows from the definition of P_{ij} that $P_{ij}(-\omega) = P_{ij}^*(\omega)$. Because of these symmetries $P_{ij}(\omega)$ has only 4 real degrees of freedom for each $|\omega|$.

Now consider a filtered version of the field

$$E_i^F(t) = \int \frac{d\omega}{2\pi} \tilde{E}_i(\omega) F(\omega) e^{-i\omega t} \quad (7.32)$$

for which the two point auto-correlation function tensor is

$$\langle E_i^F(\tau) E_j^F(0) \rangle = \int \frac{d\omega}{2\pi} \int \frac{d\omega'}{2\pi} F(\omega) F^*(\omega') \langle \tilde{E}_i(\omega) \tilde{E}_j^*(\omega') \rangle e^{-i\omega\tau} \quad (7.33)$$

or, on invoking the definition of $P_{ij}(\omega)$

$$\langle E_i^F(\tau) E_j^F(0) \rangle = \int \frac{d\omega}{2\pi} |F(\omega)|^2 P_{ij}(\omega) e^{-i\omega\tau} \quad (7.34)$$

and, if $F(\omega)$ is a narrow band filter, accepting only frequencies in a narrow range about $\omega = \pm\omega_0$,

$$\xi_{ij}^F(\tau) = \langle E_i^F(\tau) E_j^F(0) \rangle = P_{ij}(\omega_0) e^{-i\omega_0\tau} + P_{ij}^*(\omega_0) e^{i\omega_0\tau}. \quad (7.35)$$

which is manifestly real.

The four real components of the tensor P_{ij} can be taken to be the *Stokes parameters* I , Q , U , V defined by

$$\begin{aligned} I &= P_{xx} + P_{yy} \\ Q &= P_{xx} - P_{yy} \\ U &= P_{xy} + P_{yx} \\ iV &= P_{xy} - P_{yx} \end{aligned} \quad (7.36)$$

The first three of these can be expressed in terms of the zero-lag auto-correlation function tensor:

$$\begin{aligned} I &= \frac{1}{2}(\xi_{xx}^F(0) + \xi_{yy}^F(0)) \\ Q &= \frac{1}{2}(\xi_{xx}^F(0) - \xi_{yy}^F(0)) \\ U &= \xi_{xy}^F(0) \end{aligned} \quad (7.37)$$

while the fourth involves the correlation at a lag of $1/4$ of a wave period:

$$V = \xi_{yx}^F(\tau = \pi/2\omega_0) \quad (7.38)$$

- The Stokes parameters have the following significance: I gives the total energy flux, or intensity. Q and U describe the two states of linear polarization; for $Q = I$, $U = V = 0$ the field is fully linearly polarized with electric field parallel to the x -axis, for $Q = -I$, $U = V = 0$ the field is fully polarized along the y -axis. For $U = \pm I$ (and $Q = V = 0$) the field is polarized along the diagonals. The *circularity* V describes circular polarization, with $V = \pm I$ corresponding to complete circular polarization in the two helicity states.
- One can show that in general $I^2 \geq Q^2 + U^2 + V^2$. For $Q = U = V = 0$ the beam is said to be unpolarized.
- The linear polarization terms Q , U may be measured by passing the beam through a polarizing filter (e.g. a grid of wires for microwaves) and measuring the intensity as a function of angle.

- Equation (7.38) suggests how to measure circular polarization; split a beam, introduce a $1/4$ wave lag in one arm, recombine and measure the intensity.
- The Stokes parameters are additive for a superposition of independent beams. This is not the case if the beams are correlated; the superposition of two opposite helicity circularly polarized beams gives a resultant which is linearly polarized, for example.
- The general radiative transfer equation can be stated much as for the simple unpolarized case, but involves the 4-vector $\{I, Q, U, V\}$, and the scattering coefficient becomes a tensor quantity. See Chandrasehkar's "Radiative Transfer" for details.
- A general beam can be decomposed into two components

$$\begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix} = \begin{bmatrix} I - \sqrt{Q^2 + U^2 + V^2} \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \sqrt{Q^2 + U^2 + V^2} \\ Q \\ U \\ V \end{bmatrix} \quad (7.39)$$

ie as a superposition of a unpolarized beam and a pure elliptically polarized beam.

- One can define the *degree of polarization* as

$$\Pi = \frac{\sqrt{Q^2 + U^2 + V^2}}{I} \quad (7.40)$$

7.6 Problems

7.6.1 Maxwell's equations

Write down Maxwell's equations for the fields \mathbf{E}, \mathbf{B} produced by a distribution of charge with density ρ and current density \mathbf{j} .

Show that these equations imply charge conservation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0 \quad (7.41)$$

Use the Lorentz force law to show that the rate per unit volume at which the fields do work on the charges is equal to $\mathbf{j} \cdot \mathbf{E}$. Now use Maxwell's equations to show that

$$\mathbf{j} \cdot \mathbf{E} + \frac{1}{8\pi} \frac{\partial (E^2 + B^2)}{\partial t} = -\nabla \cdot \mathbf{S} \quad (7.42)$$

where $\mathbf{S} \equiv c\mathbf{E} \times \mathbf{B}/4\pi$. Integrate this over some small volume, and, using the divergence theorem to convert the \mathbf{S} integral to a surface integral, interpret the three terms in the resulting equation. (You may make use of the identity $\nabla \cdot (\mathbf{E} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{B})$)

Assuming a trial solution of the wave-like form

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \quad \mathbf{B} = \mathbf{B}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \quad (7.43)$$

use Maxwell's equations to show that for plane waves propagating in vacuum the vectors $\mathbf{E}_0, \mathbf{B}_0$ and the wave-vector \mathbf{k} form a mutually orthogonal triad; that $\mathbf{E}_0 = \mathbf{B}_0$; and that the phase and group velocity for these waves are equal to c . Compute the mean energy density $\langle U \rangle$ and the mean of the Poynting vector $\langle \mathbf{S} \rangle$.

7.6.2 Energy and momentum of radiation field

Consider a charge q moving in a viscous medium with a frictional force $\mathbf{F}_{\text{visc}} = -\beta\mathbf{v}$. Suppose a circularly polarised electromagnetic wave passes through the medium so the equation of motion of the charge is

$$m \frac{d\mathbf{v}}{dt} = \mathbf{F}_{\text{visc}} + \mathbf{F}_{\text{Lorentz}} \quad (7.44)$$

now assume further that the charged particle is very light so that it's motion is such that the viscous and Lorentz forces balance: $\mathbf{F}_{\text{visc}} + \mathbf{F}_{\text{Lorentz}} \simeq 0$. Let the electric field amplitude be E and the frequency of the wave be ω .

- a. Show that to lowest order in v/c (i.e. neglecting the magnetic contribution to $\mathbf{F}_{\text{Lorentz}}$) the charge moves in a circle in a plane normal to the wave vector with speed qE/β and with radius $qE/\beta\omega$.
- b. Show that the work done by the wave per unit time is $q^2 E^2/\beta$.
- c. Now consider the magnetic force term. Show that this gives rise to a force parallel to the wave vector of amplitude $q^2 E^2/\beta c$. Thus show that in absorbing an amount of energy dW from the wave, the momentum transfer is $dP = dW/c$.
- d. Show that the torque exerted on the fluid is $q^2 E^2/\beta\omega$. What is the angular momentum transferred to the fluid in the course of absorbing an amount of energy dW ?
- e. Show that the absorption cross section is $4\pi q^2/\beta c$.
- f. If we now regard the radiation to be composed of photons of energy $E_\gamma = \hbar\omega$, show that the foregoing results require that momentum of a photon is $P = \hbar\omega/c = E_\gamma/c$ and that the angular momentum is $J = \hbar$.

Chapter 8

Geometric Optics

A perfectly planar wave is

$$f(\mathbf{r}, t) = ae^{i\psi(\mathbf{r}, t)} \quad (8.1)$$

where f might denote either E or B (and we shall not concern ourselves here with polarization), and $\psi(\mathbf{r}, t) = \omega t - \mathbf{k} \cdot \mathbf{r} + \varphi$ is the phase.

In *geometric optics*, which can be defined as the limiting case of wave optics as $\lambda \rightarrow 0$, the radiation field locally approximates a plane wave:

$$f(\mathbf{r}, t) = a(\mathbf{r}, t)e^{i\psi(\mathbf{r}, t)} \quad (8.2)$$

with slowly varying amplitude $a(\mathbf{r}, t)$ and numerically large phase, or *eikonal*, $\psi(\mathbf{r}, t)$.

Inserting the above in the wave equation

$$\frac{\partial^2 f}{\partial x_i \partial x^i} = \nabla^2 f - \frac{1}{c^2} \frac{\partial^2 f}{\partial t^2} = 0, \quad (8.3)$$

where we have defined the four-vectors $x^i \equiv (ct, \mathbf{x})$ and $x_i \equiv (-ct, \mathbf{x})$, gives

$$\frac{\partial^2 a}{\partial x_i \partial x^i} e^{i\psi} + 2i \frac{\partial a}{\partial x_i} \frac{\partial \psi}{\partial x^i} e^{i\psi} + if \frac{\partial^2 \psi}{\partial x_i \partial x^i} + \frac{\partial \psi}{\partial x_i} \frac{\partial \psi}{\partial x^i} f = 0. \quad (8.4)$$

But if ψ is very large, the last term here overwhelmingly dominates and we obtain the *eikonal equation*

$$\frac{\partial \psi}{\partial x_i} \frac{\partial \psi}{\partial x^i} = 0 \quad (8.5)$$

or

$$(\nabla \psi)^2 - \frac{1}{c^2} \dot{\psi}^2 = 0. \quad (8.6)$$

Referring to the pure plane wave, for which $\psi = \mathbf{k} \cdot \mathbf{r} - \omega t + \varphi$ and so $\nabla \psi = \mathbf{k}$ and $\dot{\psi} = -\omega$, the eikonal equation says that

$$k^2 - \omega^2/c^2 = 0 \quad (8.7)$$

which is also analogous to the relativistic energy-momentum relation for a massless particle

$$p^2 - E^2/c^2 = 0. \quad (8.8)$$

For the interesting case of $\omega = \text{constant}$, we can write $\psi = -\omega t + \psi_0(\mathbf{r})$ and the eikonal equation is

$$(\nabla \psi_0)^2 = \frac{\omega^2}{c^2} \quad (8.9)$$

The surfaces $\psi_0(\mathbf{r}) = \text{constant}$ are wavefronts. The eikonal equation tells us that they march forward across space with uniform spacing, and the energy propagates perpendicular to the wavefront.

The eikonal equation becomes more useful in situations where the refractive index varies with position.

The eikonal equation can be used as a starting point to derive *Fermat's principle of least time* according to which light rays propagate along lines of extremal time of flight. Fermat's principle can also be understood in terms of constructive interference.

8.1 Caustics

In geometric optics light rays behave much like a stream of ballistic particles, traveling along straight lines in vacuum and, more generally, propagating along extremal time curves if the refractive index varies from place to place. The density of rays will, in general, vary from place to place, and this corresponds to variation in the energy flux.

Consider a collimated uniform beam which passes through an inhomogeneous 'phase screen' (a region with spatially varying refractive index — e.g. shower glass). The phase screen results in a slight corrugation of the wavefront, and consequently the rays will be slightly deflected from their original parallel paths and this will result in spatial variation in the energy flux.

A generic feature of such light deflection is the developments of *caustic surfaces* on which the flux is infinite. To analyze this, consider a phase screen which only causes a deflection in one direction (say along the x -axis). Label the rays by their initial spatial coordinate x . We will also refer to this as the *Lagrangian coordinate*. After suffering a phase perturbation $\phi(x)$ the wavefront will be advanced by an amount $h = \lambda\phi/2\pi$ and the normal to the wavefront will be tilted by an angle $\theta(x) = \nabla h(x) = \lambda\nabla\phi(x)/2\pi$.

After propagating a distance D beyond the phase screen, the rays will have been deflected, or *mapped*, to an *Eulerian coordinate*

$$r(x) = x + D\theta(x) \quad (8.10)$$

this is called a *Lagrangian mapping*. Let the rays initially have a uniform density n_0 in Lagrangian space. The density in Eulerian coordinates is given by the *Jacobian* of the mapping since $n dr = n_0 dx$ and therefore

$$n = n_0(dr/dx)^{-1} \propto (1 + Dd\theta/dx)^{-1}. \quad (8.11)$$

This means that for those rays which passed through a region of the phase screen where $d\theta/dx < 0$ the density of rays, and therefore the energy flux, will become infinite at finite distance $D = (d\theta/dx)^{-1}$. This is the phenomenon seen on the bottom of the swimming pool on a sunny day. A simple caustic is shown in figure 8.1.

The generic nature of the resulting caustics can be found from a simple graphical argument. Figure 8.2 shows r vs x for various depths D for a simple sinusoidal deflection $\theta = \theta_0 \cos(x)$. The intermediate curve is plotted for the distance at which infinite flux first appears. At greater depth $r(x)$ has pairs of maxima and minima at which one has a 'fold catastrophe'. Observers lying between these points will see triple images of the distant source source of illumination.

The energy flux is formally infinite at positions r_c corresponding to turning points (ie $dr/dx = r' = 0$). In the vicinity of such a point, with Lagrangian coordinate x_c

$$r = r_c + \frac{1}{2}r''(x_c)(x - x_c)^2 \quad (8.12)$$

so a small interval $\Delta r = r - r_c$ neighboring a caustic corresponds to an interval in Lagrangian space of

$$\Delta x = \left(\frac{2}{r''(x_c)} \right)^{1/2} \sqrt{\Delta r} \quad (8.13)$$

and since the rays are uniformly distributed in x , the density of rays near a caustic is

$$n \propto \Delta x / \Delta r \propto 1/\sqrt{\Delta r}. \quad (8.14)$$

This is the universal scaling law for the intensity close to fold caustics — the generic type of caustic. The flux density diverges inversely as the square root of distance from the caustic surface, $F \propto 1/\sqrt{r}$, but the integrated flux $\int dr F$ from the caustic is finite.

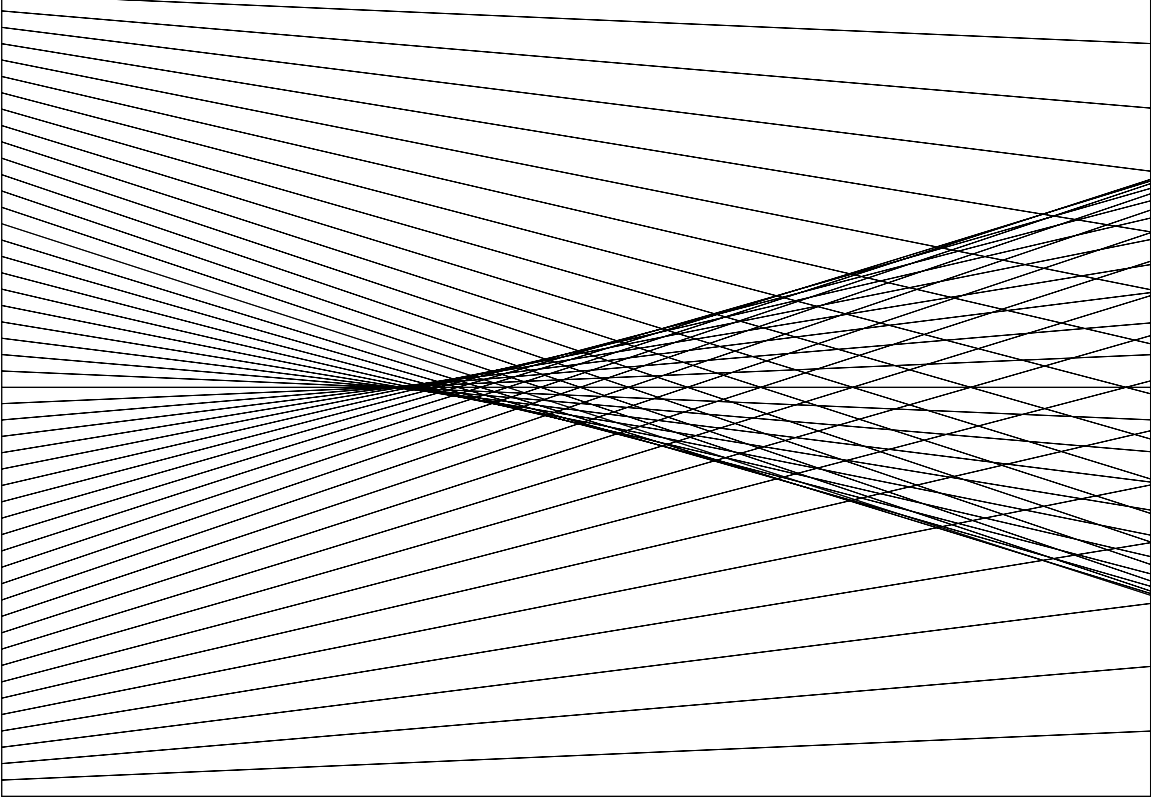


Figure 8.1: Formation of a caustic. The rays illustrated here (propagating left to right) have been subjected to a smoothly varying deflection field. This leads to focusing of rays and formation of a caustic. It is apparent that at each point inside the caustic surface there are three different ray directions, and an observer would see three images of the assumed distant source. The high spatial density of rays just inside the caustic surface correspond to high amplification of the flux.

A counter-example is well figured lens, which produces a different type of singularity - the flux becomes infinite at a point rather than on a surface. However, this is a special and degenerate case. For any finite degree of aberration of the lens the perfect conical focus will degenerate into a ‘astroid’ type caustic.

Caustics are always associated with the appearance of multiple images. As one passes through the surface, a pair of infinitely bright images will appear at the same point on the sky. They will then rapidly move apart and become fainter. Caustic surfaces may be nested within each other, and further pairs of images may appear. This leads to the ‘odd number of images’ theorem often invoked in gravitational lensing.

We have assumed here perfect geometric optics. For finite wavelength the caustics may be smeared out. We have also worked in 1-dimension for simplicity, but the results are readily extendible to a 2-dimensional deflection screen, for which the amplification, for instance, is given by the inverse of the Jacobian of the transformation from Lagrangian to Eulerian space:

$$A = \left| \frac{\partial r_i}{\partial x_j} \right|^{-1} = \left| \delta_{ij} + D \frac{\partial^2 h}{\partial r_i \partial r_j} \right|^{-1}. \quad (8.15)$$

This is readily generalized to finite source distance. Writing D_{LS} for the distance from the source to the deflecting screen (or lens) and D_{OS} from the source to the observer (ie the point where we

1-D deflection

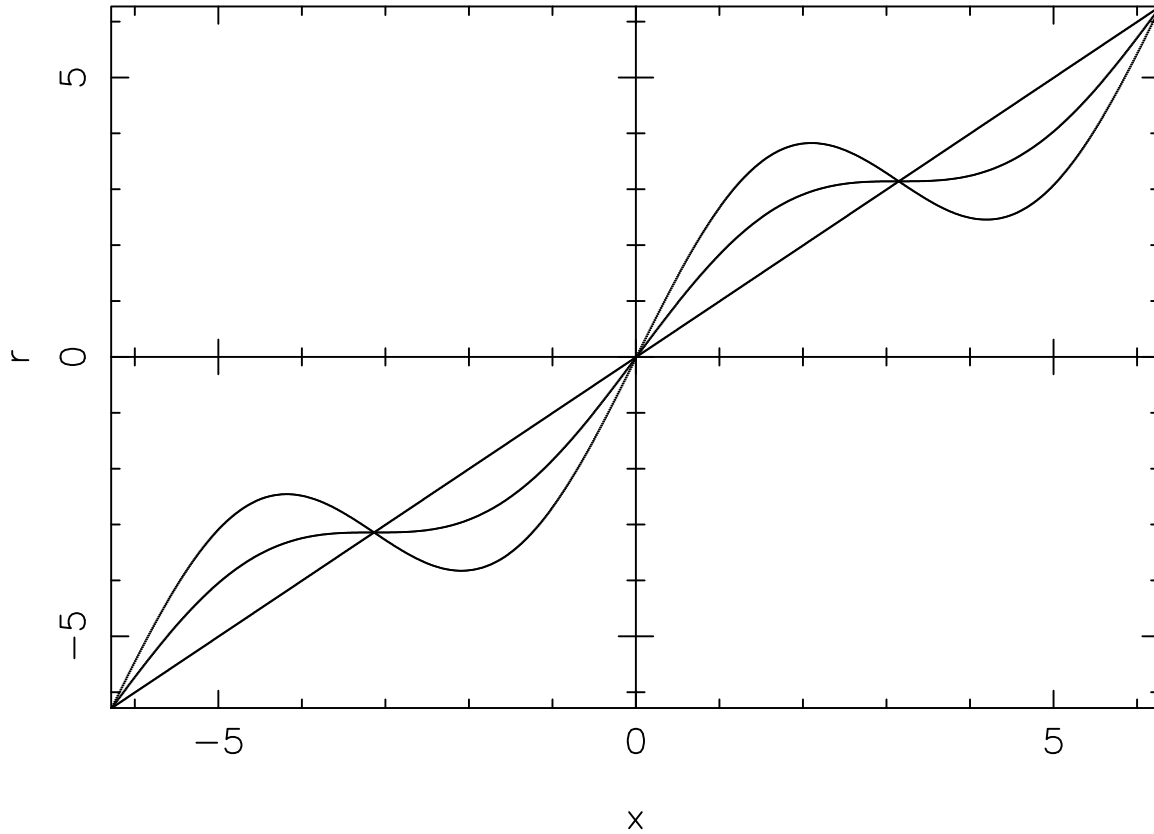


Figure 8.2: Eulerian r coordinate versus Lagrangian x for sinusoidal deflection screen for increasing distance behind the deflection screen.

measure the energy flux gives

$$A = \left| \delta_{ij} + \frac{D_{\text{LS}} D_{\text{OL}}}{D_{\text{OS}}} \frac{\partial^2 h}{\partial r_i \partial r_j} \right|^{-1}. \quad (8.16)$$

One can look at this from an alternative point of view. Consider a narrow conical bundle of rays propagating back from the observer and map these onto the source plane. For a narrow cone, these get mapped to an ellipse, with area proportional to the amplification. This is consistent with the result that the intensity, or surface brightness, of sources is not affected by static deflections; the way in which the amplification arises is by changes in the apparent *angular size* of sources.

The matrix appearing here is called the *distortion tensor* and can be used to describe the change in shapes of objects seen through inhomogeneous refractive media.

8.2 Random Caustics

In many cases of interest the phase screen is a random function. One example is gravitational ‘micro-lensing’, and other random deflecting screens arise in the interstellar medium and in the atmosphere, though in the latter cases, the situation is generally complicated by diffraction. Another physical realization of random caustics which is of great familiarity is the illumination pattern on the bottom of the swimming pool.

8.2.1 Probability for Amplification

An observer monitoring a source through a random refractive medium will see spikes in the flux from the source as the caustic surfaces sweep past (usually because of motion of the observer relative to the source). A direct corollary of the universal scaling law for the profiles of fold caustics is that there is a universal probability law for high amplifications, regardless of whether the deflecting screen is highly non-Gaussian (as in micro-lensing) or Gaussian (as in the swimming pool or the atmosphere).

The probability that a source will appear amplified by at least a factor A will be proportional to the fraction of Eulerian space in which the density of rays is $n > An_0$. Since $n \propto 1/\sqrt{\Delta r}$, this fraction is proportional to $\Delta r \propto 1/A^2$, so $P(> A) \propto 1/A^2$, and therefore the differential probability distribution is

$$p(A)dA \propto dA/A^3. \quad (8.17)$$

8.2.2 Caustics from Gaussian Deflections

For a Gaussian random phase screen one can compute the power spectrum and auto-correlation function of the illumination pattern.

Consider a set of finely, but uniformly, spaced initially parallel rays in Lagrangian space, with $x = i\Delta x$. We will work here in 1-dimension for simplicity, but the results are readily generalized to 2-dimensions.

Mapping these rays to Eulerian space yields a density field

$$n(r) = \sum_i \delta(r - x_i - D\theta(x_i)) \quad (8.18)$$

with transform

$$\tilde{n}(k) \equiv \int dr n(r) e^{ikr} = \sum_i e^{ik(x_i + D\theta(x_i))} \rightarrow \frac{1}{\Delta x} \int dx e^{ikx} e^{ikD\theta(x)}. \quad (8.19)$$

The power spectrum of the density of rays in Eulerian space is

$$P_E(k) \propto \langle |\tilde{n}(k)|^2 \rangle \quad (8.20)$$

or

$$P_E(k) \propto \int dx \int dx' \langle e^{ikD(\theta(x) - \theta(x'))} \rangle e^{ik(x - x')}. \quad (8.21)$$

For a Gaussian random variate ζ it is easy to show that $\langle e^{i\zeta} \rangle = e^{-\langle \zeta^2 \rangle / 2}$. Now the quantity $kD(\theta(x) - \theta(x'))$ is a Gaussian random variate, so

$$\langle e^{ikD(\theta(x) - \theta(x'))} \rangle = e^{-k^2 D^2 S_\theta(x)/2} \quad (8.22)$$

with $S_\theta(x) \equiv \langle (\theta(x) - \theta(0))^2 \rangle$ the structure function for the deflection angle, and so

$$P_E(k) \propto \int dx e^{-k^2 D^2 S_\theta(x)/2} e^{ikx}. \quad (8.23)$$

This gives the power-spectrum of, and therefore the auto-correlation of, the density in Eulerian space in terms of the two point function or power spectrum of the deflection in Lagrangian space, since

$$S_\theta(x)/2 = \xi_\theta(0) - \xi_\theta(x) = \int \frac{dk}{2\pi} P_\theta(k) (1 - \cos kx). \quad (8.24)$$

An interesting model is a power-law power spectrum $P_\theta(k) = P_\star (k/k_\star)^n$. The structure function is then

$$S_\theta(x)/2 = P_\star k_\star^{-n} \int \frac{dk}{2\pi} k^n (1 - \cos kx) = 2P_\star k_\star^{-n} x^{-(n+1)} I_n \quad (8.25)$$

with

$$I_n \equiv \int \frac{dy}{2\pi} y^n (1 - \cos y). \quad (8.26)$$

Convergence of this dimensionless integral at low and high frequencies requires $-3 < n < -1$, and for spectral indices in this range, the structure function is a power law in x :

$$S_\theta(x)/2 = I_n P_\star k_\star^{-n} x^{-(n+1)}. \quad (8.27)$$

As a specific example, consider the case $n = -2$. We then have $S_\theta(x)/2 = I_{-2} P_\star k_\star^2 |x|$, or equivalently $k^2 D^2 S_\theta(x)/2 = k^2 |x|/k_0(D)$ with

$$k_0(D) \equiv \frac{1}{D^2 P_\star k_\star^2 I_{-2}} \quad (8.28)$$

so

$$P_E(k) \propto \int dx e^{-k^2 |x|/k_0} e^{ikx} = \int_0^\infty dx e^{-(k^2/k_0 - ik)x} + \text{c.c.} = \frac{2k_0}{k_0^2 + k^2}. \quad (8.29)$$

This functional form is known as a *Lorentzian profile*. The power spectrum thus has a double power law spectrum, $P_E(k) \propto k^0, 1/k^2$ at low and high spatial frequencies respectively. The break between these two asymptotic power laws occurs at $k = k_0 \sim 1/(k_\star^2 D^2 P_\star)$. The significance of this characteristic scale is made more transparent if we note that the rms deflection angle on scale $\lambda \sim 1/k$ is $\Delta\theta(k) \sim \sqrt{k P(k)} = \sqrt{P_\star k_\star^2/k}$ so $\Delta\theta(k_0) \sim \sqrt{P_\star k_\star^2/k_0} = 1/k_0 D$, or equivalently, $\lambda_0 \sim D \Delta\theta(\lambda_0)$. The characteristic scale λ_0 is such that the ‘focal length’ for fluctuations on this scale is on the order of D . Since $\Delta\theta(\lambda) \propto \sqrt{\lambda}$ for this spectral index, this implies that the characteristic scale varies with distance is $\lambda_0 \propto D^2$.

It is interesting to note that the above result actually violates the ‘universal scaling law’ for high amplification $P(A) \propto 1/A^3$ defined above. To see that these are in conflict note that the second moment of the amplification computed from $P(A)$ is $\langle A^2 \rangle \int dA A^2 P(A)$ and is logarithmically divergent at large A . Computing $\langle A^2 \rangle = \int \frac{dk}{2\pi} P_E(k)$ using the above equation, in contrast, gives a perfectly finite result. The resolution of this seeming paradox is that in deriving the universal scaling we implicitly assumed that the deflection is smoothly varying, and, in particular that the mean square gradient of the deflection $\langle \theta'^2 \rangle$ is finite. For the $n = -2$ case considered above, however, the mean square deflection gradient is $\langle \theta'^2 \rangle \sim \int dk k^2 P_\theta(k)$ which is ultra-violet divergent (i.e. it grows without limit as $k \rightarrow \infty$).

If we modify the deflection spectrum and impose some kind of cut-off at large k such that $\int dk k^2 P(k)$ becomes finite then we recover the log-divergent mean square amplification predicted by $P(A) \propto A^{-3}$. To see this note that $\xi_\theta''(0) = -\int \frac{dk}{2\pi} k^2 P(k)$ is then finite. Writing the structure function as $S_\theta(x) = (\xi_\theta(0) - \xi_\theta(x))/2$ and expanding $\xi_\theta(r) = \xi_\theta(0) + \xi_\theta''(0)x^2 + \dots$ we see that $S_\theta(x) \propto x^2$ as $x \rightarrow 0$, as compared to $|x|$ for the pure $n = -2$ power law. Finally, inserting $S_\theta(x) \propto x^2$ in (8.23) and integrating over k we find $\int dk P_E(k) \sim \int dx/x$ as $x \rightarrow 0$ which is indeed logarithmically divergent.

What is happening here physically is that the high frequency deflections in the pure power law model are disrupting the caustics that would otherwise form.

The same type of analysis can be applied in 3-dimensions to the ‘Zel’dovich approximation’ for growth of cosmological structure.

Chapter 9

Diffraction Theory

Geometric optics applies in the limit $\lambda \rightarrow 0$, or effectively $\lambda \ll L$, the size of the ‘wave-packet’. For finite λ , the classical wave uncertainty principle says that a packet of size L cannot be perfectly collimated, but must have a range of momenta $\delta k/k \gtrsim \lambda/L$, and this results in spreading of the wave packet.

Diffraction theory extends geometric optics to allow for the effect of finite extent of stops, apertures, baffles, pupils etc. It is also needed for a proper treatment of scintillation, since diffraction effects will tend to smear out the geometric optics caustics discussed above.

Consider a plane wave with $\mathbf{k} \propto \hat{\mathbf{z}}$ incident on some kind of aperture or ‘pupil’ in the plane $z = 0$. A proper treatment would involve solution of Maxwell’s equations with appropriate boundary conditions at $z = -\infty$ and on the surface of the absorbing stop. Diffraction theory uses a simple, but physically appealing, approximation. We assume that on a surface covering the aperture the wave amplitude ($f = E, B$) is what one would obtain in the absence of the aperture

$$f(x, y, z = 0, t) = f_0 e^{-i\omega t} \quad (9.1)$$

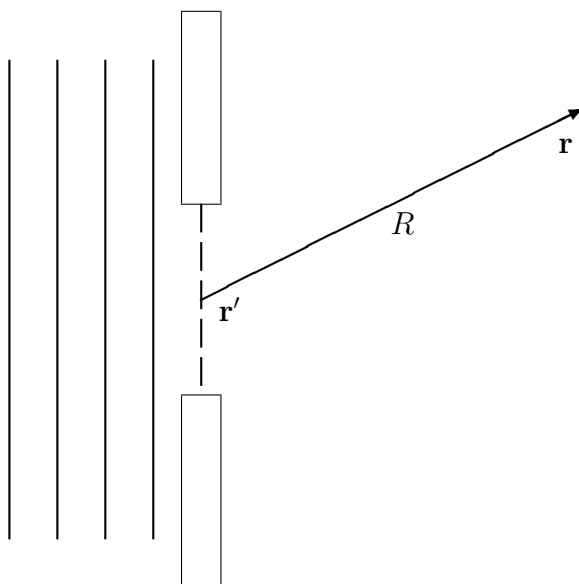


Figure 9.1: Parallel vertical lines at the left represent wavefronts in a collimated beam incident on an aperture. In diffraction theory the field amplitude at some field point \mathbf{r} behind the aperture is computed as the sum over ‘Huygens’ wavelets’ (indicated by the dashed lines) of complex phase factors $e^{i\psi}$ with ψ the path length R in units of $\lambda/2\pi$.

with $\omega = ck$, and that the amplitude of the field at some point to the right of the aperture (ie at $z > 0$) is obtained by summing over elements of the wavefront — called ‘Huygen’s wavelets’ — of the amplitude times a factor $e^{i\psi}/R$ where R is the distance from the wavelet to the point in question and $\psi = kR$ is the phase factor corresponding to path length R for a wave with this wavenumber. More precisely, the factor also contains μ , the cosine of the angle between the vector \mathbf{R} and the normal to the surface element, but here we mostly consider situations where μ is very close to unity. Mathematically, this is

$$f(x, y, z, t = 0) = \int dA \mu \frac{e^{i\psi}}{R} = \int \int dx' dy' \mu \frac{e^{ikR}}{R} \quad (9.2)$$

with $R = \sqrt{(x - x')^2 + (y - y')^2 + z^2}$. The energy density is obtained by squaring the field amplitude:

$$u(\mathbf{r}) = \left| \int dA \mu \frac{e^{i\psi}}{R} \right|^2. \quad (9.3)$$

This is for sharp-edged stops that either transmit the full amplitude or completely block it. The theory can be generalized to apertures with soft edges by introducing a continuous aperture function $A(x', y')$ in the above equations. The square of this function is the fraction of energy transmitted. A phase screen can be described by a complex aperture function.

For rigorous derivation and discussion of the domain of validity of this approximation see Born and Wolfe.

In the geometric optics limit, the effect of a sharp-edged aperture is to introduce a sharp-edged shadow. At small distances behind the aperture, the effect of diffraction is essentially to soften the edges of the shadow. At very large distances, the spreading of the beam is large compared to the width of the aperture, and all that is relevant is the distribution over direction of the wave energy. These two regimes are described as *Fresnel diffraction* and *Fraunhofer diffraction* respectively.

9.1 Fresnel Diffraction

A classic problem in Fresnel diffraction theory is to compute the shadow of a knife edge as seen on a screen placed at some distance $D \gg \lambda$ behind it. Let (x, y) denote the position in the plane of the knife-edge, and (x_0, y_0) the position on the screen (see figure 9.3). By symmetry, the intensity on the screen is independent of y_0 , so let’s calculate the intensity at $(x_0, y_0 = 0)$.

The amplitude on the screen is

$$f(x_0) = \int_{-\infty}^{\infty} dy \int_{x>0}^{\infty} dx e^{ik\sqrt{D^2 + (x-x_0)^2 + y^2}}. \quad (9.4)$$

Consider first some point at large x_0 , ie well away from the geometric shadow. The complex phase factor is stationary for $x = x_0$, $y = 0$ and has constant value e^{ikD} . Expanding the path length as we move away from $(x, y) = (x_0, 0)$ gives

$$\psi = k\sqrt{D^2 + (x - x_0)^2 + y^2} \simeq kD + k[(x - x_0)^2 + y^2]/2D \quad (9.5)$$

so the phase change is small compared to unity over a region in the knife-edge plane of size δx , $\delta y \sim \sqrt{D/k}$ but beyond this the phase change is large and increases rapidly with distance resulting in strong destructive interference. The real part of the complex phase factor is shown in figure 9.2. The upshot of this is that the contribution to the amplitude is dominated by a small region of size $\sqrt{D/k}$ around $(x, y) = (x_0, 0)$. This size is called the *Fresnel length* $r_f \sim \sqrt{D/k} \sim \sqrt{D\lambda}$, ie the geometric mean of the path length and the wavelength.

At a position x_0 much greater than the Fresnel length $x_0 \gg \sqrt{D/k}$ the effect of the knife edge will be negligible, since it cuts off the contribution from regions of the aperture plane which would

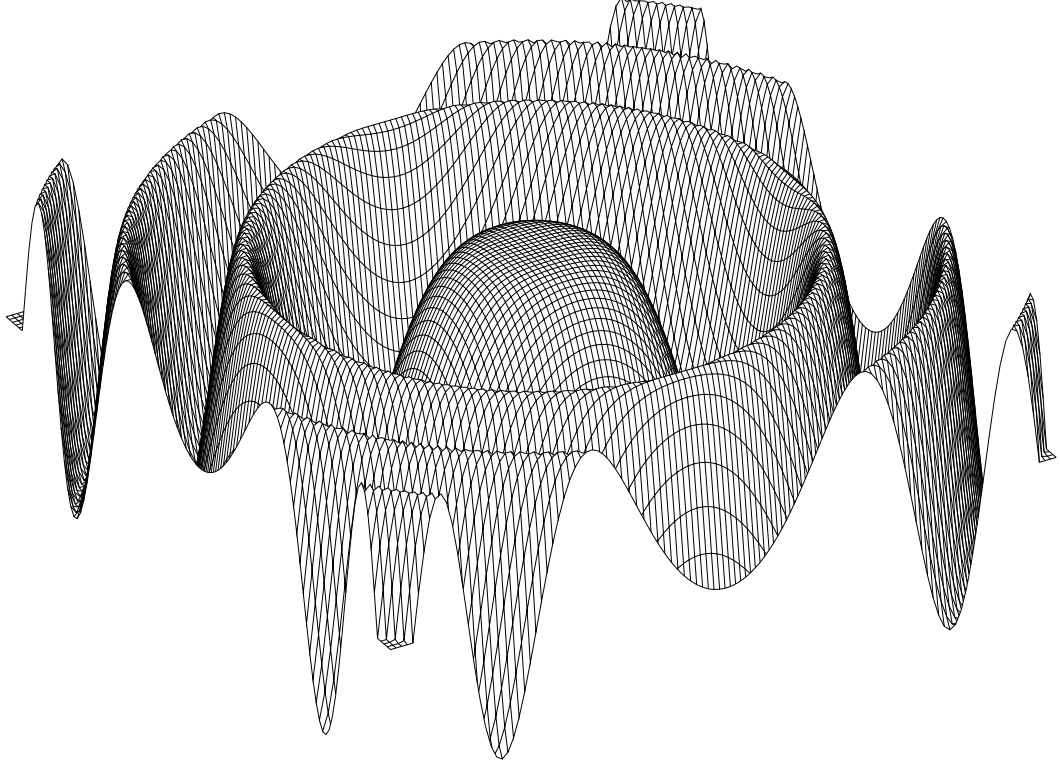


Figure 9.2: The surface shows the real part of the function e^{ix^2} which appears in Fresnel diffraction calculations.

give a negligible contribution even without the screen. For general x_0 we need to allow for the limitation on the integral from the knife-edge and we have

$$f(x_0) \propto \int dy e^{iky^2/2D} \int_{x>0} dx e^{ik(x-x_0)^2/2D}. \quad (9.6)$$

(see figure 9.3). The y -integration introduces a constant factor, which we shall ignore, since we are interested here in the *shape* of the illumination pattern on the screen. Introducing a dimensionless and shifted x -variable $\eta(x) \equiv (x - x_0)\sqrt{k/2D}$ we have for the field amplitude

$$f(x_0) \propto \int_{-w(x_0)}^{\infty} d\eta e^{i\eta^2} = \int_{-w(x_0)}^{\infty} d\eta (\cos \eta^2 + i \sin \eta^2) \quad (9.7)$$

which is clearly a dimensionless function of

$$w(x_0) = \sqrt{k/2D}x_0. \quad (9.8)$$

It is not difficult to show that for large negative w (ie well inside the geometric shadow) the amplitude is $f \propto 1/|w|$, so the intensity $I \sim |f|^2$ is proportional to $1/w^2$. For large positive w the integral is nearly constant with decaying wave-like ripples with scale $\delta w \sim 1$ or $\delta x_0 \sim \sqrt{D\lambda} \sim r_f$. The exact result for the intensity can readily be expressed in terms of the tabulated ‘Fresnel integrals’. For a more detailed and rigorous discussion of this problem see Landau and Lifshitz vol 2.

Consider now instead of a knife-edge an aperture of width L . We saw for a knife-edge that diffraction scatters light a distance $\sim \sqrt{D\lambda}$ beyond the geometric shadow. If the distance to the screen satisfies $\sqrt{D\lambda} \ll L$ then diffraction results in a relatively small modification to the edge of

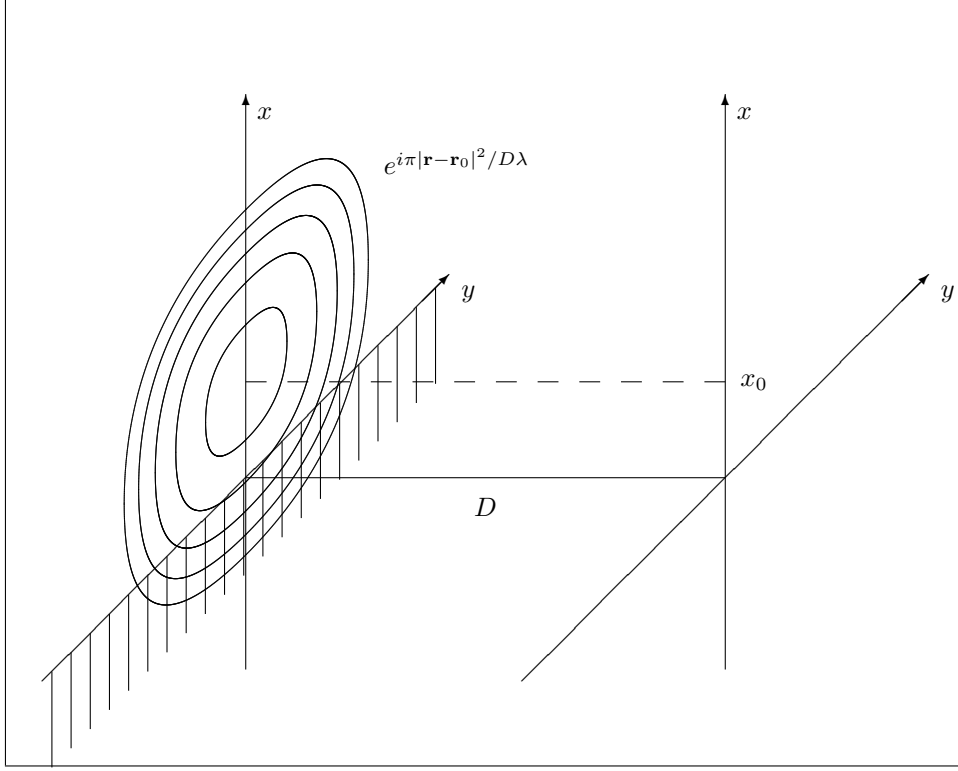


Figure 9.3: To calculate the field amplitude at the point $\mathbf{r}_0 = (x_0, 0)$ in the observer plane at the right we need to integrate the Fresnel function $e^{i\pi|\mathbf{r}-\mathbf{r}_0|^2/D\lambda}$ over the unobscured region in the knife-edge plane. The first few zero rings of the real part of the Fresnel function are illustrated schematically.

shadow pattern much as for the knife-edge. However, this condition fails to apply for sufficiently large aperture-screen distance $D \gtrsim L^2/\lambda$. In the latter regime, the Fresnel scale exceeds the width of the aperture, and a different approach is required.

9.2 Fraunhofer Diffraction

For a two dimensional aperture $A(\mathbf{r})$, the field amplitude at a point $\mathbf{r}_0 = (x_0, y_0)$ is again a 2-dimensional integral over the aperture plane:

$$f(\mathbf{r}_0) = \int d^2r A(\mathbf{r}) e^{ik\sqrt{D^2+|\mathbf{r}_0-\mathbf{r}|^2}}. \quad (9.9)$$

As before, we can expand the phase factor as

$$\psi = k\sqrt{D^2+|\mathbf{r}_0-\mathbf{r}|^2} \simeq kD + k|\mathbf{r}_0|^2/2D - k\mathbf{r}_0 \cdot \mathbf{r}/D + k|\mathbf{r}|^2/2D \quad (9.10)$$

Now if the aperture-screen distance is very large: $D \gg kL^2$, then since $r \leq L$ the last term here is always small compared to unity and may be neglected, and replacing $k \rightarrow 2\pi/\lambda$ we have

$$f(\mathbf{r}_0) = e^{ik(D+|\mathbf{r}_0|^2/2D)} \int d^2r A(\mathbf{r}) e^{-2\pi i \mathbf{r}_0 \cdot \mathbf{r}/\lambda D}. \quad (9.11)$$

The phase factor in front of the integral is irrelevant, and we recognize this as the transform of the aperture function $f(\mathbf{r}_0) = \tilde{A}(2\pi i \mathbf{r}_0/\lambda D)$.

This gives the amplitude as a function of \mathbf{r}_0 on a screen at great distance D , or equivalently the amplitude in a direction $\boldsymbol{\theta} = \mathbf{r}_0/D$, or again equivalently for the amplitude for waves scattered from

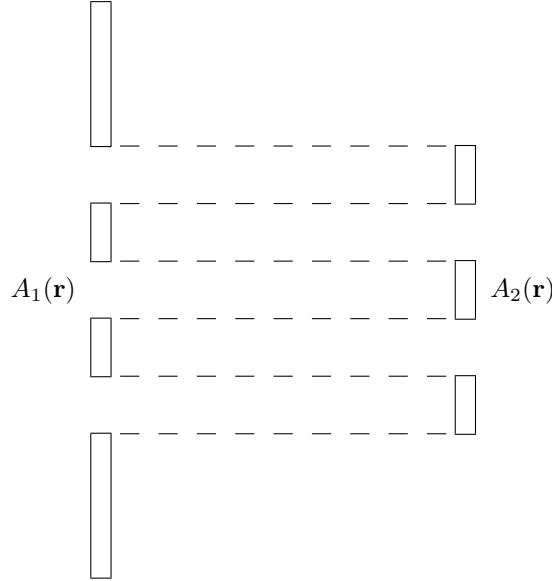


Figure 9.4: Two complementary screens $A_1(\mathbf{r})$ and $A_2(\mathbf{r})$ such that $A_1 + A_2 = 1$, so the light passed by one is blocked by the other and *vice versa*. Babinet's principle tells us that the distribution of *scattered* light is the same for the two screens. Note that in the case of A_2 — for which the blockage is finite in extent — there is a (infinitely) large non-scattered component.

initial wave-number \mathbf{k} into waves with wave-number

$$\mathbf{k}' = \mathbf{k} + \mathbf{q} = \mathbf{k} + \frac{\mathbf{r}_0}{D}k \quad (9.12)$$

and squaring the amplitude gives the intensity for scattered radiation in direction $d\Omega = d\theta_x d\theta_y = dq_x dq_y / k^2$

$$I(\Omega)d\Omega \propto |\tilde{A}(\mathbf{q})|^2 d^2q. \quad (9.13)$$

The key assumption here is that the width of the aperture is much smaller than the Fresnel scale $\sqrt{D\lambda}$. This greatly simplifies the maths, as only a small part of the Fresnel function then contributes, and we can approximate $e^{i\pi r^2/D\lambda}$ by a simple plane wave over the limited region that contributes to the field amplitude calculation.

9.2.1 Babinet's Principle

There is a useful relation between the beam pattern for complementary screens A_1 , A_2 where holes in one match obscuration in the other, or, more generally, where

$$A_1(\mathbf{r}) + A_2(\mathbf{r}) = 1 \quad (9.14)$$

(see figure 9.4). Transforming this gives $\tilde{A}_1(\mathbf{q}) + \tilde{A}_2(\mathbf{q}) \propto \delta(\mathbf{q})$. Now the radiation scattered out of the original beam has, by definition, $\mathbf{q} \neq 0$, so $\tilde{A}_1(\mathbf{q}) = -\tilde{A}_2(\mathbf{q})$ and therefore

$$I_1(\Omega) = I_2(\Omega) \quad (9.15)$$

so the intensity of radiation scattered out of the original direction is the same for complementary screens, which is *Babinet's principle*.

9.3 Telescope Resolution

An interesting application of diffraction theory is to compute the point spread function for a telescope. A well figured lens or mirror introduces a phase lag which turns plane waves from a distant

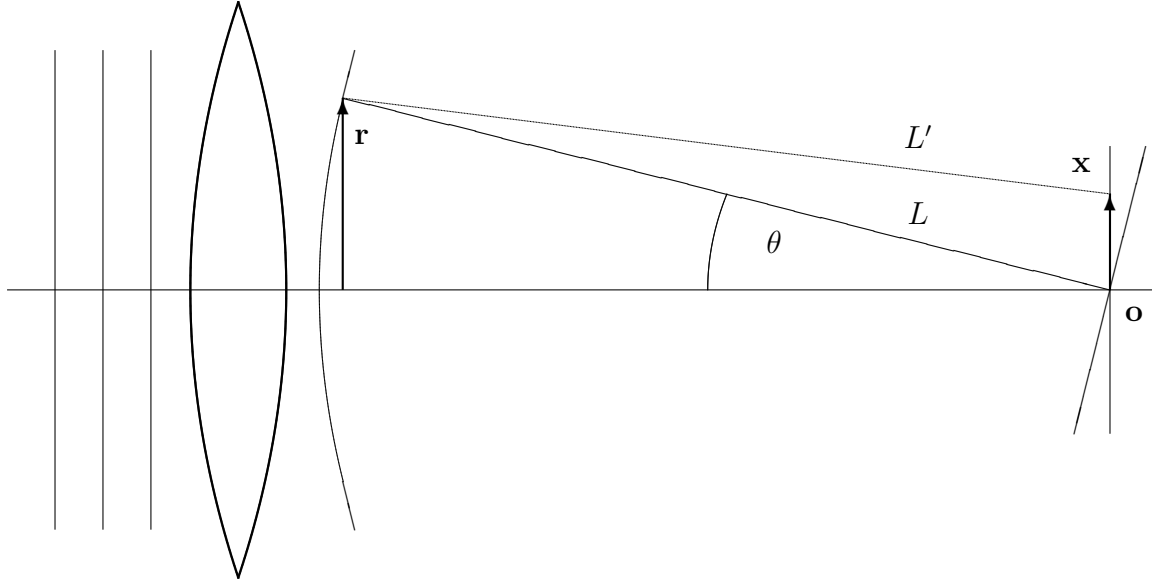


Figure 9.5: A refracting telescope converts incoming planar wavefronts to circular wavefronts converging on the focal point \mathbf{o} . The path length L' from \mathbf{r} to \mathbf{x} is shorter than L by an amount $\Delta L \simeq \theta x$ for small θ . Thus the phase factor in the integration for the field amplitude at \mathbf{x} is $e^{i2\pi\mathbf{x}\cdot\mathbf{r}/L\lambda}$.

on-axis source into spherical wavefronts. In the geometric optics limit $\lambda \rightarrow 0$ these converge to a point on the focal plane. What is the pattern of energy flux on the focal plane as a function of distance \mathbf{x} from the origin when we allow for finite wavelength?

If the focal length of the telescope is L then a shift \mathbf{x} in the focal plane corresponds to an angle $\theta = \mathbf{x}/L$ so the variation in the path length from \mathbf{x} to a point \mathbf{r} on an incoming on-axis wavefront in the *pupil plane* is the same as the distance between an on-axis wavefront and a wave-front tilted by θ , so the phase factor is $\delta\psi = 2\pi\theta \cdot \mathbf{r}/\lambda$ and the field amplitude is

$$f(\mathbf{x}) \propto \int d^2r A(\mathbf{r}) e^{2\pi i \mathbf{x} \cdot \mathbf{r} / L\lambda}. \quad (9.16)$$

Note that this is formally identical to the formula for the field amplitude in the Fraunhofer approximation. The focusing action of the lens makes the amplitude pattern on the focal plane identical (aside from a scale factor) to that computed in Fraunhofer theory for an observer at infinity for the same aperture, but without a lens.

Squaring the field amplitude, as usual, gives the focal-plane energy density pattern, or *point spread function*,

$$g(\mathbf{x}) \propto |f(\mathbf{x})|^2 \propto |\tilde{A}(2\pi i \mathbf{x} / L\lambda)|^2 \quad (9.17)$$

which says that the PSF $g(\mathbf{x})$ is the square of the transform of the aperture function evaluated at $\mathbf{k} = 2\pi i \mathbf{x} / L\lambda$. Equivalently, the PSF is the transform of the auto-correlation function of the aperture function

$$g(\mathbf{x}) \propto \int d^2z e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} \int d^2r A(\mathbf{r}) A^*(\mathbf{r} + \mathbf{z}) \quad (9.18)$$

where we have allowed for the fact that the aperture function $A(\mathbf{r})$ may contain a complex phase factor if there are aberrations for instance.

9.3.1 The Optical Transfer Function

The Fourier transform of the PSF $\tilde{g}(\mathbf{k})$ is, according to (9.18), equal to the auto-correlation of the aperture evaluated at a lag of $z = kL\lambda/2\pi$. The transform of the PSF is known as the *optical*

transfer function (OTF). Since the image $F_{\text{obs}}(\mathbf{x})$ (i.e. the energy flux) formed on the focal plane is the convolution of the true image with the PSF:

$$F_{\text{obs}}(\mathbf{x}) = (F_{\text{true}} \otimes g)_{\mathbf{x}} \quad (9.19)$$

then, by the convolution theorem, the transform of the focal plane image is

$$\tilde{F}_{\text{obs}}(\mathbf{k}) = \tilde{F}_{\text{true}}(\mathbf{k})\tilde{g}(\mathbf{k}). \quad (9.20)$$

Often, there are other sources of blurring in the system. For example, in CCDs there is further image degradation due to charge diffusion, and telescope tracking errors give a further blurring. The utility of the OTF is that in such cases the OTF for the complete system is the product of the separate OTFs due to the optics, the detector, the tracking errors etc.

The choice of argument for the OTF can sometimes be a source of confusion. The wave-vector \mathbf{k} has units of inverse length, whereas sometimes the OTF is taken to be the auto-correlation of the aperture, whose arguments has units of length. The dimensionful conversion factor is $L\lambda/2\pi$. Also, the PSF is often considered to be a function of the (vector) angle on the sky: $\text{PSF} = g(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \equiv \mathbf{x}/L$, in which case the corresponding wave-vector \mathbf{k} is dimensionless. In some cases the convention being used is clearly stated, but more often one has to infer what the argument units are by inspection.

One occasionally finds reference to the *modulation transfer function* (MTF). This is just the modulus of the OTF.

9.3.2 Properties of the Telescope PSF and OTF

Several generic properties of telescope PSFs and OTFs can be inferred from the above equations:

- If the aperture has width D_A then the PSF has width $\Delta x \sim L\lambda/D_A$ (in distance on the focal plane) or $\Delta\theta \sim \lambda/D_A$ in angle.
- Since the aperture is bounded in size, the OTF is bounded in frequency: $\tilde{g}(\mathbf{k}) = 0$ for $\mathbf{k} > 4\pi D_A/L\lambda$. This means that images formed in real telescopes always have *band limited signal* content. By virtue of the *sampling theorem*, there is then a critical sampling rate such that all of the information is preserved. Few optical telescopes, however, are sampled at the critical rate.
- Obstructions of the pupil such as secondary mirror support struts lead to extended linear features known as ‘diffraction spikes’ in the PSF. Similarly, sharp edges in the pupil lead to extended ‘wings’ of the PSF with $g(x) \propto 1/x^3$ for $x \gg \Delta x$. See figures 9.6 and 9.7 accompanying caption for more discussion.
- High frequency *figure errors* also lead to extended wings on the PSF.
- Such wings are a serious impediment for e.g. extra-solar planet searches as the target planet can be swamped by scattered light from its parent star.
- The scattered light in the PSF wings can be greatly reduced by *apodizing* the pupil, though this is not often done. Careful choice of pupil can also help. For example, as shown in figures 9.6 and 9.7, square and circular pupils both produce azimuthally average $g(x) \propto 1/x^3$, but in the former case the scattered light is confined to narrow radial ‘spikes’, and so faint companion objects can still be detected in other parts of the image.

9.3.3 Random Phase Errors

The telescope image is also affected by random phase errors. There arise either from fine-scale *mirror roughness* — imperfections arising in the grinding and/or polishing process — and *atmospheric turbulence*, when mixing of air with varying entropy results in fluctuations in the refractive index.

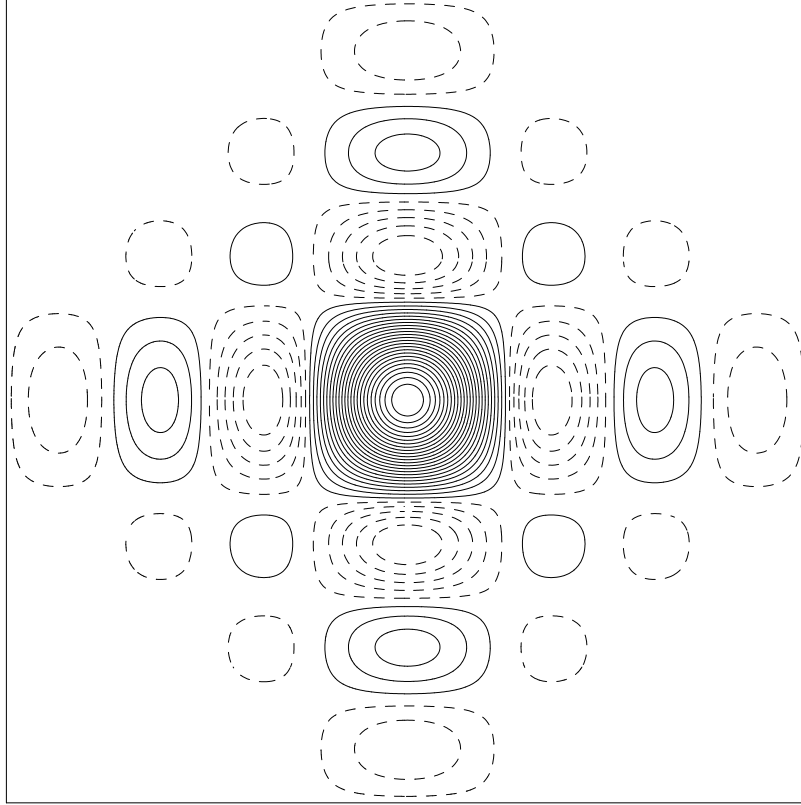


Figure 9.6: A square pupil can be considered to be the product of two one dimensional box-car functions: $A(x, y) = W(x)W(y)$ and consequently its transform is the product of two ‘sinc’ functions: $\tilde{A}(k_x, k_y) = \tilde{W}(k_x)\tilde{W}(k_y)$ with $\tilde{W}(k) = \sin(kD_A/2)/(kD_A/2)$. This (real) function is plotted above, and its square is the PSF. The straight sharp edges of the pupil give rise to extended ‘diffraction spikes’. Now the amplitude of the ripples in the sinc function fall off asymptotically as $1/k$ for large argument, so the amplitude of the diffraction spikes fall as $g(x) \propto 1/x^2$. The width of these features is independent of x , however, so if we azimuthally average the PSF, the energy scattered to large radius falls as $1/x^3$. As we shall see, the same is true for a circular aperture (see figure 9.7).

Phase errors, random or otherwise, can be treated by introducing a phase factor $C(\mathbf{r}) = e^{i\psi(\mathbf{r})}$ which multiplies the pupil function. The field amplitude is then

$$f(\mathbf{x}) \propto \widetilde{AC}. \quad (9.21)$$

Mirror Roughness

If the phase errors due to mirror roughness are small; $\psi(\mathbf{r}) \ll 1$, as is commonly the case, one can approximate $C(\mathbf{r})$ as $C(\mathbf{r}) \simeq 1 + i\phi(\mathbf{r})$. The field amplitude is then the sum of two terms; the first being simply that computed from the pupil function alone, while the second is $f(\mathbf{x}) \propto \widetilde{A\phi}$.

Mirror roughness can be an important contributor to the wings of the PSF. The field amplitude is the convolution of \tilde{A} with $\tilde{\phi}$ evaluated, as usual, at $k = 2\pi x/L\lambda$. Well outside of the diffraction limited core — i.e. at $x \gg \lambda L/D_A$ — the PSF is dominated by the mirror roughness term and

$$g(\mathbf{x}) \simeq |\tilde{\phi}(\mathbf{k} = 2\pi\mathbf{x}/L\lambda)|^2 \quad (9.22)$$

i.e. the PSF is proportional to the power spectrum of the phase fluctuations: $g(x) \propto P_\phi(k(x))$.

Empirically, the telescope PSF is often found to have an extended *aureole* with $g \propto 1/x^2$. This contribution will dominate over edge effects ($g \propto 1/x^3$) and atmospheric effects ($g \propto 1/x^{11/3}$),

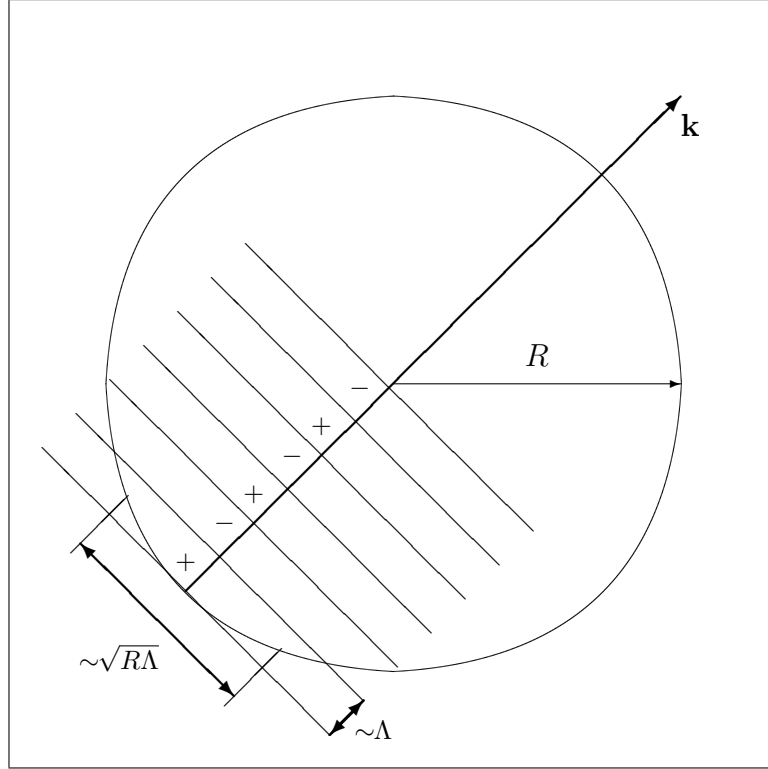


Figure 9.7: The PSF for a circular pupil can be computed exactly in terms of Bessel functions, from which one can show that, asymptotically, the wings of the PSF scale like $1/r^3$. To see how this comes about, and how this result obtains quite generally for any pupil with a sharp edge, consider computing the transform of the pupil at some wave-number $k = 2\pi/\Lambda$ with direction as indicated. The (real part of the) transform is just the product of the pupil function with a plane wave $\cos(\mathbf{x} \cdot \mathbf{k})$, and similarly for the imaginary part. If $k \gg R$ then there will tend to be very good cancellation between the positive and negative half-cycles of the wave. The strongest contribution comes from the last half-cycle, which does not get cancelled. If the pupil is circular then the length of this sliver is $\sim \sqrt{R\Lambda}$ while its width is $\sim \Lambda$, so the net contribution to the transform is $\tilde{A}(k) \sim R^{1/2}\Lambda^{3/2} \propto k^{-3/2}$. Squaring this gives the asymptotic behaviour for the PSF: $g(x) = |\tilde{A}(k = x/2\pi L\lambda)|^2 \propto 1/x^3$.

if present, at very large radii. To generate this requires a power-law phase-error power spectrum $P_\phi(k) \propto 1/k^2$. This is two-dimensional *flicker noise*; with equal contribution to the phase variance from each logarithmic range of frequencies.

In general, the the total scattered light is proportional to the phase error variance (for flicker noise this is logarithmically divergent, as is the integral $\int d^2x g(x)$). This has a non-negligible effect on aperture photometry, especially when comparing results from different telescopes with differing mirror properties.

If the fine-scale mirror surface errors can be approximated as a statistically homogeneous process then the PSF — being the power spectrum of $\phi(\mathbf{r})$ ‘windowed’ with $A(\mathbf{r})$, and it follows that the aureole is speckly with microscale equal to the diffraction limited PSF width.

Atmospheric Turbulence

Atmospheric turbulence is more difficult to analyse, as the phase fluctuations are non numerically small. For pure *Kolmogorov turbulence*, the *phase structure function* is $S_\phi(r) \equiv \langle (\phi(r) - \phi(0))^2 \rangle = 6.88(r/r_0)^{5/3}$ where r_0 is the *Fried length*. On scales larger than r_0 the phase errors are large compared to unity. A realization of a wave-front after passing through a turbulent atmosphere is

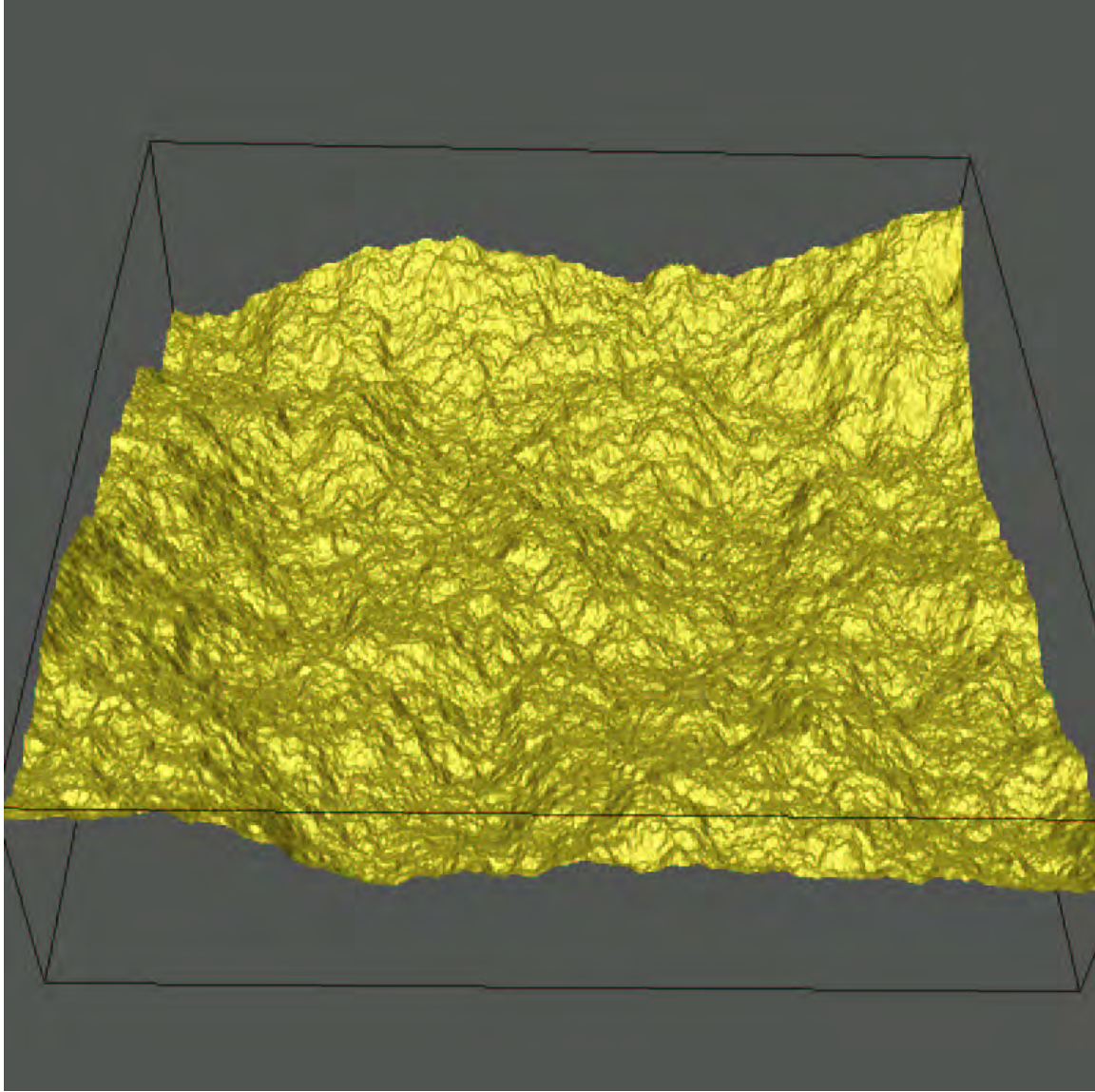


Figure 9.8: A realisation of a wavefront after passing through a turbulent atmosphere. Kolmogorov turbulence produces refractive index fluctuations δn with a power law power-spectrum $P_n(k) \propto k^{-11/3}$, and the same is true of the 2-dimensional phase (this being a projection of the refractive index). The phase fluctuations are strongly ‘infra-red divergent’, whereas the mean deflection, averaged over a patch of a given size, diverges at small scales.

shown in figure 9.8.

For big telescopes it is usually the case that the Fried length is much less than the pupil diameter D . So again $A(\mathbf{r})C(\mathbf{r})$ takes the form of an infinite random field fluctuating on scale r_c modulated, or ‘windowed’, by the aperture function. The transform of AC is therefore, on general grounds, an incoherent function \tilde{C} with overall extent $\sim 1/r_0$ convolved, or smoothed, with \tilde{A} and the squared intensity $|f(\mathbf{x})|^2$ will consist of a set of random *speckles*, each of the size of the PSF for a diffraction limited telescope, but spread over a larger area of the focal plane. A realisation of such a PSF is shown in figure 9.9.

Now the phase fluctuation screen varies with time (primarily due to being convected across the pupil by wind) and this means that the speckles tend to dance around, with a relatively short time-

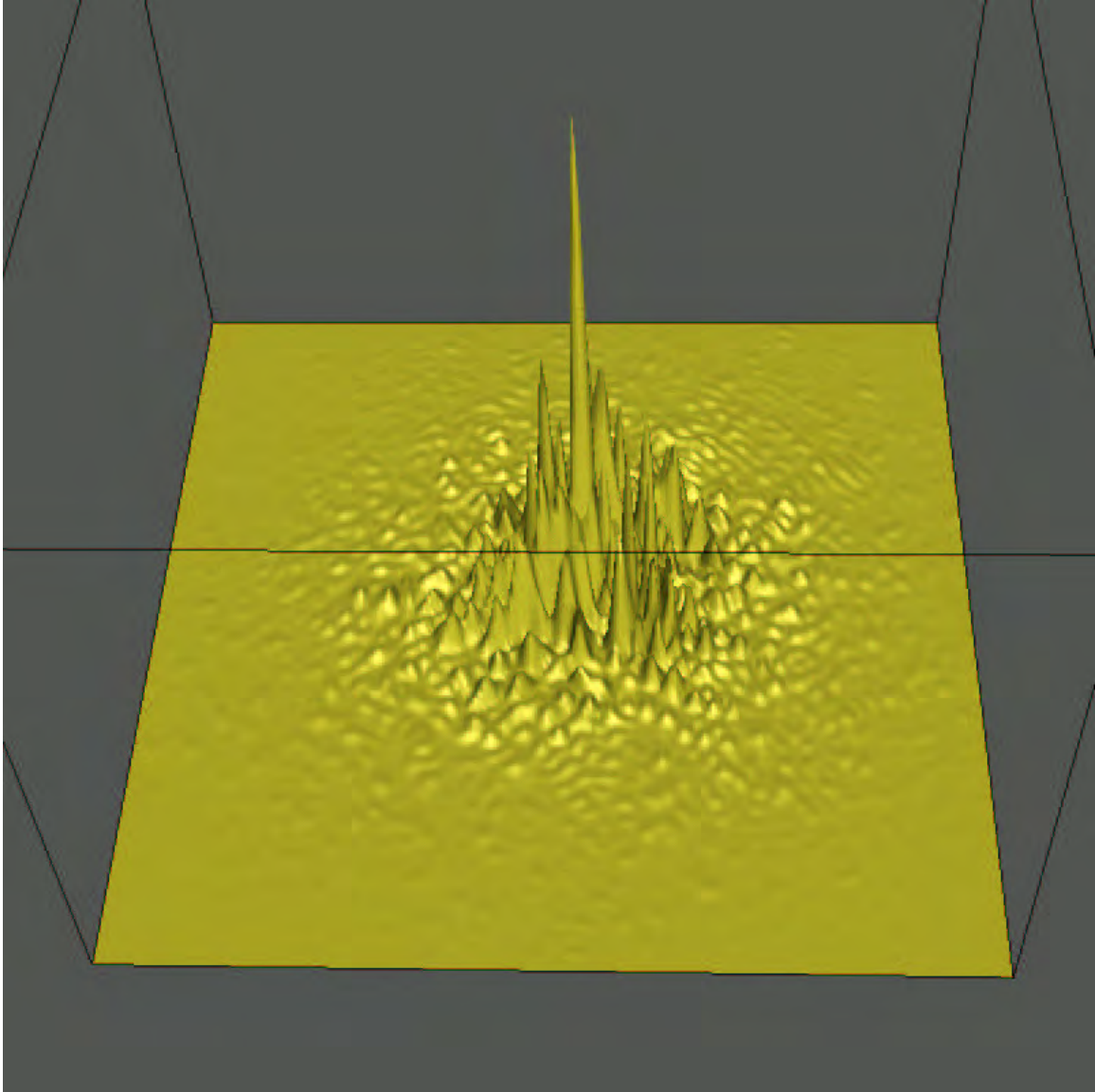


Figure 9.9: A realisation of a PSF arising from atmospheric seeing with Kolmogorov turbulence. In a diffraction limited telescope there are two important length scales; the wavelength λ and the aperture diameter D . The angular resolution, which is of course dimensionless, is $\Delta\theta \sim \lambda/D$. Refractive index fluctuations in the atmosphere introduce an additional length scale: the Fried length r_0 . This such that the root mean squared difference in phase-error for two points with separation r_0 is on the order of one radian. In the atmosphere dominated PSF, each speckle is in the order of the diffraction limit, but the overall angular width of the PSF is $\Delta\theta \sim \lambda/r_0$.

scale and if one integrates longer than this, as is usually the case, then the result is a smeared out PSF with angular resolution $\theta \sim \lambda/r_0$. This is the resolving power of a diffraction limited telescope of aperture diameter $D \simeq r_0$.

For long integrations, it is possible to subtract an azimuthally averaged PSF to help reveal faint companions for example. However, if N , the number of photons per speckle per speckle coherence time is large, then the fluctuations about the mean profile will be larger than the photon counting limit by a factor \sqrt{N} . This excess variance is called *speckle noise*.

For sufficiently bright sources it is possible to observe the instantaneous speckle pattern, and

from this deduce properties of the source on scales comparable to the diffraction limited resolution. This is called *speckle interferometry*.

9.4 Image Wander

The wave-front deformation $h(\mathbf{r})$ generated by Kolmogorov turbulence (figure 9.8) has lots of low-frequency power. The region of the wave-front covering the telescope pupil, of size D , will have a average slope $\nabla h \sim \sqrt{\langle (h(0) - h(D))^2 \rangle} / D$, corresponding to an angular deflection of the same size. The phase fluctuation is $\delta\varphi = h/\lambda$, so the typical net shift in the angular position is expected to be $\delta\theta \sim S_\varphi(D)^{1/2} \lambda / D$. With $S_\varphi(r) = 6.88(r/r_0)^{5/3}$ this is $\delta\theta \sim (\lambda/r_0)(r_0/D)^{1/6}$. The overall angular width of the PSF on the other hand is $\Delta\theta \sim (\lambda/r_0)$. Thus, for an extremely large telescope, the net deflection of the image position caused by the atmosphere becomes negligible. The ratio of the net shift to the PSF width falls off quite slowly, only as the $-1/6$ power of the diameter. For modest size telescopes, a substantial part of the atmospheric image degradation comes from the so-called *image wander*. This is important, since such image wander, as well as image motion induced by the wobbling of the wind-buffed telescope, can be taken out by *tip-tilt correction*. Traditionally this has been achieved by introducing a wobbling mirror, or glass plate, with a servo loop to freeze the motion of a continuously monitored guide star. More recently it has become possible to take out the image motion by shuffling the accumulating electronic charge around the CCD.

The above estimate of the deflection — that it is just some kind of average of the wavefront slope across the pupil — is rather hand waving. We can make this more precise if we consider the *centroid of the PSF*. The starting point is the expression for the PSF as the Fourier transform of the OTF (9.18). We will write this as

$$g(\mathbf{x}) \propto \int d^2z e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} \chi(\mathbf{z}) \quad (9.23)$$

with

$$\chi(\mathbf{z}) = \int d^2r A(\mathbf{r}) A(\mathbf{r} + \mathbf{z}) e^{i(\varphi(\mathbf{r}) - \varphi(\mathbf{r} + \mathbf{z}))} \quad (9.24)$$

i.e. the auto-correlation function of the real pupil function A times the atmospheric phase factor $e^{i\varphi}$.

The first moment of the PSF is

$$\int d^2x \mathbf{x} g(\mathbf{x}) = \int d^2z \chi(\mathbf{z}) \int d^2x \mathbf{x} e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} \quad (9.25)$$

$$= \frac{L\lambda}{2\pi i} \int d^2x \int d^2z \chi(\mathbf{z}) \nabla_z e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} \quad (9.26)$$

$$= -\frac{iL\lambda}{2\pi} \int d^2x \int d^2z e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} \nabla_z \chi \quad (9.27)$$

$$= -\frac{iL\lambda}{2\pi} \int d^2z \delta(2\pi \mathbf{z} / L\lambda) \nabla \chi \quad (9.28)$$

$$= -i \left(\frac{L\lambda}{2\pi} \right)^2 (\nabla \chi)_{\mathbf{z}=0} \quad (9.29)$$

where we have first used the fact the $\nabla_z e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} = (2\pi \mathbf{x} / L\lambda) e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda}$. We then integrated by parts to shift the gradient operator from the complex exponential to χ . We then recognized the x -integral on the third line as a Dirac δ -function.

The zeroth moment of the PSF is similarly found to be

$$\int d^2x g(\mathbf{x}) = \int d^2z \chi(\mathbf{z}) \int d^2x e^{2\pi i \mathbf{x} \cdot \mathbf{z} / L\lambda} = \frac{L\lambda}{2\pi} \chi(\mathbf{z} = 0). \quad (9.30)$$

The *centroid of the PSF* is defined as

$$\bar{\mathbf{x}} = \frac{\int d^2x \mathbf{x} g(\mathbf{x})}{\int d^2x g(\mathbf{x})} = \frac{iL\lambda}{2\pi} \left(\frac{\nabla \chi}{\chi} \right)_{\mathbf{z}=0}. \quad (9.31)$$

Now the gradient of (9.24) is

$$\nabla_z \chi = \nabla_z \int d^2r A(r) e^{i\varphi(\mathbf{r})} A(\mathbf{r} + \mathbf{z}) e^{-i\varphi(\mathbf{r} + \mathbf{z})} \quad (9.32)$$

$$= \int d^2r A(r) e^{i\varphi(\mathbf{r})} [\nabla A(\mathbf{r} + \mathbf{z}) - iA(\mathbf{r} + \mathbf{z}) \nabla \varphi(\mathbf{r} + \mathbf{z})] e^{-i\varphi(\mathbf{r} + \mathbf{z})} \quad (9.33)$$

If we now evaluate this at $\mathbf{z} = 0$ we have

$$(\nabla \chi)_{\mathbf{z}=0} = \int d^2r A(\mathbf{r}) \nabla A(\mathbf{r}) - i \int d^2r A^2(\mathbf{r}) \nabla \varphi. \quad (9.34)$$

The first term is the integral of a total derivative $A(\mathbf{r}) \nabla A(\mathbf{r}) = \nabla A^2/2$, which therefore vanishes, and, with (9.31) we finally obtain

$$\bar{\mathbf{x}} = \frac{L\lambda}{2\pi} \frac{\int d^2r W(\mathbf{r}) \nabla \varphi(\mathbf{r})}{\int d^2r W(\mathbf{r})} \quad (9.35)$$

where $W(\mathbf{r}) = A^2(\mathbf{r})$ is the square of the pupil function (and which is equal to $A(\mathbf{r})$ in the usual case that the pupil transmission is either unity or zero).

Thus the PSF centroid position is indeed a weighted average of the slope of the wavefront deformation or phase-error across the pupil. What is remarkable about this result is that while the actual PSF is a highly non-linear function of the phase-error — since the phase appears in complex exponential factors — the centroid is a linear function of φ . This means that if the phase-error has Gaussian statistics, as the central limit theorem would encourage us to believe is the case, then the centroid is also a Gaussian random variable. Given the power-law form for the power spectrum for the phase-error, one can, for instance, compute the temporal power-spectrum or auto-correlation function of the deflection $\bar{\mathbf{x}}(t)$ (problem-TBD). Given a model for the distribution of seeing with altitude one can also compute the co-variance of the deflection for different stars. This is all very convenient and elegant. However, we should point out that the centroid statistic, while mathematically appealing, is quite useless in practical applications as it has terrible noise properties because the integral must be taken over the entire noisy image. What is usually done in reality is to smooth the image with some kernel with shape similar to the average PSF and then locate the peak. Unfortunately, computing the statistical properties of this smoothed-peak motions is much more complicated, and these motions do not have precisely Gaussian statistics. The behaviour for such realistic position estimators can be found from simple numerical experiments, as one can easily generate a large realization of a phase-screen like that in figure 9.8 and then drag this across a model pupil and compute the PSF as a function of time numerically. Such experiments, and indeed analytic reasoning, suggest that the statistical properties of the smoothed-peak motions are quite similar to those of the centroid, but that the high temporal frequency behavior is rather different.

9.5 Occultation Experiments

The Fresnel scale has a useful significance in occultation experiments. Let us assume that one monitors a large number of distant stars to look for occultations by objects in the solar system. This approach is limited by diffraction effects. In the discussion of the knife-edge, we saw that the waves which interfere constructively cover a ‘Fresnel zone’ in the occulting plane of size $\sim \sqrt{D\lambda}$, and so strong occultation is expected only for objects which are larger than

$$L \simeq 1.2\text{km} \left(\frac{R}{\text{AU}} \right)^{1/2} \left(\frac{\lambda}{1\mu} \right)^{1/2}. \quad (9.36)$$

9.6 Scintillation

Consider radiation at wavelength λ from a point source propagating a distance D through a random refractive medium. Let the medium consist of clouds or domains of size r_c with random fractional refractive index fluctuations $\delta n/n$.

The light crossing time for a domain is $t_c = r_c/c$, and the fluctuations δn introduce fractional changes in the light crossing time $\delta t/t \sim \delta n/n$ leading to corrugation of the wavefronts of amplitude $\delta h \sim c\delta t \sim r_c\delta n/n$, corresponding to angular tilt (induced by a single cloud) of $\delta\theta_1 \sim \delta n/n$. The total deflection will be the sum of $N \sim D/r_c$ random deflections or

$$\delta\theta \sim \sqrt{\frac{D}{r_c}} \frac{\delta n}{n}. \quad (9.37)$$

(see figure 9.10).

If the refractive index fluctuations are sufficiently strong and/or if the path length is sufficiently long, this will lead to multi-path propagation. In the geometric optics regime, the condition for multi-path propagation is

$$D\delta\theta \gtrsim r_c. \quad (9.38)$$

For geometric optics to be valid we require that the Fresnel scale be smaller than the separation of the rays

$$r_f = \sqrt{D\lambda} \ll D\delta\theta \quad (9.39)$$

since only if this condition is satisfied do we really have well defined separate paths. If this condition is satisfied then in the vicinity of the observer there will be a superposition of waves with slightly different angles, resulting in a periodic interference pattern. If the observer is moving relative to this pattern, as is generally the case, this will result in oscillations in the apparent brightness of the source. The spatial scale of these oscillations is $\Delta s \sim \lambda/\delta\theta$.

In the case of the ISM, this mechanism — known as *diffractive inter-stellar scintillation* or DISS — can yield important information about the nature of the intervening medium even if the angular splitting is too small to be resolved. Imagine, for instance, one observes a clean sinusoidal oscillation in brightness. From this one can infer that there are two paths interfering. From the time-scale of the oscillations, together with some estimate of the observer's velocity relative to the fringe-pattern, one can infer Δs and from this the angular splitting $\delta\theta \sim \lambda/\Delta s$, and from this, together with some estimate of the distance of the source, one can infer $r_c \sim D\delta\theta$, the characteristic size of the clouds. This in turn tells us $N \sim D/r_c$, the number of clouds along a line of sight, and hence one can infer the strength of the refractive index fluctuations $\delta n/n \sim \delta\theta/\sqrt{N}$.

Note that this mechanism requires that the source be small: $\delta\theta_s \lesssim \Delta s/D$. Conversely, this mechanism can provide information about the sizes of sources which cannot be resolved.

If the geometric optics condition is violated then the diffraction pattern will be washed out. There may still be fluctuations in the source brightness, but these are caused by the net amplification averaged over a cigar shaped volume of width $\sim r_f$. This tends to be weaker, and occurs on a much longer time-scale. This is called *refractive inter-stellar scintillation* or RISS.

The atmosphere also causes scintillation; the ‘twinkling’ of stars. Atmospheric turbulence causes refractive index fluctuations with spectral index $n = -11/3$, and the two-dimensional structure function for phase fluctuations is therefore $\langle(\delta\psi)^2\rangle \sim (r/r_0)^{5/3}$. In good observing conditions, and in the optical, the phase-correlation length (or *Fried length*) is on the order of $r_0 \sim 40\text{cm}$. The amplitude of the wavefront corrugations on scale r are therefore $\delta h \sim \lambda(r/r_0)^{5/6}$ and consequently the angular deflection is $\delta\theta \sim \delta h/r \sim \lambda r_0^{-5/6} r^{-1/6}$. If the turbulence is at altitude D (with typical value $D \sim 5\text{km}$) then, in the geometric optics limit, the amplification A is on the order of

$$A(r) \sim \frac{D\delta\theta}{r} \sim D\lambda r_0^{-5/6} r^{-7/6} \quad (9.40)$$

with $A \simeq 1$ indicating the onset of caustics and multi-path propagation. However, for $\lambda \simeq 5 \times 10^{-7}\text{m}$ and $D \simeq 5\text{km}$, the Fresnel length is $r_f \simeq 5\text{cm}$ which is larger than the scale for which geometric optics gives $A = 1$ since

$$\frac{r(A=1)}{r_f} \simeq \left(\frac{r_f}{r_0}\right)^{5/7} \quad (9.41)$$

which is considerably smaller than unity for the parameters here. This very rough order of magnitude conclusion is supported by more accurate calculations (e.g. Roddier and co-workers) which show

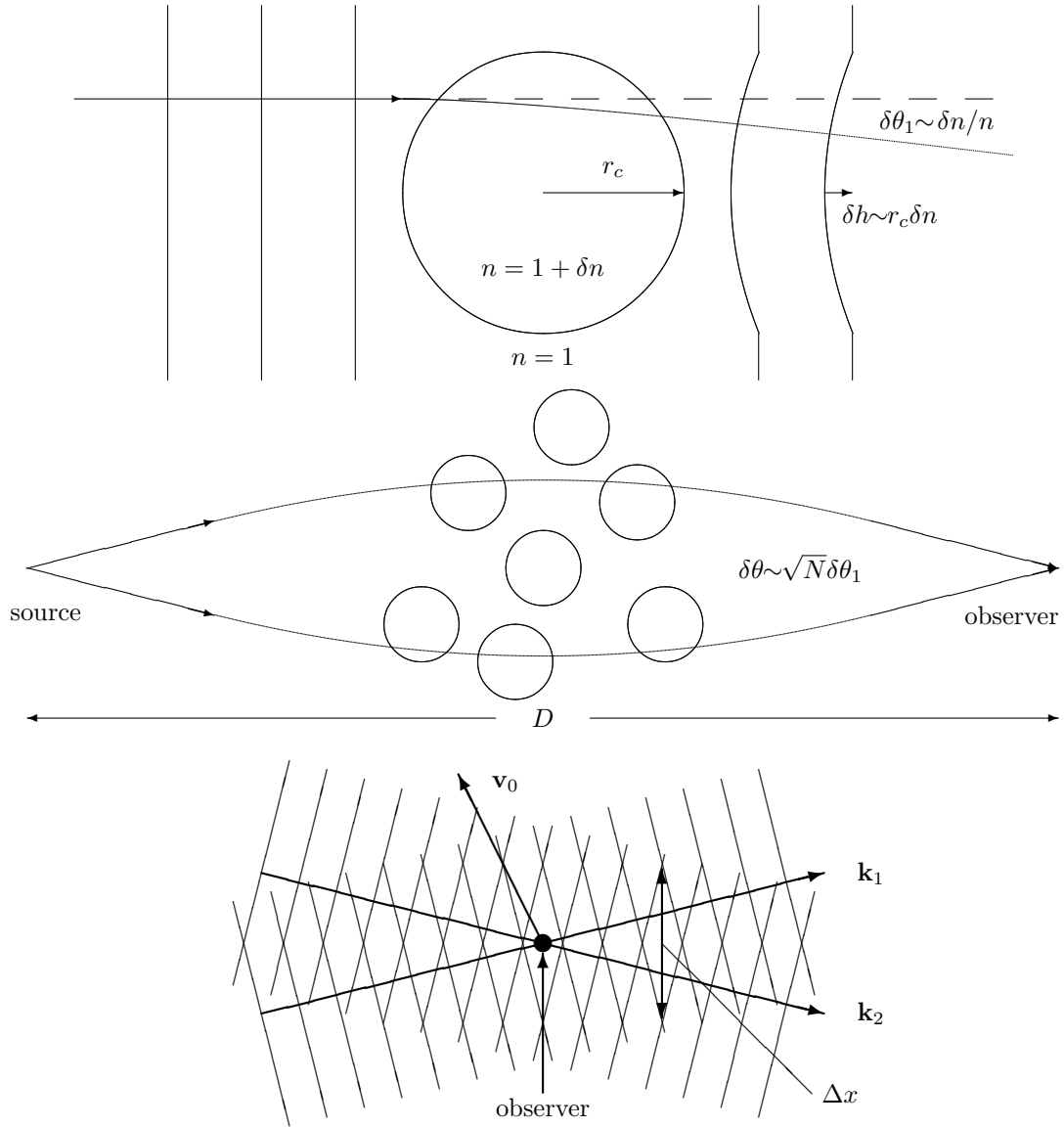


Figure 9.10: Upper panel illustrates the retardation of wavefronts δh as they pass through an overdense cloud with refractive index perturbation δn . An example ray is also shown. The middle panel illustrated the deflection of rays by a collection of randomly placed clouds. In this case the positions have contrived to deflect the rays so that the observer at right sees two images of the source. The lower panel shows a blow up of the region around the observer. Wavefronts of the two interfering signals are shown, along with the observer velocity vector and the transverse distance over which the signal from the source is modulated.

that under these ‘good seeing’ conditions the amplification should be $\lesssim 10\%$, and these theoretical results are supported by measurements of the *scintillation index*. In bad seeing, the Fried length is considerably smaller, and the scintillation can then be substantial.

9.7 Transition to Geometric Optics

Consider a simple single-element reflecting telescope whose otherwise perfect mirror suffers from a smooth *aberration* or *figure error* $h(\mathbf{r})$ as indicated in figure 9.11.

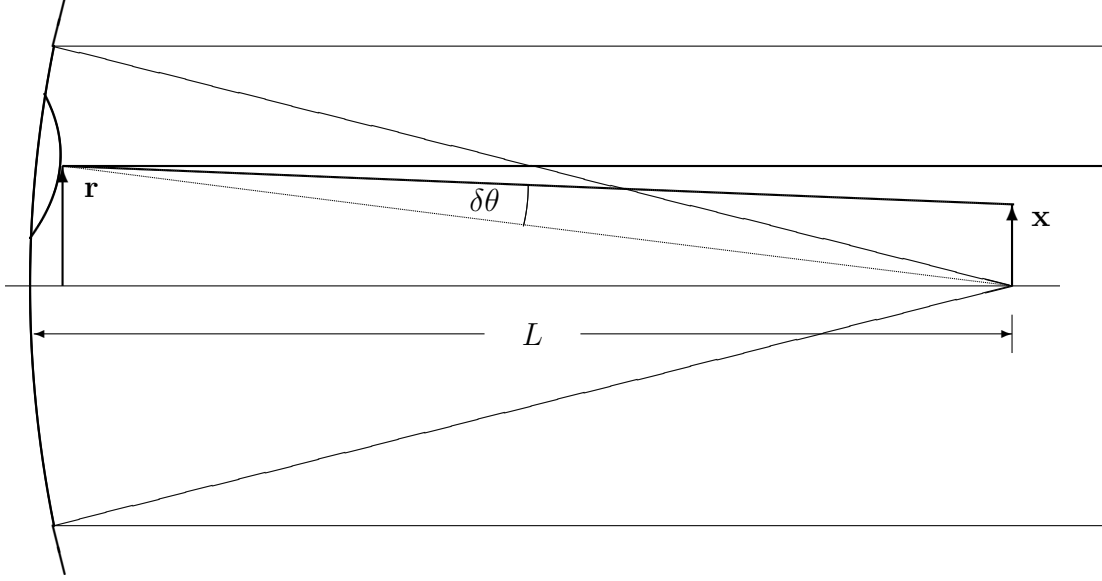


Figure 9.11: In the geometric optics limit, rays reflected off a mirror with a slight aberration suffer a small anomalous deflection $\delta\theta$ given by the gradient of the *wavefront* deformation $h(\mathbf{r})$ (so the height of the mirror deviation is $h(\mathbf{r})/2$) and arrive at the focal plane at $\mathbf{x}(\mathbf{r}) = L\nabla h(\mathbf{r})$. Caustic surfaces intersect the focal plane at the turning points of the deflection, where $|\partial^2 h / \partial r_i \partial r_j|$ vanish.

According to geometric optics, waves from a distant on-axis source reflecting off the mirror at \mathbf{r} will suffer an anomalous deflection

$$\delta\theta(\mathbf{r}) = \nabla h(\mathbf{r}) \quad (9.42)$$

and reach the focal plane at a displaced location

$$\mathbf{x}(\mathbf{r}) = L\delta\theta(\mathbf{r}) = L\nabla h(\mathbf{r}) \quad (9.43)$$

with L the focal length, and the energy flux of these rays is

$$F(\mathbf{x}) \propto \left| \frac{\partial x_i}{\partial r_j} \right|^{-1} \propto \left| \frac{\partial^2 h}{\partial r_i \partial r_j} \right|^{-1}. \quad (9.44)$$

This results in the usual caustic surfaces at points in the focal plane corresponding to turning points of the deflection.

For long wavelengths, on the other hand, the width of the PSF is $L\lambda/D \gg L\nabla h$ and the aberration is negligible.

Let us analyze this transition from wave-optics to the geometric-optics limit from the point of view of diffraction theory. The field amplitude at point \mathbf{x} on the focal plane is

$$f(\mathbf{x}) = \int d^2r A(\mathbf{r}) e^{2\pi i(h(\mathbf{r}) - \mathbf{x} \cdot \mathbf{r}/L)/\lambda} = \int d^2r A(\mathbf{r}) e^{i\psi(\mathbf{r};\mathbf{x})} \quad (9.45)$$

with $A(\mathbf{r})$ the aperture or pupil function and $\psi(\mathbf{r};\mathbf{x}) = 2\pi(h(\mathbf{r}) - \mathbf{x} \cdot \mathbf{r}/L)/\lambda$.

As the amplitude of the phase fluctuations due to the aberration become large compared to unity we expect that the greatest contribution to the field amplitude will come from points in the pupil plane where the phase ψ is nearly stationary. That is, at extrema of $\psi(r)$. These are points \mathbf{r}_0 where the gradient of the aberration function $h(\mathbf{r})$ happens to coincide with the gradient of the planar function $\mathbf{x} \cdot \mathbf{r}/L$ or where

$$(\nabla h)_{\mathbf{r}_0} = \mathbf{x}/L. \quad (9.46)$$

These are just the reflection points of geometric optics.

In the vicinity of such a point \mathbf{r}_0 we have, for the phase error,

$$\delta\psi = \frac{2\pi}{\lambda} \left(h(\mathbf{r}_0) + \frac{1}{2} r'_i r'_j \frac{\partial^2 h(\mathbf{r}_0)}{\partial r_i \partial r_j} + \dots \right) \quad (9.47)$$

with $\mathbf{r}' \equiv \mathbf{r} - \mathbf{r}_0$. Changing integration variable in the amplitude integral to \mathbf{r}' and dropping the prime gives

$$f(\mathbf{x}) \simeq e^{i2\pi h(\mathbf{r}_0)/\lambda} \int d^2r' A(\mathbf{r}_0 + \mathbf{r}') e^{i\pi h_{ij}(\mathbf{r}_0) r'_i r'_j / \lambda} \quad (9.48)$$

with $h_{ij} \equiv \partial^2 h / \partial r_i \partial r_j$. Inspecting the exponential factor in the integral we expect the dominant contribution to come from a region of size $\delta r \lesssim \sqrt{\lambda/h''}$. This ‘Fresnel zone’ becomes smaller with decreasing λ , so for most points $\mathbf{r}_0(\mathbf{x})$ we can neglect the limitation of the aperture function $A(\mathbf{r})$. Transforming to a rotated frame in which h_{ij} is diagonal,

$$h_{ij} = \begin{bmatrix} h_{xx} & 0 \\ 0 & h_{yy} \end{bmatrix} \quad (9.49)$$

and the field amplitude becomes

$$f(\mathbf{x}) = e^{i2\pi h(\mathbf{r}_0)/\lambda} \int dx e^{i\pi h_{xx} x^2 / \lambda} \int dy e^{i\pi h_{yy} y^2 / \lambda}. \quad (9.50)$$

Now, on dimensional grounds the x -integral has value $\sqrt{\lambda/\pi h_{xx}}$ times some factor of order unity, and similarly for the y -integral, so on squaring the wave amplitude we obtain the energy flux

$$|f(\mathbf{x})|^2 \propto 1/(h_{xx} h_{yy}) \quad (9.51)$$

but this is just the inverse of the determinant of the matrix h_{ij} , which is a rotational invariant, so in general we have

$$|f(\mathbf{x})|^2 \propto \left| \frac{\partial^2 h}{\partial r_i \partial r_j} \right|^{-1} \quad (9.52)$$

again in agreement with the geometric optics result (9.44).

The argument above is somewhat misleading in that we have simply squared the amplitude arising from a single Fresnel zone, whereas in general, we expect there to be a superposition of the amplitudes from several zones, each of which will have a different phase factor $e^{i2\pi h(\mathbf{r}_0)/\lambda}$, and this results in an illumination pattern on the focal plane which with a periodic pattern, with alternating zones of constructive and destructive interference, as illustrated in figure 9.12.

9.8 Problems

9.8.1 Diffraction.

You wish to construct a pinhole camera from a box of side $L = 10\text{cm}$ to take a photograph at a wavelength $\lambda = 4 \times 10^{-5}\text{cm}$. Estimate the diameter of the pinhole required to obtain the sharpest image.

9.8.2 Fraunhofer

A perfectly absorbing disk of radius R lies in a collimated beam of monochromatic radiation of wavelength λ and absorbs energy at a rate of 1 erg/s. What is the net energy flux in the radiation scattered out of the original beam direction by the disk according to Fraunhofer diffraction theory.

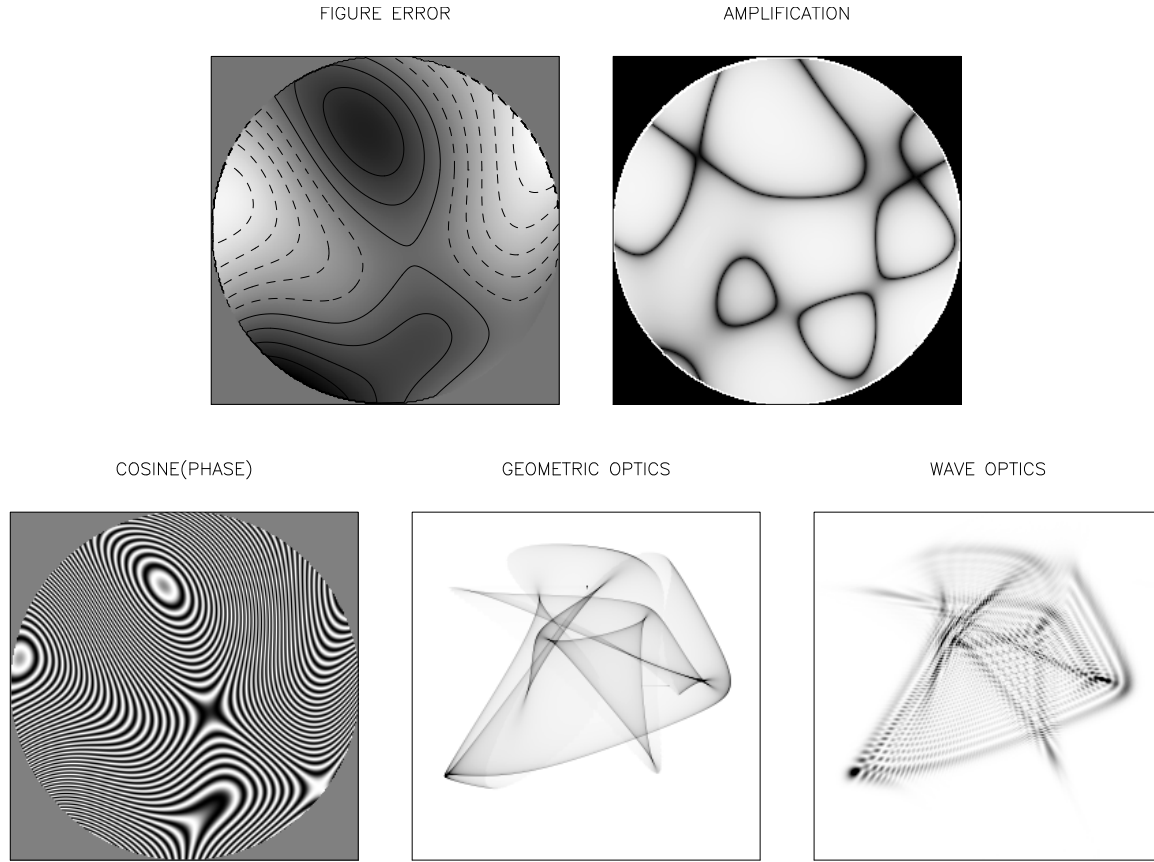


Figure 9.12: Geometric vs wave optics. Upper left panel shows an example of a quasi-random figure error $h(\mathbf{r})$. This is a sample of smoothed Gaussian random noise windowed by the circular aperture function. Upper right panel shows the amplification $A = |\partial^2 h / \partial r_i \partial r_j|^{-1}$ according to geometric optics. Lower left panel shows the cosine of the phase $\cos(\psi(\mathbf{r}))$ with $\psi(\mathbf{r}) = 2\pi h(\mathbf{r})/\lambda$. The complex field amplitude at a point \mathbf{x} in the focal plane is equal to the Fourier transform of $\cos(\psi) + i \sin(\psi)$. Lower right panel shows the PSF computed according to diffraction theory by squaring the field amplitude while the lower center panel shows the PSF for this figure error in the geometric optics limit. The sharp, well-defined features in the wave-optics PSF arise from regions in the pupil plane where $\cos(\psi)$ has a quasi-monochromatic variation over relatively extended regions, and are associated with the ‘critical-curves’ seen in the amplification pattern. These features correspond to the geometrical optical caustics. The periodic oscillations in the wave-optics PSF arise from interference between radiation arising from distinct zones on the pupil plane.

9.8.3 Telescope PSF from Fresnel Integral

Consider a large ground based telescope of aperture A and focal length D observing at wavelength λ . Assume that the optics have been configured to give diffraction limited performance. This means that planar wavefronts from a distant source on axis interfere constructively at the origin $\vec{x} = 0$ of the focal plane with negligible phase shifts (or equivalently that the optical path length from $\vec{x} = 0$ to any point \vec{r} on the entrance pupil plane is constant to a precision better than λ). Similarly, wavefronts from a source an angle $\vec{\theta}$ off axis will interfere constructively at $\vec{x} = D\vec{\theta}$.

- a. Show from the foregoing considerations that the optical path length from \vec{x} to \vec{r} is given by

$$l = \text{constant} + \vec{x} \cdot \vec{r}/D \quad (9.53)$$

- b. At any instant, the wave front from a distant point source will have been distorted by fluctua-

tions in the density of the turbulent atmosphere through which it has passed. Let the vertical displacement of the wavefront in the entrance pupil be $h(\vec{r})$. Use Fresnel diffraction theory to show that the amplitude for the radiation field on the focal plane is

$$a(\vec{x}) \propto \int d^2r C(\vec{r}) e^{2\pi i \vec{x} \cdot \vec{r} / D\lambda} \quad (9.54)$$

where $C(\vec{r}) = e^{i\varphi(\vec{r})}$ where the phase shift is $\varphi(\vec{r}) = 2\pi h(\vec{r})/\lambda$. Thus show that for a sufficiently long integration time the intensity pattern on the focal plane (the ‘point spread function’ or psf) is

$$g(\vec{x}) = \langle |a(\vec{x})|^2 \rangle = \int d^2r \xi_C(\vec{r}) e^{2\pi i \vec{x} \cdot \vec{r} / D\lambda} \quad (9.55)$$

where $\xi_C(\vec{r}) \equiv \langle C(\vec{r}' + \vec{r}) C^*(\vec{r}') \rangle$. Show that the fourier transform of the psf (the ‘optical transfer function’) is therefore given by

$$\tilde{g}(\vec{k}) \equiv \int d^2x g(\vec{x}) e^{i\vec{k} \cdot \vec{x}} = \xi_C(\vec{k} D\lambda / 2\pi) \quad (9.56)$$

- c. Assuming the distortion of the wavefront $h(\vec{r})$ takes the form of a 2-dimensional gaussian random field, one can show that $\xi_C(\vec{r}) = \langle e^{i(\varphi_1 - \varphi_2)} \rangle = e^{-S_\varphi(\vec{r})/2}$ where $S_\varphi(\vec{r}) \equiv \langle (\varphi(\vec{r}') - \varphi(\vec{r}' + \vec{r}))^2 \rangle$ is the ‘structure function’. For fully developed Kolmogorov turbulence one has $S_\varphi(\vec{r})$ is a power law $S_\varphi(r) = (r/r_0)^{5/3}$ where r_0 is the ‘Fried length’ over which the rms phase change is unity. Compute the OTF in this case, and estimate the (angular) width of the psf in terms of r_0 , λ .
- d. Assuming that the amplitude of the wavefront deformation $h(\vec{r})$ is independent of wavelength show that the resolution scales as the 1/5 power of wavelength.
- e. Derive the relation $\langle e^{i(\varphi_1 - \varphi_2)} \rangle = e^{-S_\varphi(\vec{r})/2}$ for a gaussian random field $\varphi(\vec{r})$.

Chapter 10

Radiation from Moving Charges

10.1 Electromagnetic Potentials

Maxwell's equations (7.4) involve the six variables $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{B}(\mathbf{r}, t)$. They can be reformulated in terms of 4 potentials $\phi(\mathbf{r}, t)$, $\mathbf{A}(\mathbf{r}, t)$.

The second of Maxwell's equations $\nabla \cdot \mathbf{B} = 0$ tells us that \mathbf{B} is the curl of some vector \mathbf{A} :

$$\mathbf{B} = \nabla \times \mathbf{A}. \quad (10.1)$$

Using this in Faraday's law $\nabla \times \mathbf{E} = -\frac{1}{c} \partial \mathbf{B} / \partial t$ (M3) tells us that

$$\nabla \times \left(\mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} \right) = 0 \quad (10.2)$$

so the quantity in parentheses is the gradient of some scalar function $-\phi$, or

$$\mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} - \nabla \phi \quad (10.3)$$

These potentials are not unique, since if we make the *gauge transformation*

$$\begin{aligned} \mathbf{A} &\rightarrow \mathbf{A}' = \mathbf{A} + \nabla \psi \\ \phi &\rightarrow \phi' = \phi - \frac{1}{c} \dot{\psi} \end{aligned} \quad (10.4)$$

where $\psi(\mathbf{r}, t)$ is an arbitrary function this leaves the physical fields \mathbf{E} , \mathbf{B} unchanged.

Using (10.3), Gauss' law $\nabla^2 \phi = 4\pi\rho$ becomes

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} + \frac{1}{c} \frac{\partial (\nabla \cdot \mathbf{A} + \dot{\phi}/c)}{\partial t} = -4\pi\rho \quad (10.5)$$

and Ampere's law (M4) similarly becomes

$$\nabla^2 \mathbf{A} - \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} + \nabla (\nabla \cdot \mathbf{A} + \dot{\phi}/c) = -4\pi\mathbf{j}/c \quad (10.6)$$

The terms involving the quantity $\zeta \equiv \nabla \cdot \mathbf{A} + \dot{\phi}/c$ can be removed by a suitable gauge transformation, since, under the transformation (10.4)

$$\zeta \rightarrow \zeta' = \zeta + \square \psi \quad (10.7)$$

where we have defined the *d'Alembertian* or *wave operator*

$$\square \equiv \nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \quad (10.8)$$

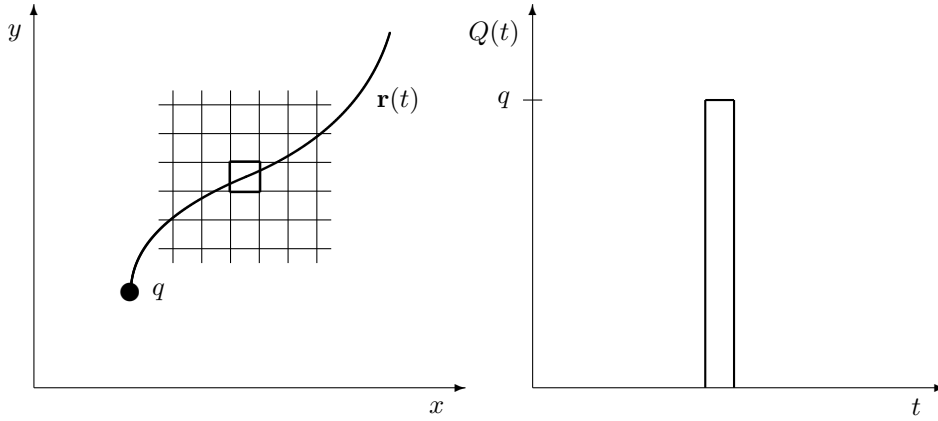


Figure 10.1: Illustration of the source term $Q(t)$ for the retarded potential. We carve space up into cells (left panel) and then solve for the field due to a single cell (shown highlighted). For a single moving charge, for example, the source term $Q(t)$ would be a short pulse (right). Having solved for the field from once such microscopic cell we then invoke the linearity of the field equations to add together the solutions for all of the cells.

Since $\psi(\mathbf{r}, t)$ is arbitrary, it can be chosen to make $\zeta = \square\psi$ vanish:

$$\nabla \cdot \mathbf{A} + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0 \quad (10.9)$$

which is called the *Lorentz gauge condition*.

With this choice of gauge, Maxwell's equations take the very simple form:

$$\begin{aligned} \square \mathbf{A} &= -4\pi \mathbf{j}/c \\ \square \phi &= -4\pi \rho \end{aligned} \quad (10.10)$$

In the absence of charges both \mathbf{A} and ϕ obey the wave equation, and the terms on the RHS are source terms.

See appendix E for a review of invariance of electromagnetism.

10.2 Retarded Potentials

The equations (10.10) are linear in the fields and in the charges. This means that the solutions for a superposition of source terms can be obtained by summing the solutions for the various components. In particular, we can describe a general charge/current distribution by carving space up into a grid of infinitesimal cubical cells and giving the charge within, and current passing through, each cell (see figure 10.1).

Consider one infinitesimal cell, which we can place at the origin of our coordinates. Let the charge within the cell be $Q = \sum q$ and the current be $\mathbf{J} = \sum q\mathbf{v}$. These will be, in general, functions of time, so Maxwell's equations are

$$\begin{aligned} \square \mathbf{A} &= -4\pi \mathbf{J}(t)\delta(\mathbf{r})/c \\ \square \phi &= -4\pi Q(t)\delta(\mathbf{r}) \end{aligned} \quad (10.11)$$

The four source terms here are clearly spherically symmetric, so $\phi(\mathbf{r}, t) = \phi(r, t)$ and $A_x(\mathbf{r}, t) = A_x(r, t)$ etc., and the solutions will share this symmetry. To find the solutions, we use the Laplacian for a spherically symmetric function

$$\nabla^2 f(r) = \frac{1}{r^2} \frac{d(r^2 df/dr)}{dr}. \quad (10.12)$$

(Problem: show this.) Applying this to the potential, and making the change of variable $\phi(r, t) = \chi(r, t)/r$ we find that for $\mathbf{r} \neq 0$ the function $\chi(r, t)$ satisfies the 1-dimensional wave equation

$$\chi'' - \ddot{\chi}/c^2 = 0 \quad (10.13)$$

where $\chi'' \equiv \partial^2 \chi / \partial r^2$. This has the general solution

$$\chi(r, t) = \chi_1(t - r/c) + \chi_2(t + r/c) \quad (10.14)$$

where χ_1, χ_2 are two arbitrary functions which must be chosen to satisfy appropriate boundary conditions. The first term here describes an outward propagating wave, which is physically reasonable, so we discard the latter.

Finally, we need to find $\chi_1(t)$ which gives the correct Coulombic behavior in the immediate vicinity of the source: $\phi(r, t) \rightarrow Q(t)/r$ as $r \rightarrow 0$. This requires $\chi_1(t) = Q(t)$. Similarly, requiring that the magnetic potential tend to the magneto-static form gives the $\mathbf{A}(r, t) \rightarrow \mathbf{J}(t)/cr$ and the solutions to Maxwell's equations are

$$\begin{aligned} \mathbf{A}(\mathbf{r}, t) &= \mathbf{J}(t - r/c)/cr \\ \phi(\mathbf{r}, t) &= Q(t - r/c)/r \end{aligned} \quad (10.15)$$

Problem: show that $\mathbf{A}(\mathbf{r}) = \mathbf{J}/r$ reproduces the Biot-Savart law with $\mathbf{B} = \nabla \times \mathbf{A}$.

These are the solutions for a single infinitesimal source cell located at the origin. Summing over all cells give the potentials at some arbitrary 'field point' \mathbf{r} as 3-dimensional integrals

$$\begin{aligned} \mathbf{A}(\mathbf{r}, t) &= \frac{1}{c} \int d^3 r' \frac{\mathbf{j}(\mathbf{r}', t - |\mathbf{r} - \mathbf{r}'|/c)}{|\mathbf{r} - \mathbf{r}'|} \\ \phi(\mathbf{r}, t) &= \int d^3 r' \frac{\rho(\mathbf{r}', t - |\mathbf{r} - \mathbf{r}'|/c)}{|\mathbf{r} - \mathbf{r}'|} \end{aligned} \quad (10.16)$$

These are called the *retarded potentials*. They explicitly give the potentials ϕ, \mathbf{A} produced by an arbitrary charge distribution, and, with (10.1), (10.3) the fields \mathbf{E}, \mathbf{B} .

The potentials (10.16) are just like the electro- and magneto-static potentials, but where the source is evaluated at the retarded time.

To check that these solutions do in fact obey the Lorentz gauge conditions, rewrite these with a change of integration variable to $\mathbf{r}'' = \mathbf{r} - \mathbf{r}'$ so

$$\nabla \cdot \mathbf{A} = \frac{\nabla}{c} \int d^3 r'' \frac{\mathbf{j}(\mathbf{r} - \mathbf{r}'', t - r'')}{r''} = \frac{1}{c} \int d^3 r'' \frac{(\nabla \cdot \mathbf{j})_{\mathbf{r} - \mathbf{r}'', t - r''}}{r''} \quad (10.17)$$

and similarly $\dot{\phi} = \int d^3 r'' \frac{\dot{\rho}(\mathbf{r} - \mathbf{r}'', t - r'')}{r''}$ and so

$$\nabla \cdot \mathbf{A} + \dot{\phi}/c = \frac{1}{c} \int d^3 r'' \frac{(\nabla \cdot \mathbf{j} + \dot{\rho})_{\mathbf{r} - \mathbf{r}'', t - r''}}{r''} \quad (10.18)$$

which vanishes due to charge conservation.

One can add to these any solution of the homogeneous wave equations, to describe, for instance, radiation from external sources.

10.3 Lienard-Wiechart Potentials

The *Lienard-Wiechart potentials* are the specialization of (10.16) to the case of a single particle of charge q moving along a path $\mathbf{r}_0(t)$ with velocity $\mathbf{u} = \dot{\mathbf{r}}_0$. To obtain these, we first rewrite (10.16) as 4-dimensional space-time integrals by introducing a δ -function. For the scalar potential this gives

$$\phi(\mathbf{r}, t) = \int d^3 r' \int dt' \frac{\rho(\mathbf{r}', t') \delta(t' - t + |\mathbf{r} - \mathbf{r}'|/c)}{|\mathbf{r} - \mathbf{r}'|}. \quad (10.19)$$

For a point charge, the charge density is

$$\rho(\mathbf{r}', t') = q \delta(\mathbf{r}' - \mathbf{r}_0(t')) \quad (10.20)$$

which allows us to perform the spatial integration to give

$$\phi(\mathbf{r}, t) = q \int dt' \frac{\delta(t' - t + |\mathbf{r} - \mathbf{r}_0(t')|/c)}{|\mathbf{r} - \mathbf{r}_0(t')|} \quad (10.21)$$

or, with $\mathbf{R}(t') \equiv \mathbf{r} - \mathbf{r}_0(t')$ and $R = |\mathbf{R}|$, we can write this as

$$\phi(\mathbf{r}, t) = q \int dt' \frac{\delta(t' - t + R(t')/c)}{R(t')}. \quad (10.22)$$

This equation says that the potential seen by an observer at point \mathbf{r} and time t only ‘knows’ about the state of the particle at the specific instant t' on its trajectory. That instant is when $c(t - t')$ happens to coincide with the distance to the particle $R(t')$. This is the point in space-time at which the particle’s trajectory intersects the *past light cone* of the observer (see figure 10.2). It should be obvious from this figure that, provided the particle’s velocity does not exceed the speed of light, there can only be one such point. This time, the solution of the equation $t' + R(t')/c = t$, is called the *retarded time* $t' = t_{\text{ret}}(t)$.

The usual procedure is now to formally manipulate the δ -function using the property of the Dirac δ -function that $\delta(F(t') - t) = \delta(t' - F^{-1}(t))/F'(t')$ where $F(t)$ is an arbitrary continuous function, F^{-1} is the inverse function, and F' denotes the derivative of F . (To prove this simply perform a Taylor expansion of $F(t')$ around $t' = t_{\text{ret}}(t) \equiv F^{-1}(t)$.) This gives an integral over dt' involving $\delta(t' - t_{\text{ret}}(t))$ which is then easily performed. Here we will take a more pedestrian, though entirely equivalent, approach. Let’s calculate the potential at the observer’s location, averaged over a small (eventually infinitesimal) time interval $0 < t < \tau$:

$$\langle \phi \rangle = \frac{1}{\tau} \int_0^\tau dt \phi(\mathbf{r}, t) = \frac{q}{\tau} \int \frac{dt'}{R(t')} \int_0^\tau dt \delta(t' - t + R(t')/c). \quad (10.23)$$

The second integral has value unity if $0 < t' + R(t') < \tau$, and zero otherwise (since in the latter case the δ -function falls outside of the domain of integration). The time averaged potential is therefore

$$\langle \phi \rangle = \frac{q}{\tau} \int_{t_{\text{ret}}(0)}^{t_{\text{ret}}(\tau)} \frac{dt'}{R(t')} \simeq \frac{q}{R(t_{\text{ret}})} \frac{\Delta t'}{\tau} \quad (10.24)$$

with $\Delta t' = t_{\text{ret}}(\tau) - t_{\text{ret}}(0)$, which is the coordinate time it takes the particle to pass from the past light cone of the point $(\mathbf{r}, 0)$ to that of (\mathbf{r}, τ) . This approximate equality becomes exact as the averaging interval $\tau \rightarrow 0$.

We need therefore to calculate the ratio $\Delta t'/\tau$, since the actual potential is just the regular coulombic potential $\phi = q/R$ times this factor. Inspection of figure 10.2 shows that for a stationary charge $\Delta t'/\tau = 1$; for a charge moving away from the observer at $v = c$, $\Delta t'/\tau = 1/2$ while for an observer moving towards the observer, $\Delta t'/\tau \rightarrow \infty$ as $v \rightarrow c$. To obtain the general result, simply insert $t_{\text{ret}}(\tau) = t_{\text{ret}}(0) + \Delta t'$ into the $t_{\text{ret}}(\tau) + R(t_{\text{ret}}(\tau))/c = \tau$ (which is the equation defining what we mean by the retarded time) to give

$$t_{\text{ret}}(0) + \Delta t' + R(t_{\text{ret}}(0) + \Delta t')/c = \tau. \quad (10.25)$$

Doing a Taylor expansion: $R(t_{\text{ret}}(0) + \Delta t') \rightarrow R(t_{\text{ret}}(0)) + \Delta t' dR/dt'$ and using $t_{\text{ret}}(0) + R(t_{\text{ret}}(0))/c = 0$ yields

$$\Delta t' = \left(1 + \frac{1}{c} \frac{dR}{dt'} \right)^{-1} \tau. \quad (10.26)$$

Now since by definition $R(t') = \sqrt{(\mathbf{r} - \mathbf{r}_0(t')) \cdot (\mathbf{r} - \mathbf{r}_0(t'))}$, the derivative here is $dR(t')/dt' = -\dot{\mathbf{R}}(t') \cdot \dot{\mathbf{r}}_0(t') = -\dot{\mathbf{R}}(t') \cdot \mathbf{v}(t')$, and we have

$$\frac{\Delta t'}{\tau} = \frac{1}{1 - \dot{\mathbf{R}}(t') \cdot \mathbf{v}(t')/c}. \quad (10.27)$$

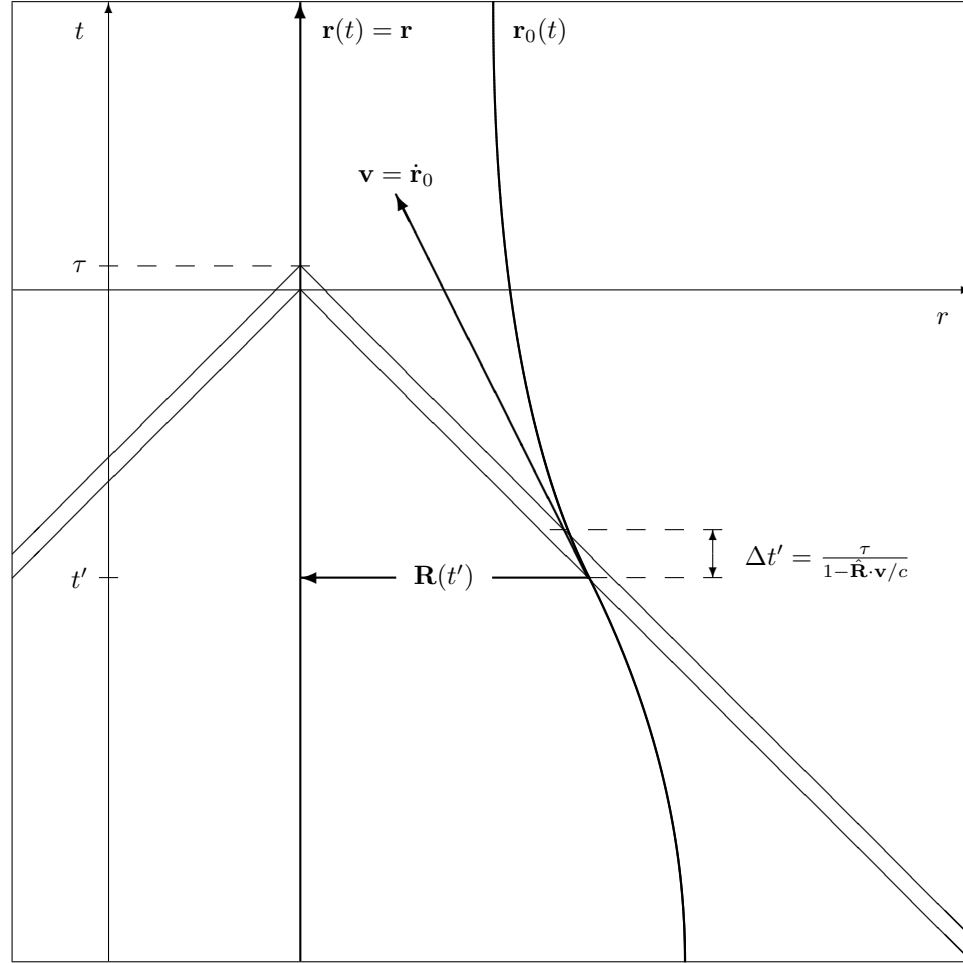


Figure 10.2: Space-time diagram used to derive the Lienard-Wiechart potentials. The bold vertical line represents the observer — taken here to be stationary — and the heavy curve represents the moving charged particle. The lines at 45 degrees are the past light cones of the points on the observer's world line $(\mathbf{r}, t = 0)$ and $(\mathbf{r}, t = \tau)$. The charged particle intercepts the past light cone of the point (\mathbf{r}, t) only once, at a *retarded time* t' such that $t' + R(t') = t$ and at distance $R(t')$. We show in the text that the potential seen by the observer is just equal to the Coulomb potential $\phi = q/R(t')$ times the factor $\Delta t'/\tau$, with $\Delta t'$ the (coordinate) time taken to pass between the two past light cones. This factor ranges from 1/2, for a particle moving very rapidly ($v \simeq c$) away from the observer, to arbitrarily large for a relativistic particle moving directly towards the observer. The general formula for this ratio is indicated.

Clearly, this agrees with the three special cases above. Finally then, for $\tau \rightarrow 0$ the average field tends to the instantaneous field $\langle \phi \rangle \rightarrow \phi(\mathbf{r}, t = 0)$. There is nothing special about the choice of $t = 0$ and we have the general formula for the LW scalar potential

$$\phi(\mathbf{r}, t) = q \left[\frac{1}{(1 - \hat{\mathbf{R}} \cdot \mathbf{v}/c)R} \right] \quad (10.28)$$

where the quantity in square brackets is to be evaluated at the retarded time.

Identical reasoning can be applied to the vector potential to give

$$\mathbf{A}(\mathbf{r}, t) = \frac{q}{c} \left[\frac{\mathbf{v}}{(1 - \hat{\mathbf{R}} \cdot \mathbf{v}/c)R} \right] \quad (10.29)$$

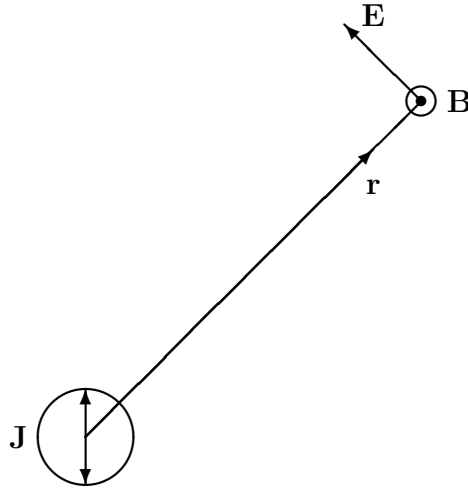


Figure 10.3: The geometry for calculation of dipole radiation from a small region of space (circle) containing charges such that the net current \mathbf{J} is parallel to the z -axis. The ‘field point’ at large distance \mathbf{r} is indicated. In the calculation it is assumed that the charge distribution is effectively static over the time it takes for radiation to cross the system.

A key feature of the L-W potentials (10.28), (10.29) is that the potentials, and hence also the fields, become very large when $1 - \hat{\mathbf{R}} \cdot \mathbf{v}/c$ is very small. This happens when a highly energetic particle is moving almost directly towards the observer, as one would expect from relativistic beaming. The potentials also provide a useful starting point to discuss such issues as the transition from coulomb potential to radiation at large distances. We will use them in the treatment of synchrotron and Cerenkov radiation.

10.4 Dipole Radiation

Imagine we have some oscillating charge distribution within a small region of space, with the net current aligned with the z -axis as illustrated in figure 10.3. The potential and field solutions will approximate those for a single microscopic cell (10.15). If the current is

$$\mathbf{J}(t) = J_0 \hat{\mathbf{z}} e^{i\omega t} \quad (10.30)$$

so the magnetic potential is

$$\mathbf{A}(\mathbf{r}, t) = J_0 \hat{\mathbf{z}} e^{i\omega(t-r/c)}/cr \quad (10.31)$$

The magnetic field is given by taking the curl: $\mathbf{B} = \nabla \times \mathbf{A}$. For a static field ($\omega \rightarrow 0$), the spatial derivatives act only on the $1/r$ factor, giving a field falling as $B \propto 1/r^2$ (the *Biot-Savart* law). With a time varying current, the spatial derivatives also act on the complex exponential, so $\partial A \sim (i\omega/c)J_0/r$ giving only a $B \propto 1/r$ fall-off. At sufficiently large distances the $1/r$ component comes to dominate, and this is the radiation field, with energy density $\propto E^2 = B^2 \propto 1/r^2$.

More precisely, we have $\partial_x e^{-i\omega r/c} = (-i\omega x/cr) e^{-i\omega r/c}$ and hence

$$\mathbf{B} = \nabla \times \mathbf{A} = \frac{J_0 e^{i\omega t}}{cr} \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial_x & \partial_y & \partial_z \\ 0 & 0 & e^{-i\omega r/c} \end{vmatrix} = \frac{-i\omega J_0 e^{i\omega t}}{c^2} \left[\frac{\hat{\mathbf{x}} y}{r^2} - \frac{\hat{\mathbf{y}} x}{r^2} \right]. \quad (10.32)$$

- The magnetic field is perpendicular to the current ($\mathbf{B} \cdot \mathbf{z} = 0$), and also to \mathbf{r} , the direction of propagation (as required for transverse waves).
- The amplitude of the magnetic field is greatest for points with $z = 0$ — ie around the equator — and is zero for points along the z -axis. The energy density is $\propto |\mathbf{B}|^2$ and is proportional to $\sin^2 \theta$ — with θ the polar angle.

- The electric field is trickier to calculate this way, but we know that it must be perpendicular to \mathbf{B} and to the line of sight \mathbf{r} , and equal in magnitude. It is oriented parallel to the projection of the current vector on the sky.
- The pre-factor $i\omega$ tells us that the field is proportional to the rate of change of the current. Since the current is proportional to the charge velocity, this means that the radiation field is proportional to the charge acceleration.
- The current is given by $\mathbf{J} = \sum q\mathbf{v} = \partial(\sum q\mathbf{r})/\partial t$ so the field is proportional to the second time derivative of the *dipole moment* $\mathbf{d} = \sum q\mathbf{r}$.

For an arbitrary small system of charges — where the current need not align with the z -axis — the spectral decomposition of the radiation field is then

$$\mathbf{B}_\omega(\mathbf{r}) = \frac{\ddot{\mathbf{d}}_\omega \times \mathbf{r}}{c^2 r^2}. \quad (10.33)$$

10.5 Larmor's Formula

If the current is due to a single charge, we have for the Poynting vector

$$\mathbf{S} = \frac{c}{8\pi}(E^2 + B^2) = \frac{c}{4\pi}B^2 = \frac{j^2}{4\pi r^2 c^3} \sin^2 \theta = \frac{q^2 \dot{u}^2}{4\pi r^2 c^3} \sin^2 \theta \quad (10.34)$$

with \mathbf{u} the charge velocity (we have derived this for a single harmonic component, but it is true in general by Parseval's theorem).

The energy radiated into direction $d\Omega$ per unit time is $Sr^2 d\Omega$ hence

$$\frac{dW}{dt d\Omega} = \frac{q^2 \dot{u}^2}{4\pi c^3} \sin^2 \theta \quad (10.35)$$

and integrating over direction gives the total power

$$P = \frac{dW}{dt} = \frac{q^2 \dot{u}^2}{4\pi c^3} 2\pi \int_{-1}^1 d\mu (1 - \mu^2) \quad (10.36)$$

the integral is elementary and leads to *Larmor's formula* for the power radiated by an accelerated charge:

$$P = \frac{2q^2 \dot{u}^2}{3c^3}. \quad (10.37)$$

10.6 General Multi-pole Expansion

We showed above that a small region containing a collection of moving charges generates a radiation field proportional to the second time derivative of the dipole moment of the charge. In obtaining this we implicitly assumed that the variation of the system over the time it takes for light to cross it is negligible, so the result is valid only for sufficiently small regions and velocities. This *dipole approximation* is just one term in a more general expansion for the radiation from a collection of charges.

To derive the general multi-pole expansion we start with (10.16) and rewrite the expression for the magnetic potential as a 4-dimensional space time integral by introducing a temporal δ -function:

$$\mathbf{A}(\mathbf{r}, t) = \frac{1}{c} \int d^3 r' \int dt' \frac{\mathbf{j}(\mathbf{r}', t') \delta(t' - t + |\mathbf{r} - \mathbf{r}'|/c)}{|\mathbf{r} - \mathbf{r}'|} \quad (10.38)$$

Next, define the temporal Fourier transforms of the current and potential to be

$$\mathbf{j}_\omega(\mathbf{r}) = \int dt \mathbf{j}(\mathbf{r}, t) e^{i\omega t} \quad (10.39)$$

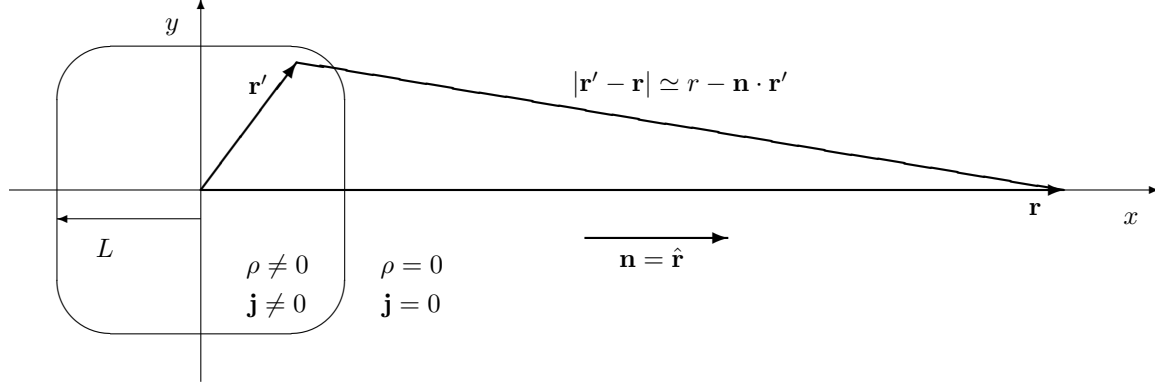


Figure 10.4: For a bounded charge distribution of size L and a distant observer at $r \gg L$ one can approximate the distance $|\mathbf{r}' - \mathbf{r}| \simeq r - \mathbf{n} \cdot \mathbf{r}'$.

and

$$\mathbf{A}_\omega(\mathbf{r}) = \int dt \mathbf{A}(\mathbf{r}, t) e^{i\omega t} \quad (10.40)$$

to obtain

$$\begin{aligned} \mathbf{A}_\omega(\mathbf{r}) &= \frac{1}{c} \int dt \int d^3 r' \int dt' \mathbf{j}(\mathbf{r}', t') \delta(t' - t + |\mathbf{r} - \mathbf{r}'|/c) e^{i\omega t} \\ &= \frac{1}{c} \int d^3 r' \frac{e^{ik|\mathbf{r} - \mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|} \int dt' \mathbf{j}(\mathbf{r}', t') e^{i\omega t'} \\ &= \frac{1}{c} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') \frac{e^{ik|\mathbf{r} - \mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|} \end{aligned} \quad (10.41)$$

The transform of the potential at some frequency ω is thus the convolution of $\mathbf{j}_\omega(\mathbf{r})$ with the circular wave function e^{ikr}/r . This is reminiscent of diffraction theory.

Now let's place the origin of coordinates within the charge distribution of extent $r' \sim L$, and place the field point at some distance $r \gg L$ as illustrated in figure 10.4. For $r \gg L$ we can replace $|\mathbf{r} - \mathbf{r}'|$ in the denominator with r and expand the distance factor in the complex exponential as $|\mathbf{r} - \mathbf{r}'| \simeq r - \mathbf{n} \cdot \mathbf{r}'$, with \mathbf{n} the unit vector in the direction of the field point, to obtain

$$\mathbf{A}_\omega(\mathbf{r}) = \frac{e^{ikr}}{cr} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') e^{-ik\mathbf{n} \cdot \mathbf{r}'} \quad (10.42)$$

and expanding the exponential in the integral we have

$$\mathbf{A}_\omega(\mathbf{r}) = \frac{e^{ikr}}{cr} \sum_{n=0}^{\infty} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') (-ik\mathbf{n} \cdot \mathbf{r}')^n \quad (10.43)$$

This is a series expansion in $k\mathbf{n} \cdot \mathbf{r}' \sim L/\lambda$. This parameter will be small (and so the series will rapidly converge) provided $L \ll \lambda$, or equivalently if the light crossing-time for the system is small compared to the period of the radiation, which will usually be the case if the charges are moving non-relativistically.

The lowest order term in the expansion is $n = 0$ for which

$$\mathbf{A}_\omega^{(0)}(\mathbf{r}) = \frac{e^{ikr}}{cr} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') = \frac{e^{ikr}}{cr} \int dt e^{i\omega t} \mathbf{d}(t) = \frac{e^{ikr}}{cr} i\omega \mathbf{d}_\omega \quad (10.44)$$

with $\mathbf{d}(t) \equiv \sum q\mathbf{r}(t)$ the dipole moment of the charge distribution.

The next term in the series is

$$\mathbf{A}_\omega^{(1)}(\mathbf{r}) = \frac{-ike^{ikr}}{cr} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') \mathbf{n} \cdot \mathbf{r}' = \frac{-ike^{ikr}}{cr} \sum q\mathbf{r}'(\mathbf{n} \cdot \mathbf{r}') \quad (10.45)$$

which involves the quadrupole moment of the charge distribution and this radiation is called quadrupole radiation.

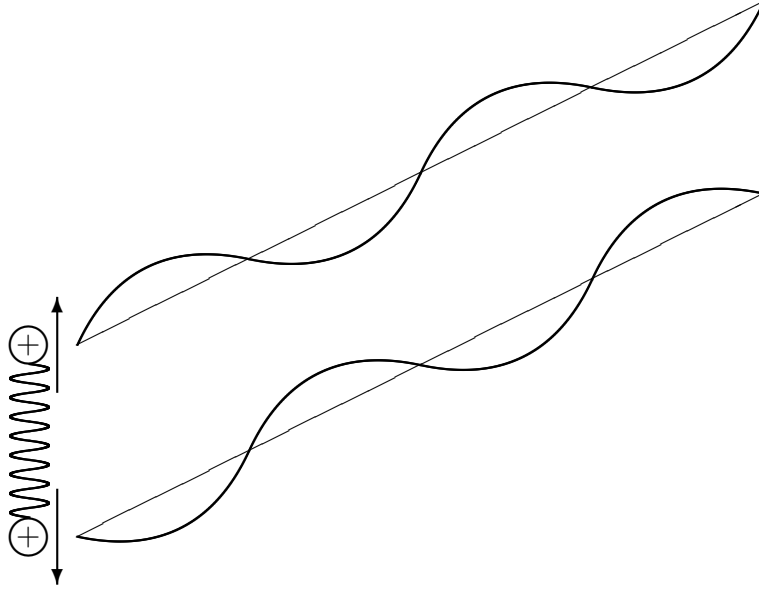


Figure 10.5: A simple antenna consists of two equal charges on a spring. The dipole moment for this system vanishes. If the wavelength is very long compared to the size of the antenna then there is very strong destructive interference of the radiation from the two charges. However, this cancellation is not perfect, and once we allow for the finite light-crossing time of the system, then distant observers will, in general, see some radiation. The strength of the field is typically smaller than the dipole field that one of these charges would create by a factor $\sim L/\lambda$. The radiation is strongest towards the poles, since the phase difference is then maximized, and zero for observers around the equator. The energy has a $\cos^2(\theta)$ or ‘quadrupolar’ dependence.

For non-relativistic charge distributions, the lowest order non-vanishing term will dominate. In many cases this is the dipole term. In some cases the dipole may vanish. For example, consider a system consisting of a spring with equal charges at each end and undergoing linear oscillation, as illustrated in figure 10.5. The net current in this system vanishes, as does the dipole moment. In the dipole approximation this system does not radiate. This is because to lowest order, the fields generated by the opposing currents cancel. However, this neglects the finite time it takes for the radiation to cross the system. In general, the field at large distance will be the sum of that from the two charges, but they will not be perfectly out of phase, and there will be some net radiation in the quadrupole term. Another example is a collision between two equal charges, for which there is also no net dipole moment, and the dominant radiation term is again the quadrupole component.

The above formulae show that a given temporal frequency of the radiation field is generated by the corresponding frequency in the relevant moment of the charge distribution. Note that these frequencies need not correspond to the frequency of motion of the charges. Consider, for example, a system consisting of two charges rotating about their center of mass. In this case the period of the quadrupole moment is half the rotation period.

10.7 Thomson Scattering

Larmor’s formula gives the power radiated by an accelerated charge. Consider a point charge of mass m and charge q lying in a beam of linearly polarized radiation of frequency ω_0 with electric field

$$\mathbf{E} = E_0 \mathbf{e} \sin \omega_0 t \quad (10.46)$$

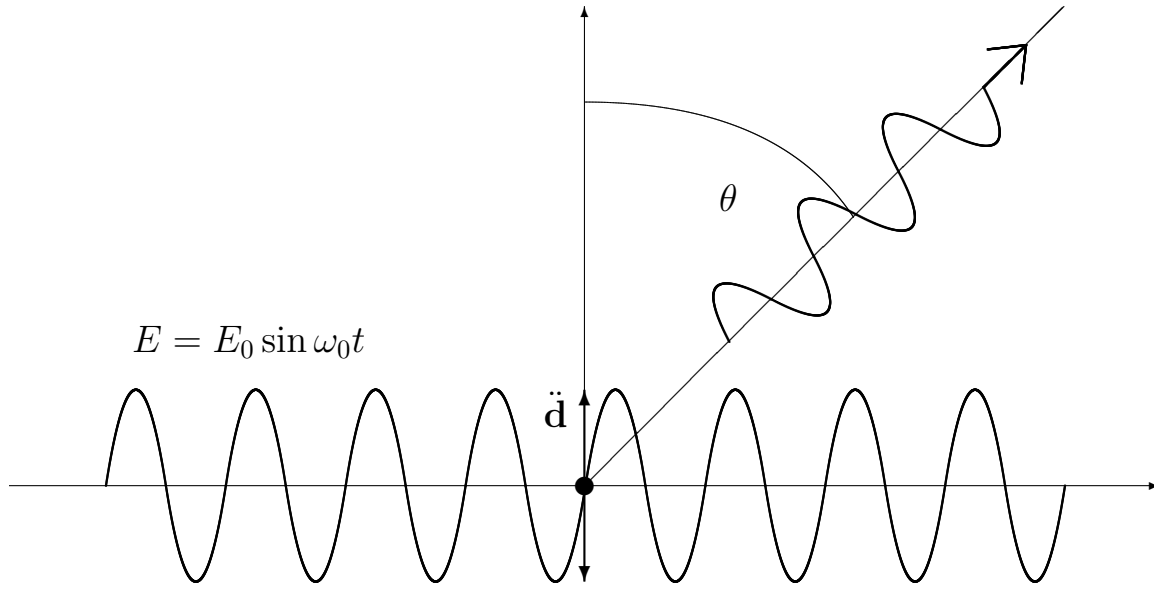


Figure 10.6: Geometry for the calculation of Thomson scattering. A linearly polarized beam of radiation enters from the left and illuminates an electron, generating an oscillating dipole moment as indicated. This dipole generates an outgoing wave of radiation with cylindrical symmetry.

as illustrated in figure 10.6. The force on the particle is $\mathbf{F} = q\mathbf{E}$ and results in an acceleration $\ddot{\mathbf{r}} = q\mathbf{E}/m$ and therefore a 2nd derivative of the dipole moment

$$\ddot{\mathbf{d}} = q\ddot{\mathbf{r}} = \frac{q^2 E_0 \mathbf{e}}{m} \sin \omega_0 t \quad (10.47)$$

so, according to Larmor, the power radiated (averaged over a cycle) is

$$\frac{dP}{d\Omega} = \frac{\langle |\ddot{\mathbf{d}}|^2 \rangle}{4\pi c^3} \sin^2 \theta = \frac{q^4 E_0^2}{8\pi m^2 c^3} \sin^2 \theta \quad (10.48)$$

and the total power radiated (or scattered) is

$$P = \frac{q^4 E_0^2}{3m^2 c^3} \quad (10.49)$$

The incident flux is $\langle S \rangle = c\langle E^2 \rangle/4\pi = cE_0^2/8\pi$, and if we define a differential cross-section $d\sigma/d\Omega$ for scattering such that

$$\frac{dP}{d\Omega} = \langle S \rangle \frac{d\sigma}{d\Omega} \quad (10.50)$$

then we have

$$\frac{d\sigma}{d\Omega} = \frac{q^4}{m^2 c^4} \sin^2 \theta. \quad (10.51)$$

For an electron, $q = e$, this is

$$\frac{d\sigma}{d\Omega} = r_0^2 \sin^2 \theta \quad (10.52)$$

where we have defined

$$r_0 = \frac{e^2}{m_e c^2} = 2.82 \times 10^{-13} \text{ cm} \quad (10.53)$$

as the *classical radius of the electron* (ie the radius of a cloud of charge e such that its electrostatic potential energy is equal to the electron rest-mass energy).

The total cross section is

$$\sigma \equiv \int d\Omega \frac{d\sigma}{d\Omega} = 2\pi r_0^2 \int_{-1}^{+1} d\mu (1 - \mu^2) = \frac{8\pi}{3} r_0^2 \quad (10.54)$$

For an electron this gives the *Thomson cross-section*

$$\sigma_T = 0.665 \times 10^{-24} \text{cm}^2. \quad (10.55)$$

The limitations on this result are

- The energy per photon should be much less than the rest-mass energy of the electron $h\nu \ll m_e c^2$ or $h\nu \ll 0.5 \text{MeV}$. At higher energies we need to take the recoil of the electron into account.
- The field should not be so strong as to accelerate the charge to relativistic velocity.

The foregoing was for an incident linearly polarized beam. Scattering of natural (unpolarized) radiation can be computed by considering a superposition of two incoherent linearly polarized beams with perpendicular polarization states.

Features of electron scattering:

- The cross-section is independent of frequency.
- The distribution of scattered energy is forward-backward symmetric.
- The total cross-section for scattering of natural radiation is the same as for a polarized beam.
- The scattered radiation is strongly linearly polarized (100% for the radiation scattered at right-angles to the incoming beam).

10.8 Radiation Reaction

Consider a simple harmonic oscillator consisting of a spring with a charged mass on the end. The energy radiated will tap the internal energy of this system and will cause the amplitude of the oscillations to decay.

This may be described, at least phenomenologically, as a *radiation reaction force*.

To order of magnitude, the time-scale for the oscillation to decay is $t_{\text{decay}} \sim mv^2/P_{\text{Larmor}} \sim (mc^3/e^2)t_{\text{orbit}}^2 \sim ct_{\text{orbit}}^2/r_0$ with $t_{\text{orbit}} = v/\dot{v}$. Thus the decay time will greatly exceed the orbital time or oscillator period provided $t_{\text{orbit}} \gg \tau$ where τ is the time for light to cross the classical radius of the electron: $\tau = 2e^2/3mc^3 \sim 10^{-23} \text{s}$. If $t_{\text{orbit}} \gg \tau$, as is usually the case, radiation reaction can be treated as a small perturbation.

The radiation reaction force \mathbf{F} must satisfy

$$\int dt \mathbf{F} \cdot \mathbf{u} = \int dt P_{\text{Larmor}} = \int dt \frac{2e^2 \dot{u}^2}{3c^3} \quad (10.56)$$

where the integrations are over one or more periods of the oscillation. An acceptable choice is

$$\mathbf{F} = \frac{2e^2 \ddot{\mathbf{u}}}{3c^3} = m\tau \ddot{\mathbf{u}}. \quad (10.57)$$

thought this should be treated with caution (see R+L).

10.9 Radiation from Harmonically Bound Particles

Including the radiation reaction force, the equation of motion for an electron tethered by a spring with spring constant k , and corresponding frequency $\omega_0 = \sqrt{k/m_e}$, is

$$\frac{d^2x}{dt^2} + \omega_0^2 x - \tau \frac{d^3x}{dt^3} = 0 \quad (10.58)$$

Evaluating the radiation reaction term using the undamped solution $x \propto \cos \omega_0 t$ gives $d^3x/dt^3 = -\omega^2 dx/dt$ and the E.O.M becomes

$$\ddot{x} + \omega_0^2 \tau \dot{x} + \omega_0^2 x = 0 \quad (10.59)$$

in which the radiation reaction term now takes the same form as would a frictional force.

Searching for solutions of the form $x \propto \exp(\alpha t)$ (with α complex to allow for decaying oscillations) converts this to an algebraic equation

$$\alpha^2 + \omega_0^2 \tau \alpha + \omega_0^2 = 0 \quad (10.60)$$

with solutions

$$\alpha = \pm i\omega_0 - \frac{1}{2}\omega_0^2 \tau + \dots \quad (10.61)$$

With initial conditions $x(0) = x_0$ and $\dot{x}(0) = 0$ the solution is

$$x(t) = x_0 e^{-\Gamma t/2} \cos \omega_0 t \quad (10.62)$$

which is a damped oscillation with decay rate

$$\Gamma = \omega_0^2 \tau = \frac{2e^2 \omega_0^2}{3mc^3} \quad (10.63)$$

in agreement with the order of magnitude argument above.

The transform of this decaying oscillator is

$$x(\omega) = \int dt x(t) e^{i\omega t} = \left[\frac{1}{\Gamma/2 - i(\omega - \omega_0)} + \frac{1}{\Gamma/2 - i(\omega + \omega_0)} \right] \quad (10.64)$$

which is small, save in the vicinity of $\omega = \pm \omega_0$. The radiated power is $dW/d\omega \sim \omega^4 |x(\omega)|^2$ or, more precisely

$$\frac{dW}{d\omega} = \frac{8\pi\omega^4 e^2 x_0^2}{3c^3 (4\pi)^2} \frac{1}{(\omega - \omega_0)^2 + (\Gamma/2)^2} \quad (10.65)$$

The radiated power thus has a Lorentzian profile.

Radiation damping therefore imposes a minimum width on spectral lines for electronic oscillators: $\Delta\omega = \Gamma = 2e^2 \omega_0^2 / 3mc^3$ or equivalently a width in wavelength of $\Delta\lambda = (\lambda/\omega_0) \delta\omega = 2\pi c\tau \simeq 1.2 \times 10^{-4} \text{ \AA}$.

10.10 Scattering by Bound Charges

In the case of scattering by free charges — Thomson scattering — the incident electric field produces an *acceleration* proportional to the field, and hence $\ddot{\mathbf{d}} \propto \mathbf{E}$ and therefore a scattered radiation field with amplitude proportional to the incoming field, resulting in a scattering cross-section which is independent of frequency.

In the case of scattering by a charge bound in a quadratic potential well with free oscillation frequency ω_0 , we expect the cross-section to be equal to σ_T for $\omega \gg \omega_0$. For low-frequency incident radiation $\omega \ll \omega_0$ the situation is different; here the incident field causes the charge to move to a displaced position such that the electric field is balanced by the spring constant force, but this

results in a displacement, and therefore a dipole moment, which is proportional to the field (rather than the second time derivative of the dipole being proportional to the field).

For a free charge, the dipole amplitude is $\mathbf{d} \propto \mathbf{E}/\omega^2$ whereas for the bound charge and $\omega \ll \omega_0$, the dipole amplitude is $\mathbf{d} \sim \mathbf{E}/\omega_0^2$, which is smaller by a factor $(\omega/\omega_0)^2 \ll 1$, and the scattered radiation intensity — being proportional to the square of the dipole — is smaller than for free charges by a factor $(\omega/\omega_0)^4$. Thus for low frequency radiation we expect

$$\sigma(\omega) \simeq (\omega/\omega_0)^4 \sigma_T \quad (10.66)$$

A more detailed analysis shows that there is a *resonance* at $\omega = \omega_0$ where the cross section becomes very large. In the absence of damping, this resonance is infinitely sharp, and the cross-section becomes formally infinite. Including radiation, or other, damping results in finite width to the resonance $\delta\omega \simeq \Gamma$.

10.11 Problems

10.11.1 Antenna beam pattern

Two oscillating dipole moments (antennae) $\mathbf{d}_1, \mathbf{d}_2$ are oriented vertically and are a horizontal distance L apart. They oscillate in phase with the same frequency ω . Consider radiation emitted in a direction at angle θ with respect to the vertical and in the vertical plane containing the two dipoles.

- a. Show that at large distances $D \gg L$ the angular distribution of the radiated power is

$$\frac{dP}{d\Omega} = \frac{w^4 \sin^2 \theta}{4\pi c^3} (d_1^2 + 2d_1 d_2 \cos \delta + d_2^2) \quad (10.67)$$

where the ‘phase angle’ is $\delta = \omega L \sin \theta / c$.

- b. Show that when $L \ll \lambda$ the radiation is the same as from a single oscillating dipole of amplitude $\mathbf{d}_1 + \mathbf{d}_2$
- c. Generalise to a 2-dimensional array $n(x, y) = \sum_i \delta(x - x_i, y - y_i)$ containing a large number of antennae laid out in the plane $z = 0$, now with their dipoles lying in the horizontal plane (parallel to the x -axis) say. Show that the ‘synthesised beam pattern’ — which we define to be the power radiated as a function of direction $\boldsymbol{\theta}$ — is just the square of the modulus of the fourier transform of the array pattern n :

$$P(\theta_x, \theta_y) \propto \left| \int \int dx dy n(x, y) e^{2\pi i(x\theta_x + y\theta_y)/\lambda} \right|^2 = \left| \sum_i e^{2\pi i x_i \cdot \boldsymbol{\theta} / \lambda} \right|^2 \quad (10.68)$$

(you may adopt the small angle approximation).

- d. The VLA is a Y-shaped array of radio receivers (which could also be used as a highly directional transmitter). Sketch the resulting beam pattern.

10.11.2 Multipole radiation 1

At large distances R_0 from a relatively compact distribution of radiating currents the vector potential can be written as

$$\mathbf{A} = \frac{1}{cR_0} \int d^3 r \mathbf{j}_{t'} + \mathbf{r} \cdot \mathbf{n} / c \quad (10.69)$$

where $t' = t - R_0/c$ and \mathbf{n} is a unit vector in the direction from the radiating system to the observer.

- a. Expand in powers of $\mathbf{r} \cdot \mathbf{n} / c$ to obtain

$$\mathbf{A} = \frac{1}{cR_0} \int d^3 r \mathbf{j}_{t'} + \frac{1}{c^2 R_0} \frac{\partial}{\partial t'} \int d^3 r (\mathbf{r} \cdot \mathbf{n}) \mathbf{j}_{t'} \quad (10.70)$$

- b. Show that, for a system of point charges, this becomes

$$\mathbf{A} = \frac{\sum q\mathbf{v}}{cR_0} + \frac{1}{c^2 R_0} \frac{\partial}{\partial t} \sum q\mathbf{v}(\mathbf{r} \cdot \mathbf{n}) \quad (10.71)$$

- c. Show that

$$2\mathbf{v}(\mathbf{r} \cdot \mathbf{n}) = \frac{\partial}{\partial t} \mathbf{r}(\mathbf{r} \cdot \mathbf{n}) + (\mathbf{r} \times \mathbf{v}) \times \mathbf{n} \quad (10.72)$$

and thus obtain

$$\mathbf{A} = \frac{\dot{\mathbf{d}}}{cR_0} + \frac{1}{2c^2 R_0} \frac{\partial^2}{\partial t^2} \sum q\mathbf{r}(\mathbf{n} \cdot \mathbf{r}) + \frac{1}{cR_0} (\dot{\mathbf{m}} \times \mathbf{n}) \quad (10.73)$$

where the dipole moment is $\mathbf{d} \equiv \sum q\mathbf{r}$ and the magnetic moment is $\mathbf{m} \equiv \sum q\mathbf{r} \times \mathbf{v}/c$. Thus, in this approximation, the potential contains terms proportional to the first time derivative of the dipole (dipole radiation); the second time derivative of the quadrupole moment (quadrupole radiation) and the first time derivative of the magnetic moment (magnetic dipole radiation).

- d. Show that the radiation field intensity from a bar magnet spinning about an axis perpendicular to the line joining its poles is on the order of $E, B \sim \ddot{m}/c^2 R_0$ and that the radiated power is $P \sim \ddot{m}^2/c^3$.

10.11.3 Multipole radiation 2

Starting with the expression for the retarded vector potential in the form

$$\mathbf{A}(\mathbf{r}, t) = \int d^3 r' \int dt' \frac{\mathbf{j}(\mathbf{r}', t')}{|\mathbf{r} - \mathbf{r}'|} \delta(t' - t + |\mathbf{r} - \mathbf{r}'|/c) \quad (10.74)$$

- a. Show that the temporal fourier transforms $\mathbf{A}_\omega(\mathbf{r}) \equiv \int dt \mathbf{A}(\mathbf{r}, t) \exp(i\omega t)$ and $\mathbf{j}_\omega(\mathbf{r}) \equiv \int dt \mathbf{j}(\mathbf{r}, t) \exp(i\omega t)$ are related by

$$\mathbf{A}_\omega(\mathbf{r}) = \frac{1}{c} \int d^3 r' \frac{\mathbf{j}_\omega(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} e^{ik|\mathbf{r} - \mathbf{r}'|} \quad (10.75)$$

where $k = \omega/c$. Note that this gives a one-to-one relationship between the temporal fourier components of \mathbf{A} and \mathbf{j} . Note also, that the field $\mathbf{A}_\omega(\mathbf{r})$ is just a convolution of the source $\mathbf{j}(\mathbf{r})$ with the convolution kernel $e^{ik|\mathbf{r}|}/r$.

- b. For sources of size L and field points at distances $r \gg L$ we can take $|\mathbf{r} - \mathbf{r}'| \simeq r - \mathbf{n} \cdot \mathbf{r}'$ in the exponential (where \mathbf{r}' is measured relative to a spatial origin inside the source) and one can take $|\mathbf{r} - \mathbf{r}'| \simeq r$ in the demoninator. Thus show that in this approximation

$$\mathbf{A}_\omega(\mathbf{r}) = \frac{e^{ikr}}{cr} \sum_{m=0}^{\infty} \frac{1}{m!} \int d^3 r' \mathbf{j}_\omega(\mathbf{r}') (-ik\mathbf{n} \cdot \mathbf{r}')^m \quad (10.76)$$

For sources with dimension $L \ll \lambda$, this is an expansion in the small dimensionless quantity $k\mathbf{n} \cdot \mathbf{r}'$. The $n = 0$ component gives the dipole radiation, $n = 1$ gives the quadrupole term etc.

- c. Now specialise to the case of a point charge q moving in a circle of radius r_0 in the $z = 0$ plane with frequency ω_0 :

$$\mathbf{j}(\mathbf{r}, t) = q\mathbf{v}\delta(\mathbf{r} - \mathbf{R}(t)) \quad (10.77)$$

where $\mathbf{R} = \{r_0 \cos \omega_0 t, r_0 \sin \omega_0 t, 0\}$, $\mathbf{v} = \{-\omega_0 r_0 \sin \omega_0 t, \omega_0 r_0 \cos \omega_0 t, 0\}$. Show that the dipole radiation is non-zero only at frequency $\omega = \omega_0$, the quadrupole radiation appears at $\omega = 2\omega_0$ and so on.

10.11.4 Electron scattering

Larmor's formula gives the power radiated by an accelerated charge as $P \sim q^2 a^2 / c^3$ where q is the charge and a is the acceleration.

- To order of magnitude, estimate the cross section for scattering of radiation by a free electron in terms of the electronic charge, mass and the speed of light.
- Now consider an electron bound in a parabolic potential well such that the free oscillation frequency is ω_0 . Estimate the cross section for scattering of light as a function of frequency ω for the two limiting regimes $\omega \gg \omega_0$ and $\omega \ll \omega_0$.

10.11.5 Thomson drag

Consider a radiation field which, in a particular frame of reference, appears isotropic with specific intensity I_ν .

- Show that a photon with frequency ν in the isotropic frame will appear to an observer moving at velocity $v = \beta c \ll c$ with respect to this frame to have frequency $\nu' = (1 - \beta\mu)\nu$ where μ is the cosine of the angle between the photon momentum and the direction of motion.
- Invoke Lorentz invariance of I_ν/ν^3 to show that the specific intensity seen by the moving observer is

$$I'_\nu = I_\nu - \beta\mu(3I_\nu - \nu\partial I_\nu/\partial\nu) \quad (10.78)$$

and therefore that

$$I' = I - 4\beta\mu I \quad (10.79)$$

- Show that an electron moving through this radiation will feel a drag force $F = (4/3)\beta U\sigma_T$ where $U = 4\pi I/c$ is the energy density and σ_T is the Thomson cross-section.
- Specialise to a black-body radiation field and show that the velocity of an electron will decay exponentially with $v \propto \exp(-t/\tau_e)$ on a timescale $\tau_e \sim m_e c / (a T^4 r_0^2)$.
- Compute the decay time τ_p for the velocity of a blob of fully ionized plasma moving relative to an isotropic black body radiation field.
- Compute these timescales for electrons or plasma moving through the microwave background ($T = 2.7\text{K}$) and compare to the age of the Universe $t \sim H_0^{-1} \sim 10^{10}$ years. Conclude that these effects are of little consequence at the present. However, in the expanding universe, the drag rates scale as $\tau^{-1} \propto T^4 \propto (1+z)^4$ while the Hubble rate scales as $H \propto (1+z)^{3/2}$ (matter dominated). Estimate the redshifts at which $H\tau_e = 1$ and $H\tau_p = 1$. At redshifts exceeding the latter, plasma will be effectively locked to the frame of isotropy of the microwave background. At redshifts exceeding the former, the microwave background will act as an efficient coolant for hot ionized gas.

10.11.6 Polarisation

Sketch the pattern of linear polarisation expected for a non polarised point source embedded in a optically thin cloud of ionized gas (assume the scattering is dominated by electron scattering).

Chapter 11

Cerenkov Radiation

Most radiation generation involves accelerated charges, since a uniformly moving charge does not radiate. An exception to this rule is a charge which is moving faster than the speed of light. This super-luminal motion may arise in two ways; ordinarily when a high energy particle passes through a medium with refractive index $n > 1$, such as a cosmic ray entering the atmosphere. Alternatively, one may have a charge *concentration* which moves super-luminally (think of caterpillar legs). This is not as crazy as it sounds; a group in Oxford are building a device which produces a charge pattern with super-luminal motion.

Cerenkov radiation can be likened to the *sonic boom* from a supersonic airplane. Figure 11.1 shows how a supersonic, or super-luminal, particle will outrun any disturbance it makes, and that one would expect a conical pulse of radiation, with the normal to the surface of the cone making an angle $\cos^{-1}(c/v)$ with the direction of motion of the particle.

We will analyze the resulting *Cerenkov* or *Heaviside* radiation first using the retarded potential, and then using the LW potentials.

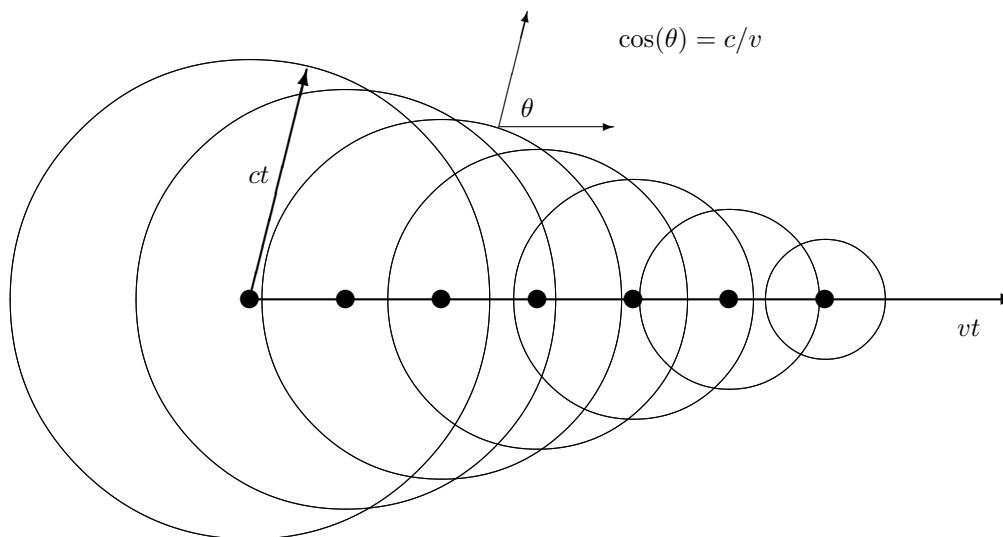


Figure 11.1: A supersonic plane excites a ‘sonic-boom’; a wave pulse propagating out with conical wave fronts with angle, relative to the direction of motion, $\theta = \cos^{-1}(c/v)$.

11.1 Retarded Potential

For a field point very distant from the source of the radiation, the spectral decomposition of the magnetic potential can be written as

$$\mathbf{A}_\omega(\mathbf{r}) = -\frac{e^{ikr}}{cr} \int dt' \int d^3r' \mathbf{j}(\mathbf{r}', t') e^{i\omega t'} e^{-ik\mathbf{n}\cdot\mathbf{r}'} \quad (11.1)$$

where $\mathbf{n} = \mathbf{r}/r$.

If we now introduce a current density which describes a charge (or a charge *concentration*) moving at a uniform velocity v along the z -axis:

$$\mathbf{j}(\mathbf{r}, t) = \hat{\mathbf{z}} qv \delta(x) \delta(y) \delta(z - vt) \quad (11.2)$$

the potential is

$$\mathbf{A}_\omega(\mathbf{r}) = -\frac{e^{ikr}}{cr} qv \hat{\mathbf{z}} \int dt e^{i\omega(1-n_z v/c)t}. \quad (11.3)$$

Now the integral here is a representation of the δ -function: $2\pi\delta(\omega(1-n_z v/c))$. Since $n_z \leq 1$, the argument of the delta function is non-zero for all ω if $v \leq c$, and there is no radiation. However, if $v > c$ the argument of the δ -function vanishes for the direction $n_z = c/v$, and the potential is very strong (formally infinite). This is just the direction of the outgoing conical wave.

In addition to being highly directional, the sonic-boom analogy suggests that the radiation would have a rather blue spectrum. This is indeed the case. To determine the spectrum, consider a relativistic particle propagating superluminally for finite time T , with $v > c$ (as, for example as a energetic particle passes through a slab with refractive index $n > 1$. See figure 11.2). The time integral is now bounded, and becomes a ‘sinc’ function, rather than a δ -function, and the magnetic potential is then

$$\mathbf{A}_\omega(\mathbf{r}) = -\frac{qv\hat{\mathbf{z}}T}{c} \frac{e^{ikr}}{r} \text{sinc}(\omega(1-n_z v/c)T/2) \quad (11.4)$$

with $\text{sinc}(x) \equiv \sin(x)/x$. The field is obtained in the usual way from the potential as $\mathbf{B} = \nabla \times \mathbf{A}$. Again as usual, the most rapidly spatially varying factor here is e^{ikr} . When we apply the gradient operator we can effectively ignore the variation in r provided $r \gg 1/k$. The sinc function also depends on \mathbf{r} since $n_z \equiv \hat{r}_z$, but this contribution to the spatial gradient is also negligible for $r \gg T$. If these conditions are satisfied, the field is then

$$\mathbf{B}_\omega(\mathbf{r}) = -\frac{i\omega qv\hat{\mathbf{z}} \times \hat{\mathbf{r}}T}{c^2} \frac{e^{ikr}}{r} \text{sinc}(\omega(1-n_z v/c)T/2). \quad (11.5)$$

The Poynting flux is $S = c(E^2 + B^2)/8\pi = cB^2/4\pi$ and is directed along $\hat{\mathbf{r}}$. The total energy emitted into solid angle $d\Omega$ is

$$dW = r^2 d\Omega \int dt S(t) = \frac{cr^2 d\Omega}{4\pi} \int dt B^2(t) = \frac{cr^2 d\Omega}{2\pi} \int_0^\infty \frac{d\omega}{2\pi} B_\omega^2(t) \quad (11.6)$$

or equivalently $dS_\omega = cB_\omega^2 d\omega/(2\pi)^2$. The system is cylindrically symmetric, so the solid angle is $d\Omega = 2\pi \sin\theta d\theta$ and the amount of energy emerging per unit frequency per unit angle is

$$d^2W = 2\pi r^2 dS_\omega d\theta \sin\theta = \frac{q^2 v^2 \omega^2 \sin^3\theta d\theta d\omega}{4\pi c^4} \text{sinc}^2(\omega(1-n_z v/c)T/2) \quad (11.7)$$

where we have used $|\hat{\mathbf{z}} \times \hat{\mathbf{r}}| = \sin\theta$.

Equation (11.7) tells us how the emergent energy is distributed over frequency and over angle. The sinc^2 factor limits the contribution to a small range in angle $\Delta\theta \sim 1/\omega T \tan\theta_0$ around $\theta_0 = \cos^{-1} c/v$. Integrating over direction to obtain the total spectrum, we can ignore the variation of the

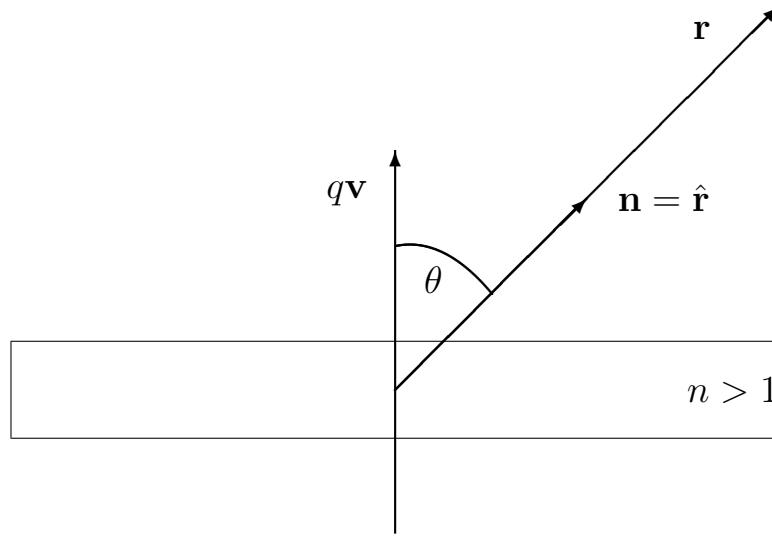


Figure 11.2: Geometry for Cerenkov radiation calculation.

$\sin^3 \theta$ term. Changing integration variable from θ to $\zeta = \omega(1 - n_z v/c)T/2 \simeq \omega \tan \theta_0 (\theta - \theta_0)T/2$ so $d\theta = 2d\zeta/\omega T \tan \theta_0$ gives

$$dW = \frac{q^2 v (1 - c^2/v^2) T \omega d\omega}{\pi c^2} \int d\zeta \sin^2 \zeta / \zeta^2. \quad (11.8)$$

The dimensionless integral here introduces a factor of order unity. Equation (11.7) shows that the energy radiated is proportional to the time of flight, as one might have expected, and therefore to the path length through the refractive medium. It also shows that the spectrum is blue, with $dW/d\omega \propto \omega$.

11.2 LW Potentials

The treatment above provides a useful illustration of the use of retarded potentials. It is also illustrative to obtain the form of the radiation pulse using the Lienard-Wiechart potentials.

Recall that the LW potentials are just equal to the electro-static and magneto-static potentials times a factor $1/(1 - \hat{\mathbf{R}} \cdot \mathbf{v})$. This means that the potential becomes very large for a particle moving directly towards the observer (which can also be understood in terms of relativistic beaming). An entirely analogous situation arises here where the particle world line can ‘graze’ the observer’s light cone, as illustrated in figure 11.3. This is very different from the situation for a particle moving slower than c , for which the particle world line pierces the past light cone of any point on the observers world line, and does so exactly once. There is therefore a specific instant on the observer’s world line when the particle grazes it’s past light cone. At that instant $\hat{\mathbf{R}} \cdot \mathbf{v} = 1$ and the effective charge of the particle becomes infinite. At later observer times, the particle pierces the light cone twice, and observer perceives a potential which is the sum of that from two particles of finite charge.

As another way of looking at this, we saw that the velocity dependent boost factor $1/(1 - \hat{\mathbf{R}} \cdot \mathbf{v})$ was equivalent to $dt/d\tau$ where dt is the coordinate time interval spent by the particle between two observer past light cones whose apices differ by time $d\tau$. Clearly, as the particle grazes the critical light cone, the observer time τ is stationary with respect to the coordinate time, and $dt/d\tau \rightarrow \infty$.

Let the charged particle move along the trajectory

$$\mathbf{r}(t) = (0, 0, vt) \quad (11.9)$$

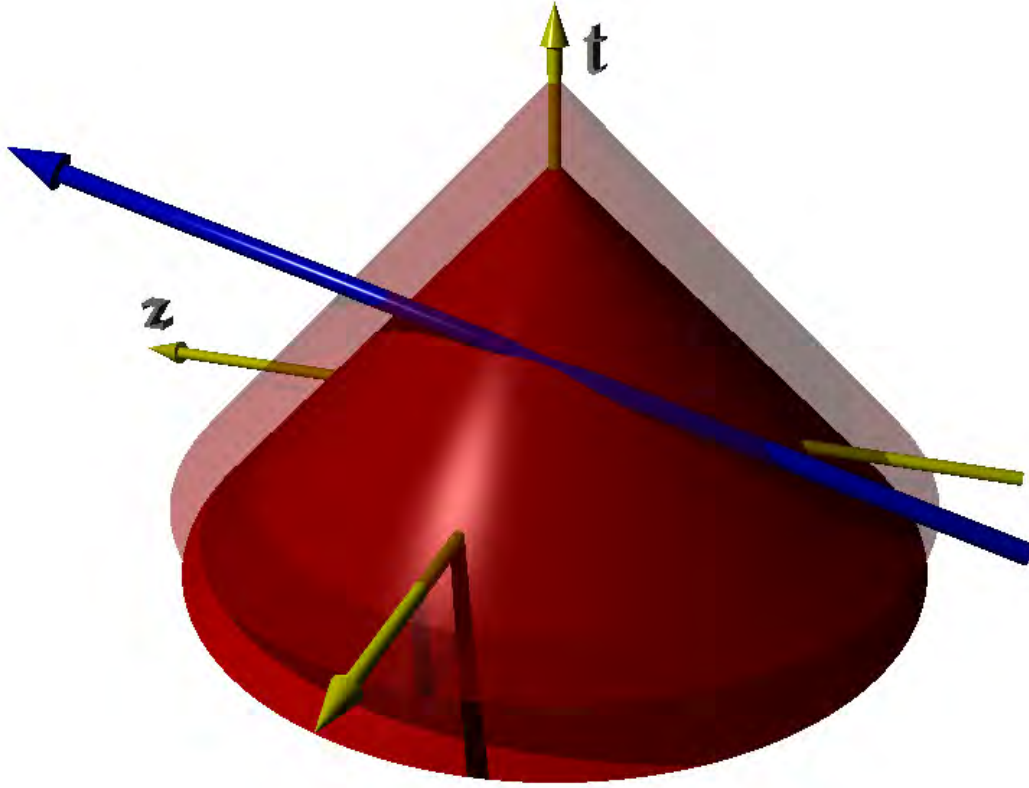


Figure 11.3: Intersection of the world-line of a super-luminal charged particle (or charge disturbance) and the past light cone of an observer. For sufficiently early times ($t < t_*$) the particle world-line does not intersect the light cone at all, and the observer sees no field whatsoever. For $t = t_*$ the particle world-line just grazes the light cone, as illustrated by the opaque cone. For later times (e.g. semi-transparent cone) the world line intersects the past light cone at two points.

and let us place the observer at location $\mathbf{r}_{\text{obs}} = (x_0, 0, 0)$. The intersection of the past light cone of the event $(\tau, x_0, 0, 0)$ and the plane $x = 0, y = 0$ is

$$z^2 = c^2(t - \tau)^2 - x_0^2. \quad (11.10)$$

The intersection of the particle's trajectory and the past light cone occurs at coordinate times t such that $z = vt = \sqrt{c^2(t - \tau)^2 - x_0^2}$ or

$$t = \frac{-\tau \pm \sqrt{\tau^2 v^2 / c^2 - (v^2 / c^2 - 1)x_0^2 / c^2}}{v^2 / c^2 - 1}. \quad (11.11)$$

This gives two real roots only for $\tau > \tau_*$ where

$$c\tau_* = x_0 \sqrt{1 - c^2 / v^2}. \quad (11.12)$$

This is the instant of grazing. Note that we need to use $\tau_* > 0$. The alternate case corresponds to the particle grazing the future light cone.

Now with $\mathbf{v} = (0, 0, v)$ and $\mathbf{R} = (x_0, 0, -vt)$ one can readily show that at the instant of grazing, $t = \tau_*/(v^2/c^2 - 1)$, the factor $1 - \hat{\mathbf{R}} \cdot \mathbf{v}$ vanishes, as claimed above. Alternatively, and more simply,

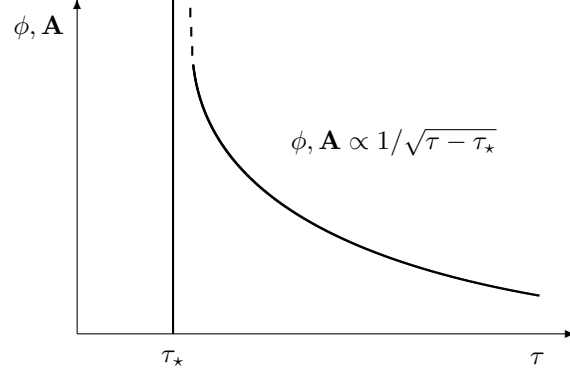


Figure 11.4: The form of a pulse of Cerenkov radiation.

taking the derivative of (11.11) with respect to τ we find

$$\frac{dt}{d\tau} = \frac{1}{v^2/c^2 - 1} \left[-1 \pm \frac{v}{c} \frac{\tau}{\sqrt{\tau^2 - \tau_*^2}} \right] \quad (11.13)$$

The observer feels the potential from two sources $\mathbf{A} = \mathbf{A}_+ + \mathbf{A}_-$:

$$\mathbf{A}_\pm(\tau) = \frac{q\mathbf{v}}{cr(t_\pm)} \left| \frac{dt}{d\tau} \right|_\pm. \quad (11.14)$$

Specialising to the behaviour of the pulse close to the leading edge (i.e. $\tau \simeq \tau_*$), we have $\sqrt{\tau^2 - \tau_*^2} = \sqrt{(\tau + \tau_*)(\tau - \tau_*)} \simeq \sqrt{2\tau_*} \sqrt{\tau - \tau_*}$, and the potential is

$$\mathbf{A}(\tau) \simeq \frac{q\mathbf{v}}{cr} \frac{v/c}{v^2/c^2 - 1} \sqrt{\frac{2\tau_*}{\tau - \tau_*}} \quad (11.15)$$

with $r = x_0/\sin\theta$.

The form of the resulting potential wave pulse is therefore as illustrated schematically in figure 11.4. The potential diverges as $1/\sqrt{\Delta\tau}$ as τ approaches τ_* . The characteristic form is identical to that of a ‘fold-caustic’.

The magnetic field is $\mathbf{B} = \nabla \times \mathbf{A}$. Since $\mathbf{v} = v\hat{\mathbf{z}}$, the field at $(x_0, 0, 0)$ is $\mathbf{B} = \hat{\mathbf{y}}\partial A/\partial x_0$. Now x_0 appears in $\tau_* = x_0 \sin\theta$ and also in r . Close to the pulse edge, however, the most rapid variation is in the factor $1/\sqrt{\tau - x_0 \sin\theta/c}$ and we can therefore replace $\partial A/\partial x_0$ by $-c^{-1} \sin\theta \partial A/\partial \tau$. The magnetic field is then

$$B = \frac{2^{3/2}q\sqrt{1 - c^2/v^2}}{\sqrt{x_0 c^3 (\tau - \tau_*)^3}}. \quad (11.16)$$

The field therefore diverges as $1/\Delta\tau^{3/2}$ as $\tau \rightarrow \tau_*$. This implies a formally infinite (integrated) Poynting flux: $\int d\tau B^2 \propto 1/\Delta\tau^2$. This divergence is the real-space analog of the divergent energy computed from the power spectrum over $dW \propto \omega d\omega \propto \omega_{\max}^2$. This divergence would be removed, for example, by introducing a finite width for the shower of particles in an atmospheric Cerenkov shower, for example.

One can also compute the power spectrum of the radiation, and the dependence on v/c , from the potential (11.15). Taking the temporal transform we find

$$\mathbf{A}_\omega = \int dt \mathbf{A}(t) e^{i\omega t} \sim \frac{q\mathbf{v}}{r} \frac{v}{v^2 - c^2} \sqrt{\tau_*/\omega} e^{i\omega\tau_*} \quad (11.17)$$

where we have dropped some dimensionless factors of order unity. Since $cB(t) = -\sin\theta \partial A/\partial t$, $B_\omega = -i \sin\theta A_\omega/(\omega)$ and the Poynting flux is then

$$dS_\omega \sim B_\omega^2 d\omega \sim \frac{q^2 \sin^2 \theta \tau_* \omega d\omega}{c^2 r^2 (1 - c^2/v^2)^2}. \quad (11.18)$$

If we erect a cylinder of length L and radius x_0 around the particle trajectory, the component of the energy flux in the direction normal to the surface is $\sin \theta dS$. The area area of the cylinder is $2\pi x_0 L$, so the energy crossing the surface (per unit frequency) is $dW = 2\pi x_0 L \sin \theta dS_\omega$. With $x_0 = r \sin \theta$, $c\tau_\star = x_0 \sin \theta$ and with $L = vT = cvT/\cos \theta$ we obtain

$$dW \sim \frac{q^2 v (1 - c^2/v^2) T \omega d\omega}{c^2} \quad (11.19)$$

in agreement with (11.8).

Astronomical applications include cosmic-ray detection *via* atmospheric Cerenkov radiation, and also in solid state Cerenkov detectors. The other scenario for generating radiation from a superluminal charge or current disturbance may have implications for radiation from pulsars.

Chapter 12

Bremsstrahlung

Bremsstrahlung, or ‘braking-radiation’, also known as *free-free emission*, is produced by collisions between particles in hot ionized plasmas.

- Bremsstrahlung arises predominantly from collisions between electrons and ions. Electron-electron collisions are ineffective as they produce no dipole radiation. Collisions between ions with different charge-to-mass ratio are capable of generating dipole radiation, but their low accelerations render them also unimportant.
- In an electron-ion collision we can take the ion to be unaccelerated.
- Precise results require quantum treatment, but useful approximate results can be obtained from a classical calculation of the dipole radiation, with plausible cut-offs.

In what follows we will first compute the radiation power spectrum from a single collision with given electron velocity and impact parameter. We then integrate over impact parameter to get the emission from a single-speed electron component, and then integrate over a thermal distribution of electron velocities to obtain the thermal bremsstrahlung emissivity. We briefly mention thermal bremsstrahlung absorption and the emission from a plasma with relativistic electron velocities.

12.1 Radiation from a Single Collision

The geometry for the collision of an electron with an ion of charge $+Ze$ is shown in figure 12.1.

The energy radiated into a range of directions $d\Omega$ around $\hat{\mathbf{r}}$ is, as usual,

$$dW = r^2 d\Omega \int dt S(t) \quad (12.1)$$

with Poynting flux

$$S(t) = \frac{cE^2(t)}{4\pi} = \frac{cB^2(t)}{4\pi} \quad (12.2)$$

and at very large distances the radiation field is

$$\mathbf{B}(t) = \frac{\hat{\mathbf{r}} \times \ddot{\mathbf{d}}(t - r/c)}{c^2 r} \quad (12.3)$$

Integrating over directions gives

$$W = \frac{1}{c^3} \int dt \int \frac{d\Omega}{4\pi} |\hat{\mathbf{r}} \times \ddot{\mathbf{d}}(t)|^2 = \frac{2}{3c^3} \int dt |\ddot{\mathbf{d}}(t)|^2 = \frac{2}{3c^3} \int \frac{d\omega}{2\pi} |\tilde{\ddot{\mathbf{d}}}(\omega)|^2 \quad (12.4)$$

by Parseval’s theorem, or

$$dW = \frac{2|\tilde{\ddot{\mathbf{d}}}(\omega)|^2}{3\pi c^3} d\omega. \quad (12.5)$$

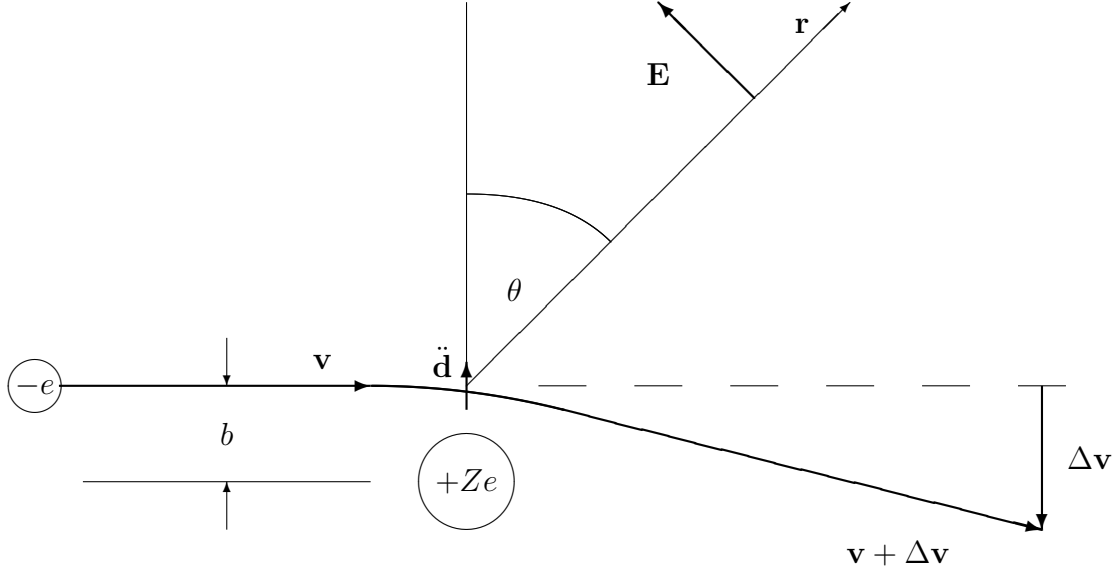


Figure 12.1: Geometry of an electron-ion collision. See text.

The collision velocity and impact parameter define a characteristic time $t_* = b/v$ for changes in the electron velocity and a corresponding characteristic frequency $\omega_* \sim v/b$ at which we expect most of the radiation to be emitted. The 2nd derivative of the dipole moment is $\ddot{\mathbf{d}} = -e\dot{\mathbf{v}} = -e\mathbf{a}$, and the acceleration is an impulse of strength $a_{\max} = Ze^2/b$ and duration $\sim t_*$. For $\omega \ll \omega_*$ then we have $\int dt \dot{\mathbf{v}} e^{i\omega t} \simeq \int dt \dot{\mathbf{v}} = \Delta \mathbf{v}$. It is not difficult to show that, for small deflection angles, the net impulse is

$$\Delta \mathbf{v} = \frac{2Ze^2}{mbv} \quad (12.6)$$

and we find

$$\frac{dW(b)}{d\omega} \simeq \begin{cases} \frac{8Z^2e^6}{3\pi c^3 m^2 v^2 b^2} & \text{for } \omega \ll \omega_* = v/b \\ 0 & \omega \gg \omega_* = v/b \end{cases} \quad (12.7)$$

Thus, for given v and b the spectrum of the radiation is flat, with strength $\propto 1/v^2 b^2$ up to the cut off ω_* .

The total energy emitted is $dW \sim \omega_* dW/d\omega$ and is inversely proportional to the velocity. This is because the slower particles are accelerated for a longer time. The more rapid collisions then spread the energy over a greater bandwidth resulting in $dW/d\omega \propto 1/v^2$.

12.2 Photon Discreteness

Classically, the total energy emitted in a collision is $dW \sim e^2 \dot{u}^2 t_*/c^3 \sim e^6/m^2 c^3 b^3 v$ and emerges at frequency $\omega \simeq v/b$. It is interesting to compare this to the energy of a quantum of this frequency:

$$\frac{dW}{h\nu} \simeq \left(\frac{e^2}{hc} \right)^3 \left(\frac{h\nu}{mv^2} \right)^2 \quad (12.8)$$

The first factor here is the cube of the fine structure constant and is on the order of 10^{-6} , and the second factor can be at most of order unity for any physically allowed reaction, so this number is always very small.

The energy emitted according to the classical dipole calculation is therefore less than the *typical* energy of any photons which are actually emitted by a huge factor. The classical calculation however correctly gives the *mean* energy emitted in a collision. Evidently, the rate of collision events that actually generate a photon is much less than the classical frequency of collisions.

12.3 Single-Speed Electron Stream

For a single electron with velocity \mathbf{v} passing through a cloud of static ion targets with space density n_i , the rate of collisions with impact parameter in the range b to $b + db$ is $dN/dt = 2\pi n_i v b db$.

For a stream of electrons with space density n_e then the rate of collisions per unit volume with impact parameter in this range is $dN/dV dt = 2\pi n_i n_e v b db$. Integrating over impact parameter gives the energy radiated per unit bandwidth per unit volume per unit time of

$$\frac{dW}{d\omega dV dt} = 2\pi n_i n_e v \int db b \frac{dW(b)}{d\omega} \simeq \frac{16Z^2 e^6}{3c^3 m^2 v} \ln(b_{\max}/b_{\min}). \quad (12.9)$$

The upper limit b_{\max} arises from the requirement $\omega < \omega_*$, or

$$b \lesssim b_{\max}(\omega, v) = v/\omega. \quad (12.10)$$

What sets the lower limit b_{\min} ? At low energies it is the condition that the scattering angle be small, but for energies $\gtrsim 10\text{eV}$ the lower limit is set by quantum mechanics: an electron with velocity v has a de Broglie wavelength $\lambda_{\text{dB}} \sim h/mv$ and it makes no sense to treat the electron as point-like for impact parameters less than λ_{dB} , and it is reasonable to adopt a cut-off

$$b \gtrsim b_{\min}(v) = \frac{h}{mv}. \quad (12.11)$$

Equation (12.9) together with (12.10, 12.11) suggests that the radiated power spectrum is only weakly (logarithmically) frequency dependent. This is true at sufficiently low frequencies, but note that b_{\max} decreases with increasing frequency, while b_{\min} depends only on the electron velocity. For a given electron velocity, there is a cut-off in the power spectrum at frequency ω where $b_{\max}(\omega, v) = b_{\min}(v)$, or $h/mv = v/\omega$, or equivalently

$$\omega \lesssim \omega_{\max}(v) \sim mv^2/h. \quad (12.12)$$

This is very reasonable, since it states that the energy of the emitted photon had better be smaller than the total kinetic energy of the electron.

The above argument gives a useful approximation to the radiated power

$$\frac{dW}{d\omega dV dt} \simeq \begin{cases} \frac{16n_e n_i Z^2 e^6}{3c^3 m^2 v} & \text{for } \omega \ll mv^2/h \\ 0 & \omega \gg mv^2/h \end{cases} \quad (12.13)$$

It is conventional to encapsulate the details of a proper calculation in the ‘Gaunt-factor’ and write the power as

$$\frac{dW}{d\omega dV dt} = \frac{16n_e n_i Z^2 e^6}{3\sqrt{3}c^3 m^2 v} g_{\text{ff}}(v, \omega) \quad (12.14)$$

see e.g. R+L for more details.

Note that the low-frequency asymptotic power scales inversely with velocity. This is because while fast and slow electrons spend the same fraction of time being accelerated, the faster ones spread their radiation over a greater bandwidth.

12.4 Thermal Bremsstrahlung

Having computed the radiated power for a stream of electrons with a single velocity all that remains to compute the power radiated by an thermally equilibrated plasma is to average over a thermal (Maxwellian) distribution of velocities:

$$d^3p(v) = d^3v \exp(-mv^2/2kT) \propto v^2 \exp(-mv^2/2kT) dv \quad (12.15)$$

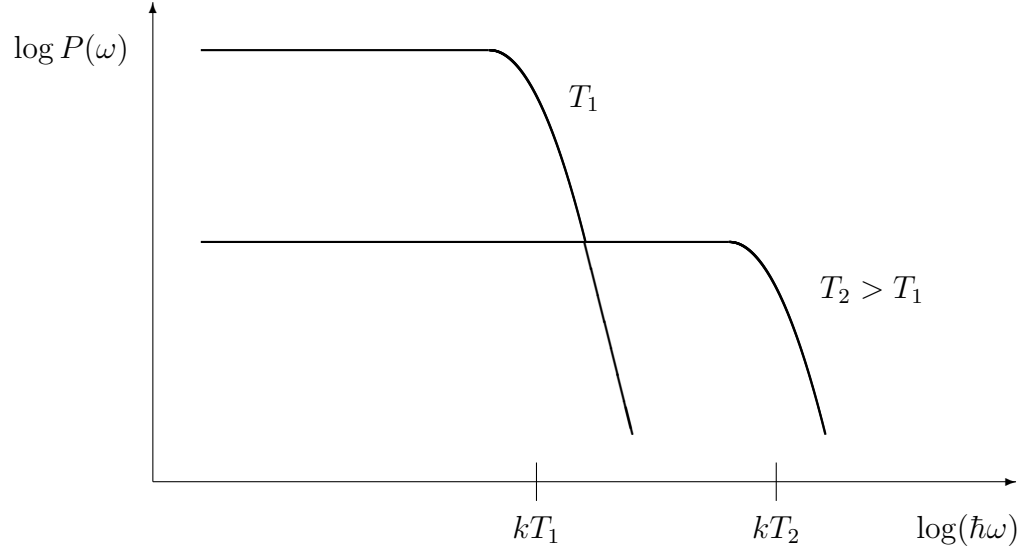


Figure 12.2: Spectra for thermal bremsstrahlung for two different temperatures (though assuming the same density).

To a rough approximation we can account for the cut-off by inserting the low-frequency asymptotic form for the single speed $dW/d\omega dV dt$ but limit the integration to $v > v_{\min}(\omega)$ such that $mv_{\min}^2/2 = \hbar\omega$ to give

$$\frac{dW}{d\omega dV dt} \simeq \frac{16n_en_iZ^2e^6}{3c^3m^2} \frac{\int_{v_{\min}} dv v e^{-mv^2/2kT}}{\int dv v^2 e^{-mv^2/2kT}} \sim \frac{Z^2n_en_ie^6}{(mc^2)^{3/2}} (kT)^{-1/2} e^{-\hbar\omega/kT}. \quad (12.16)$$

For low frequencies the emission scales inversely with square root temperature, consistent with the $1/v$ scaling above since the typical thermal velocity scales as \sqrt{T} .

Thermal bremsstrahlung spectra are sketched in figure 12.2.

Integrating this over frequency gives

$$\frac{dW}{dV dt} \simeq Z^2n_en_ie^6\sqrt{kT}(mc^2)^{-3/2} \quad (12.17)$$

so the bolometric emission scales as \sqrt{T} .

The emission scales as the square of the density.

The *free-free emissivity* for a plasma is then

$$\epsilon^{\text{ff}} = \frac{dW}{dt dV} = 1.4 \times 10^{-27} (T/\text{K})^{1/2} Z^2 n_e n_i \bar{g} \quad (12.18)$$

in cgs units. The appropriately averaged Gaunt factor here is very close to unity for realistic conditions.

This ‘thermal-bremsstrahlung’ spectrum is valid for sufficiently high temperatures, ie higher than the temperature corresponding to atomic transitions for the ions in question. At lower temperatures line emission becomes important.

12.5 Thermal Bremsstrahlung Absorption

The inverse reaction to that considered above results in *thermal bremsstrahlung absorption*, which can be obtained from the emissivity using Kirchoff’s law: $\alpha_\nu = j_\nu/B_\nu(T)$. The bremsstrahlung emissivity is asymptotically flat at low frequencies, whereas $B_\nu \propto \nu^2$, so the absorption is strongly frequency dependent $\alpha_\nu \propto 1/\nu^2$ and is therefore most effective at low frequencies.

12.6 Relativistic Bremsstrahlung

We now elucidate the general features of free-free emission from a plasma where the electron velocities are relativistic.

First, let's review the steps leading to the non-relativistic result.

- **1 electron — 1 ion.** In a collision, the electron sees an impulse with $\dot{u} \sim Ze^2/mv^2$ lasting a time $t_\star \sim b/v$ resulting in the emission of mean amount of energy $dW \sim e^2 \dot{u}^2 t_\star / c^3 \sim Z^2 e^6 / m^2 c^3 b^3 v$ with frequency $\omega \sim v/b$.
- **Energy cut-off.** The energy of the emitted quanta should not exceed kinetic energy of electron, so $h\nu \lesssim mv^2$ or equivalently $b > b_{\min} = h/mv$.
- **1 electron — many ions.** The frequency of collisions with impact parameter $< b$ is $dN/dt \sim n_i v b^2 \propto b^2$ whereas energy $dW \propto 1/b^3$, so most energy is emitted in collisions with $b \simeq b_{\min}$. The rate of such collisions is $dN/dt \sim n_i v b_{\min}^2$ and the power output of a single electron is therefore $dW/dt \sim n_i v Z^2 e^6 / mc^3 h$.
- **Many electrons — many ions.** The power per unit volume is $dW/dVdt = n_e dW/dt \sim n_e n_i v Z^2 e^6 / mc^3 h$.

We now generalize this argument to the relativistic case. The approach is to compute the emission in the rest frame of the electron — which sees the highly foreshortened and amplified electric field of a relativistic ion — and then transform back to the observer frame.

- **1 electron — 1 ion.** The electron sees the electric field of a rapidly moving ion as an impulse with $E \sim \gamma Ze^2/b^2$, so $\dot{u} \sim \gamma Ze^2/mv^2$, and lasting a time $t_\star \sim b/\gamma c$. The net energy radiated is therefore $dW \sim e^2 \dot{u}^2 t_\star / c^3 \sim \gamma Z^2 e^6 / m^2 c^4 b^3$ with frequency $\omega_\star \sim \gamma c/b$.
- **Energy cut-off.** The energy of the emitted quanta in the observer frame should not exceed the electron energy $\sim \gamma mc^2$, or, in the electron frame, $h\nu \lesssim mc^2$. This gives $b_{\min} \sim \gamma h/mc$.
- **1 electron — many ions.** The electrons see the volume occupied by the ions foreshortened, and therefore they see an oncoming stream of ions with $v \sim c$ and density γn_i (where n_i is the ion density in the observer frame or in rest frame of the ions). The rate of collisions with $b \sim b_{\min}$ is $dN/dt \sim \gamma n_i b_{\min}^2 c$ and therefore the power output in the electron frame is $dW/dt \sim \gamma n_i Z^2 e^6 / mc^2 h$.
- **Many electrons — many ions.** The 1-electron power dW/dt is Lorentz invariant, and electrons have space-density n_e , so the net power per unit volume is

$$\frac{dW}{dVdt} \sim \frac{\gamma n_i n_e e^6}{mc^2 h} \sim \frac{n_i n_e e^6 E}{h(mc^2)^2} \quad (12.19)$$

with most of the energy emitted at frequency $\hbar\omega \sim \gamma mc^2 = E$.

A characteristic property of relativistic Bremsstrahlung is that the emissivity is proportional to the electron energy, and therefore to the temperature T for thermalized plasma, as compared to emissivity $\propto \sqrt{T}$ in the non-relativistic case.

There is an interesting and illuminating alternative way to look at bremsstrahlung. We saw that the mean energy radiated in a collision (in the electron rest frame) is $dW' \sim \gamma Z^2 e^6 / (m^2 c^4 b^3)$ and that the energy in the observer frame is therefore $dW = \gamma dW'$. Now we can write this as

$$dW = \gamma^2 \left(\frac{Ze}{b^2} \right)^2 \left(\frac{e^2}{mc^2} \right)^2 b \quad (12.20)$$

but we recognize the first factor in parentheses as the square of the ion's coulomb field (and therefore on the order of the energy density of the ion's field at distance b) and the second factor as the square of the classical radius of the electron r_0^2 or, equivalently on the order of the Thomson cross section.

Thus, we can write the energy radiated as

$$dW \simeq \gamma^2 b \sigma_T U_{\text{field}} \quad (12.21)$$

with U_{field} the energy density of the external field (in the rest frame of the ion, that is). Crudely speaking, we can describe the bremsstrahlung process as though the electron, passing close to the ion, knocks out the energy density in a volume $b\sigma_T$ and in the process boosts this by a factor γ^2 .

As we shall see, one can express the radiation power from synchrotron and from Compton scattering in exactly the same way; it is as though the electron has size σ_T and impacts the ambient field — be it the field of an ion for bremsstrahlung, a magnetic field in the case of synchrotron emission, or a randomly fluctuating ambient radiation field in the case of Comptonization — and ejects it with a γ^2 energy boost. As we shall see when we discuss Compton scattering, this γ^2 boost factor is simply the result of a pair of Lorentz transforms.

12.7 Applications of Thermal Bremsstrahlung

12.7.1 Low Frequency Emission from Ionized Gas Clouds

Bremsstrahlung is important at low frequencies since the spectrum is flat, with $j_\nu \sim n^2 T^{-1/2}$. As discussed, at sufficiently low frequency the clouds may become optically thick to thermal bremsstrahlung absorption. The signature of such clouds is a spectrum with $I_\nu \propto \nu^2$ at very low frequencies flattening to $I_\nu \propto \nu^0$ at higher frequencies. This effect is seen in radio observations of HII regions.

12.7.2 Clusters of Galaxies

- Diffuse X-ray emission from very massive clusters of galaxies looks like thermal bremsstrahlung with $kT \simeq 10\text{keV}$. Lower mass clusters have bremsstrahlung-like continuum with iron lines superposed.
- The inferred temperature of $\sim 10^8\text{K}$ is consistent with hot gas in hydrostatic equilibrium in the same potential well depth as inferred from the virial theorem and the observed velocity dispersion of $\sigma_v \simeq 1000\text{km/s}$. For these values, the kinetic energy per unit mass is the same for a galaxy as for an electron-ion pair.
- Since emissivity scales as n^2 the surface brightness is the integral of n^2 along the line of sight $I \propto \int dl n^2$. This is known as the *emission measure*. This tends to be very centrally concentrated (as compared to the projected density of galaxies for instance); X-ray observations have become the preferred method for detecting distant clusters where optical searches become subject to confusion.
- Bremsstrahlung emission can become an effective cooling mechanism in centrally concentrated clusters. The net rate of energy loss is $dE/dt \propto E^{1/2}n$ so the cooling time is $t_{\text{cool}} \sim E/\dot{E} \propto E^{1/2}n^{-1}$. This has the consequences that cooling is most effective in the center of clusters and, for a given density, tends to be less effective in the hotter, more massive clusters.
- Many clusters have central cooling times on the order of the age of the Universe, and are reasonably well modeled as ‘cooling flows’ in which gas pressure support is removed from the gas at the center, and the outer atmosphere quasi-statically adjusts. However, a puzzle is where the gas which ‘drops out’ ends up. There are cooling flows with mass disappearing at the rate of several hundred solar masses per year, but this gas does not end up as luminous stars.

12.7.3 Bremsstrahlung from High Energy Electrons

Gamma ray emission is detected from our galaxy which is thought to arise from Bremsstrahlung from high energy electrons.

- The electrons are typically modeled as having a power-law distribution of energies.
- The radiative energy is carried by photons with $h\nu \simeq E_e$.
- Gamma-rays from the galaxy with energies in the range 30 – 100 MeV suggest an abundance of relativistic electrons with $\gamma \sim 100$.

12.8 Problems

12.8.1 Bremsstrahlung

In the following assume that the electron is non-relativistic in the ion frame but that $mv^2/2 \gg 1$ Rydberg.

- Using Larmor's formula (see problem 5), give an order of magnitude estimate for the energy dW radiated as an electron flies past a singly charged ion with velocity v and impact parameter b . What is the characteristic frequency of the radiation emitted?
- The foregoing is valid only for sufficiently soft encounters that the photon energy satisfies $\hbar\omega < m_e v^2/2$. Compute, for given v , the impact parameter b_{\min} such that this condition is marginally satisfied.
- Compute the frequency of collisions (per unit volume) with $b \sim b_{\min}$ for a beam of electrons with space density n_e passing through a cloud of ions with space density n_i , and (combining this with the answer from part a) compute the power per unit volume.
- Replacing the single speed v with the characteristic velocity for electrons with a thermal distribution of velocities at some temperature T , obtain an approximation to the thermal bremsstrahlung bolometric emissivity ϵ_{ff} in terms of n_i , n_e , T and fundamental constants.
- Estimate the mean number of photons generated per collision with $b \sim b_{\min}$. Express your answer in terms of the 'fine structure constant' $\alpha = q^2/\hbar c \simeq 1/137$.

Chapter 13

Synchrotron Radiation

Synchrotron radiation is generated by the acceleration of electrons spiraling around static magnetic fields. It is often called ‘non-thermal’ radiation (to distinguish it from emission from thermal electrons, rather than from black-body radiation).

13.1 Equations of Motion

The relativistic Lorentz force law gives the rate of change of the four-momentum. The time component is

$$\frac{dE}{dt} = \frac{d\gamma m}{dt} = -e\mathbf{v} \cdot \mathbf{E} = 0 \quad (13.1)$$

so the Lorentz factor γ , and therefore also the speed for the electron are constant in time.

The space component is

$$\frac{d\mathbf{P}}{dt} = \frac{d(\gamma m \mathbf{v})}{dt} = -e\mathbf{B} \times \mathbf{v}/c \quad (13.2)$$

or

$$m\gamma \frac{d\mathbf{v}}{dt} = -\frac{e}{c} \mathbf{B} \times \mathbf{v} \quad (13.3)$$

Note that t here is coordinate time.

The rate of change of \mathbf{v}_{\parallel} , the component of the velocity parallel to the magnetic field, vanishes, so $\mathbf{v}_{\parallel} = \text{constant}$.

The solution of the equations of motion for the component of the velocity perpendicular to the field correspond to circular motion. If the field is aligned with the z -axis

$$\mathbf{r}_{\perp}(t) = \begin{bmatrix} r_x(t) \\ r_y(t) \end{bmatrix} = r \begin{bmatrix} \cos \omega_B t \\ \sin \omega_B t \end{bmatrix} \quad \Rightarrow \quad \mathbf{v}_{\perp} = r\omega_B \begin{bmatrix} -\sin \omega_B t \\ \cos \omega_B t \end{bmatrix}. \quad (13.4)$$

This trajectory is a helix and the angular frequency of rotation about the field axis is

$$\omega_B = \frac{|\mathbf{v}_{\perp}|}{r} = \frac{eB}{\gamma mc} \quad (13.5)$$

which is known as the *relativistic (angular) gyro-frequency*.

For low velocities $v \ll c$, $\gamma \rightarrow 1$ and the gyro-frequency becomes independent of the particle energy. The (non-relativistic) gyro-frequency is

$$\omega_G \equiv \frac{eB}{mc}. \quad (13.6)$$

13.2 Total Power Radiated

The power radiated is, according to Larmor, proportional to the square of the proper acceleration $P = 2e^2 a_0^2 / 3c^3$, or, in terms of the coordinate acceleration in the observer frame

$$P = \frac{2e^2}{3c^3} \gamma^4 a_\perp^2 = \frac{2e^2}{3c^3} \gamma^4 \left(\frac{eB}{\gamma mc} \right)^2 v_\perp^2 \quad (13.7)$$

or, in terms of the classical radius of the electron $r_0 = q^2 / mc^2$,

$$P = \frac{2}{3} r_0^2 c \beta_\perp^2 \gamma^2 B^2. \quad (13.8)$$

An alternative path to this result is to transform the magnetic field into the instantaneous frame of rest of the electron. The electron sees an electric field and the power can be computed much as for Thomson scattering.

This is power in the rest-frame, but since dipole radiation is front-back symmetric, the radiated power is Lorentz invariant.

Equation (13.8) applies for a electrons of fixed ‘pitch angle’, defined such that $\sin \alpha = v_\perp / v$. Averaging (13.8) over pitch angle assuming an isotropic distribution gives

$$\langle P \rangle = \left(\frac{2}{3} \right)^2 r_0^2 c \beta^2 \gamma^2 B^2. \quad (13.9)$$

One can also express the power as

$$P = 2\sigma_T c U_{\text{mag}} \beta^2 \gamma^2 \sin^2 \alpha. \quad (13.10)$$

where $U_{\text{mag}} = B^2 / 8\pi$ is the magnetic field energy, and $\sigma_T = 8\pi r_0^2 / 3$ is the Thomson cross-section. Averaging over pitch angle and taking the highly relativistic limit,

$$\langle P \rangle = \frac{4}{3} c \sigma_T \gamma^2 U_{\text{mag}}. \quad (13.11)$$

13.3 Synchrotron Cooling

The energy loss rate for a relativistic electron is $dE/dt = P \propto E^2$. One can define a *cooling rate* \dot{E}/E and a corresponding *cooling time* $t_{\text{cool}} = E/\dot{E}$.

More precisely,

$$\frac{dE}{E^2} = -\frac{4\sigma_T U_{\text{mag}}}{3m^2 c^3} dt \quad (13.12)$$

which can be integrated to give

$$\frac{1}{E_f} - \frac{1}{E_i} = \frac{4\sigma_T U_{\text{mag}}}{3m^2 c^3} t. \quad (13.13)$$

This sets an upper limit to the electron energy as a function of the time since the electrons were injected. Even if the electrons were initially infinitely energetic they will have cooled to a finite temperature $E_{\text{max}}(t) = (3m^2 c^3 / 4\sigma_T U_{\text{mag}}) t^{-1}$ after time t and electrons of lower initial energy will have $E < E_{\text{max}}$.

Observing electrons of a certain energy in a given magnetic field then gives an upper limit to the age of the electrons (ie time since they were accelerated and injected).

13.4 Spectrum of Synchrotron Radiation

At low energies, $v \ll c$ one can compute the radiation from the spiraling electrons using the dipole formula, and one finds that the radiation is emitted at the gyration frequency.

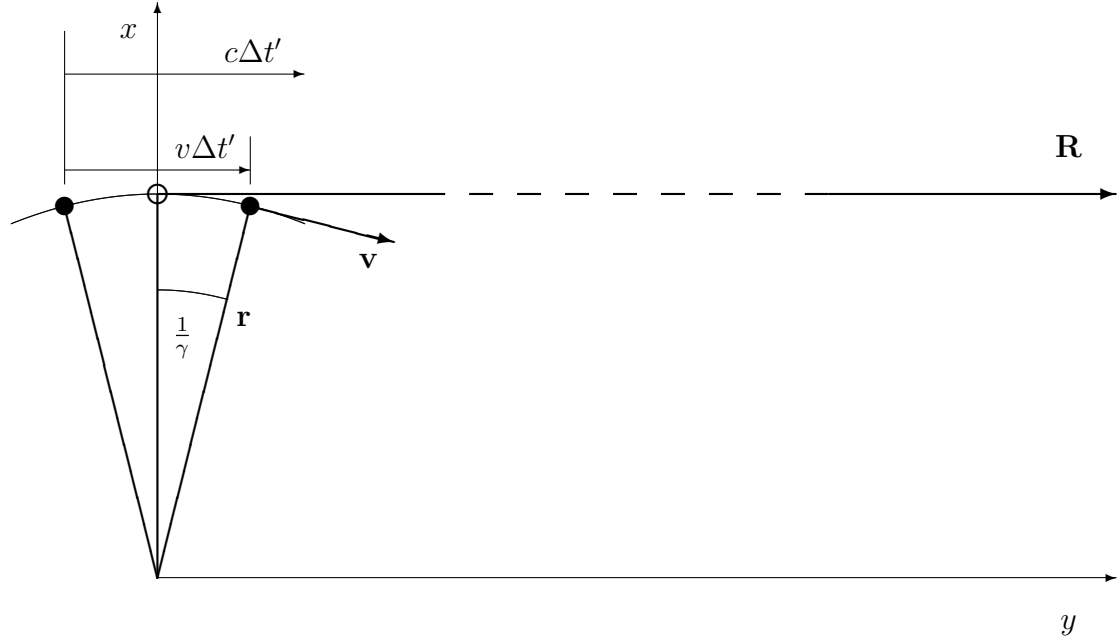


Figure 13.1: Because of relativistic beaming, a distant observer at **R** receives appreciable synchrotron radiation from the electron only for a small fraction ($\sim 1/\gamma$) of its orbit. The leading edge of the pulse is emitted as the particle enters the active zone at the left, and the trailing edge of the pulse is emitted a time $\Delta t' \sim 1/(\gamma\omega_B)$ later as the particle leaves the active zone at the right. The leading edge of the pulse has meanwhile propagated a distance $c\Delta t'$ whereas the particle has moved a distance $v\Delta t'$, so it has almost kept up with the leading edge. Consequently, the interval between reception of the pulse edges is on the order of $\Delta t = \Delta t' \times (1 - v/c)$ or, since $\Delta t' \sim 1/(\gamma\omega_B)$ and $(1 - v/c) \simeq 1/(2\gamma^2)$ we have $\Delta t \sim 1/(\gamma^3\omega_B)$. Thus the frequency of the radiation detected is larger than the orbital frequency by a factor $\sim \gamma^3$.

For large electron energies we expect the radiation to be strongly beamed along the direction of motion with opening angle for the beam $\Delta\theta \sim 1/\gamma$. An observer will then receive pulses of radiation of period $\tau = 2\pi/\omega_B$ but of duration $\Delta t \ll \tau$. In fact, the time-scale for the pulses is $\Delta t \sim \tau/\gamma^3$ and consequently the radiation emerges at frequency

$$\omega_c \simeq \gamma^3 \omega_B. \quad (13.14)$$

This can be understood qualitatively as follows. The beaming effect means that a given observer will see radiation from the particle only for a small fraction $\sim 1/\gamma$ of its orbit ie for $|\theta| \lesssim 1/\gamma$ in figure 13.1. This is when it is moving almost directly towards the observer, and consequently there is a big Doppler effect: the emission of the leading edge of the pulse precedes the emission of the trailing edge by a coordinate time interval $\Delta t' \sim 1/(\omega_B\gamma)$, but the latter event is closer to the observer by an amount $\Delta r = v\Delta t'$, so the interval between the reception of the leading and trailing edges is

$$\Delta t = (1 - \beta)\Delta t' \sim (1 - \beta)/(\omega_B\gamma) \sim 1/(\omega_B\gamma^3) \quad (13.15)$$

where we have used $(1 - \beta) \simeq 1/(2\gamma^2)$ for $\gamma \gg 1$.

Thus, we expect to receive radiation up to frequencies at most of order ω_c given by (13.14), which is often called the *critical frequency*.

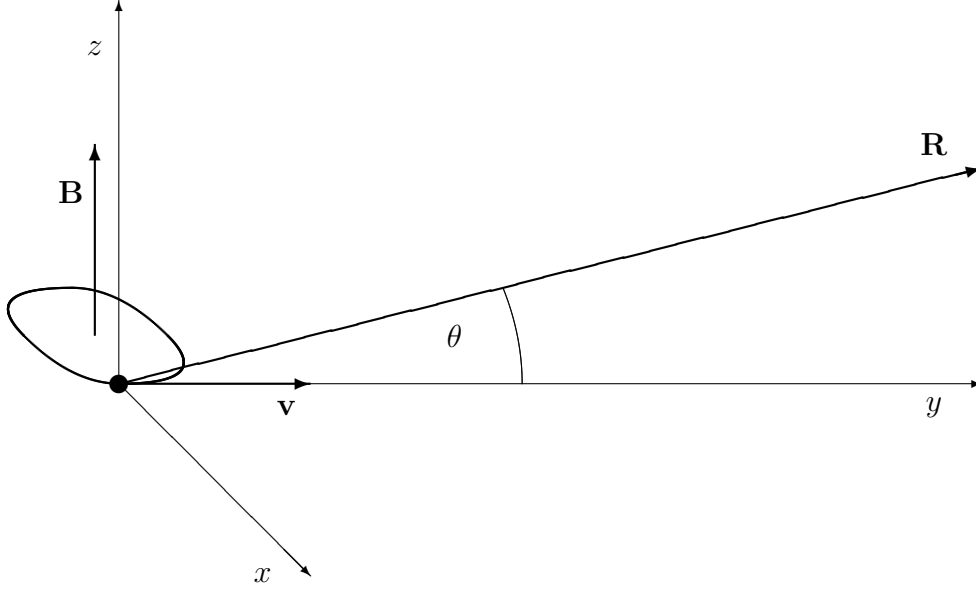


Figure 13.2: Geometry for calculation of the spectrum of synchrotron radiation. An electron orbits in the $x - y$ plane with velocity \mathbf{v} . The vector \mathbf{R} points to the distant observer.

13.4.1 Pulse Profile

This result can also be obtained using the Lienard-Wiechart potentials, and this also allows us to infer the shape of the spectrum of synchrotron radiation.

For an observer at great distances the magnetic potential is

$$\mathbf{A}(\mathbf{R}, t) = \frac{q}{c} \left[\frac{\mathbf{v}}{(1 - \mathbf{n} \cdot \mathbf{v}/c)|\mathbf{R} + \mathbf{r}|} \right]_{\text{ret}} \simeq \frac{q}{cR} \left[\frac{\mathbf{v}(t')}{1 - \mathbf{n} \cdot \mathbf{v}(t')/c} \right] \quad (13.16)$$

where $\mathbf{r}(t')$ is the trajectory of the electron, $\mathbf{v} \equiv \dot{\mathbf{r}}$, $\mathbf{n} = \hat{\mathbf{R}}$ and where the retarded time t' is the solution of

$$t' = t - R/c + \mathbf{n} \cdot \mathbf{r}(t')/c \quad (13.17)$$

where R is now considered constant.

For the geometry shown in figure 13.2 we have $\mathbf{r} = r(\cos \omega_B t', \sin \omega_B t', 0)$ and $\mathbf{n} = (0, \cos \theta, \sin \theta)$ so $\mathbf{n} \cdot \mathbf{r}(t') = r \cos \theta \sin \omega_B t'$. The relation between observer time t and retarded time t' is then

$$t - R/c = t' - \frac{\beta}{\omega_B} \cos \theta \sin \omega_B t'. \quad (13.18)$$

This function is plotted in figure 13.3.

The ‘Doppler factor’ appearing in the LW potentials is

$$\kappa = 1 - \hat{\mathbf{R}} \cdot \mathbf{v} = \frac{dt}{dt'} = 1 - \sqrt{1 - (1/\gamma)^2} \cos \theta \cos \omega_B t' \quad (13.19)$$

where we have used $\beta = \sqrt{1 - (1/\gamma)^2}$. Evidently, the factor κ becomes very small — and consequently the potential becomes very large — provided $\sqrt{1 - 1/\gamma^2}$, $\cos \theta$ and $\cos \omega_B t'$ are all very close to unity. Equivalently, a small κ requires that $1/\gamma$, θ and $\omega_B t'$ are all very small compared to unity. If so,

$$\kappa \simeq \frac{1}{2} [(1/\gamma)^2 + \theta^2 + \omega_B^2 t'^2] \quad (13.20)$$

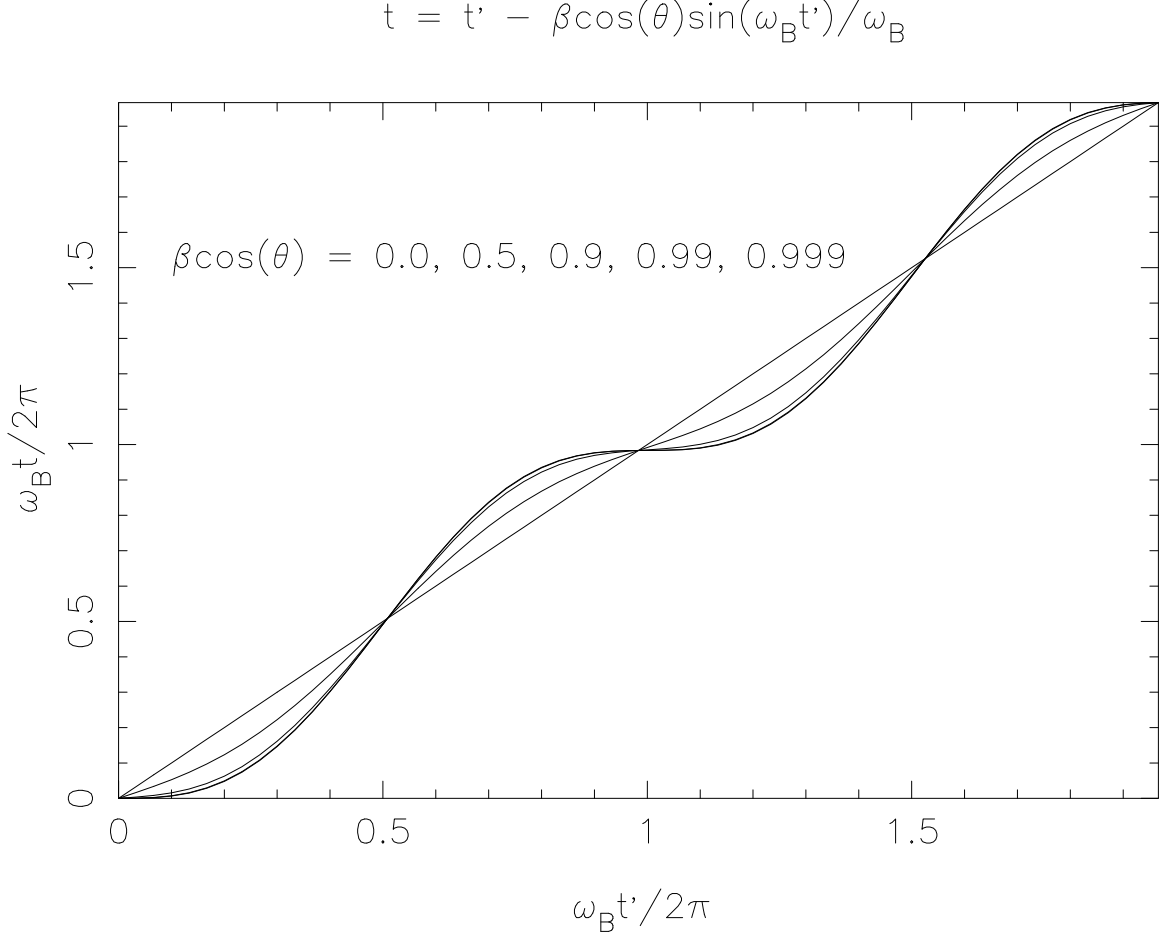


Figure 13.3: Observer time as a function of retarded times as given by (13.18). Two complete rotations are shown, for various values of $\beta \cos \theta$. At high velocities and small angles, observer time becomes almost stationary with respect to retarded time. Consequently, the function κ becomes very small, and the potentials and fields become very strong, resulting in a short pulse of radiation. It appears that the function tends to a well defined limit as $\beta \cos \theta \rightarrow 1$. However, this is somewhat misleading; if we examine in detail the nearly stationary region, we find that the behaviour is very sensitive to how close $\beta \cos \theta$ is to unity.

where we have expanded the trigonometric functions in (13.19).

Performing the analogous expansion on (13.18) gives

$$t - R/c \simeq \frac{1}{2}(\gamma^{-2} + \theta^2)t' \left(1 + \frac{1}{3} \frac{\omega_B^2 t'^2}{\gamma^{-2} + \theta^2} \right). \quad (13.21)$$

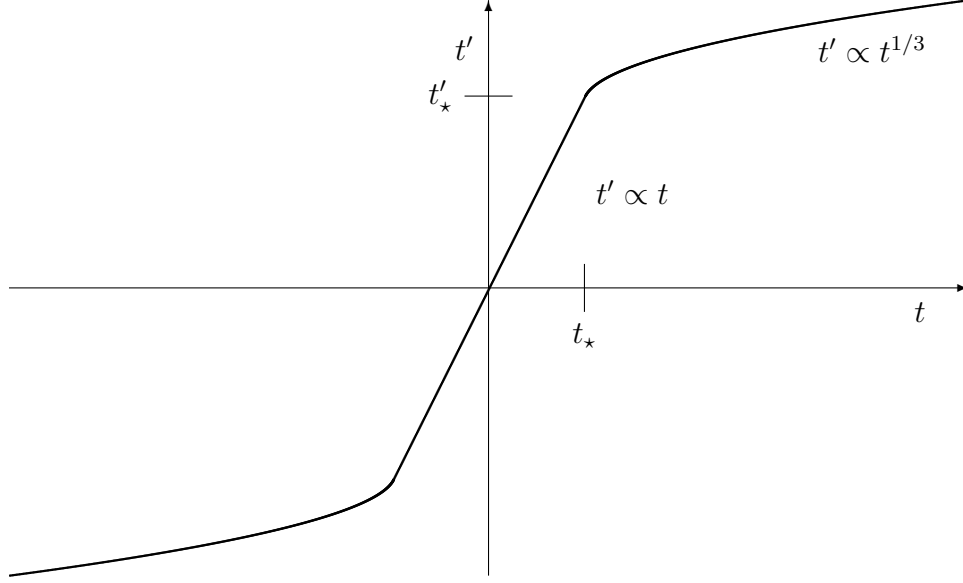
from which we see that there is a characteristic time-scale

$$t'_*(\gamma, \theta) = \sqrt{\gamma^{-2} + \theta^2} / \omega_B. \quad (13.22)$$

For $t' \ll t'_*$ we have a linear relation $t \propto t'$ whereas for $t' \gg t'_*$, $t \propto t'^3$. More specifically,

$$t \sim \begin{cases} (\gamma^{-2} + \theta^2)t' & t' \ll t'_* \\ \omega_B^2 t'^3 & t' \gg t'_* \end{cases} \quad (13.23)$$

The time-scale t'_* corresponds to a characteristic observer time interval $t_* = (\gamma^{-2} + \theta^2)^{3/2} / \omega_B$. Note that for $\theta = 0$ (or, more generally, $\theta \lesssim 1/\gamma$) this is just the inverse of the critical frequency: $t_* \simeq 1/\omega_c$.

Figure 13.4: Relation between retarded time t' and observer time t .

We can turn (13.23) around to obtain the asymptotic form for the retarded time as a function of observer time:

$$t' \simeq \begin{cases} t/(\gamma^{-2} + \theta^2) & t \ll t_* \\ (t/\omega_B^2)^{1/3} & t \gg t_* \end{cases} \quad (13.24)$$

as as sketched in figure 13.4.

We can now compute, for example, the x -component of the potential A_x . The numerator in (13.16) then contains the x -component of the velocity $v_x = v \sin \omega_B t' \simeq v \omega_B t'$ since we can safely assume that $t' \ll 1/\omega_B$. The potential is therefore given by

$$A_x(t) = \frac{qv\omega_B t'}{cR} \frac{2}{(1/\gamma)^2 + \theta^2 + \omega_B^2 t'^2} \quad (13.25)$$

with t' given by (13.21). This has the asymptotic behaviour

$$A_x \simeq \frac{qv}{cR} \times \begin{cases} 2\omega_B t/(\gamma^{-2} + \theta^2)^2 & t \ll t_* \\ 2(\omega_B t)^{-1/3} & t \gg t_* \end{cases} \quad (13.26)$$

as sketched in figure 13.5.

The magnetic field is $\mathbf{B} = \nabla \times \mathbf{A}$. At large distances, and for our geometry, this gives $B_z = (1/c)\partial A_x/\partial t$. The field is a time symmetric pulse with

$$B \simeq \frac{qv}{c^2 R} \times \begin{cases} 2\omega_B/(\gamma^{-2} + \theta^2)^2 & t \ll t_* \\ -(2/3)\omega_B^{-1/3} t^{-4/3} & t \gg t_* \end{cases} \quad (13.27)$$

The form of the field is shown in figure 13.6. The field in the negative ‘wings’ falls off rapidly, and the contribution to the net energy flux from $t \gg t_*$ is small. The characteristic frequency in the spectrum of a single pulse from an electron of Lorentz factor γ as viewed by an observer along a direction at angle θ out of the orbital plane is therefore

$$\omega_* = 1/t_* \simeq \frac{\omega_B}{(\gamma^{-2} + \theta)^{3/2}}. \quad (13.28)$$

Since there are no features in the potential or the field on scales smaller than t_* the power falls rapidly for $\omega \gg 1/t_*$.

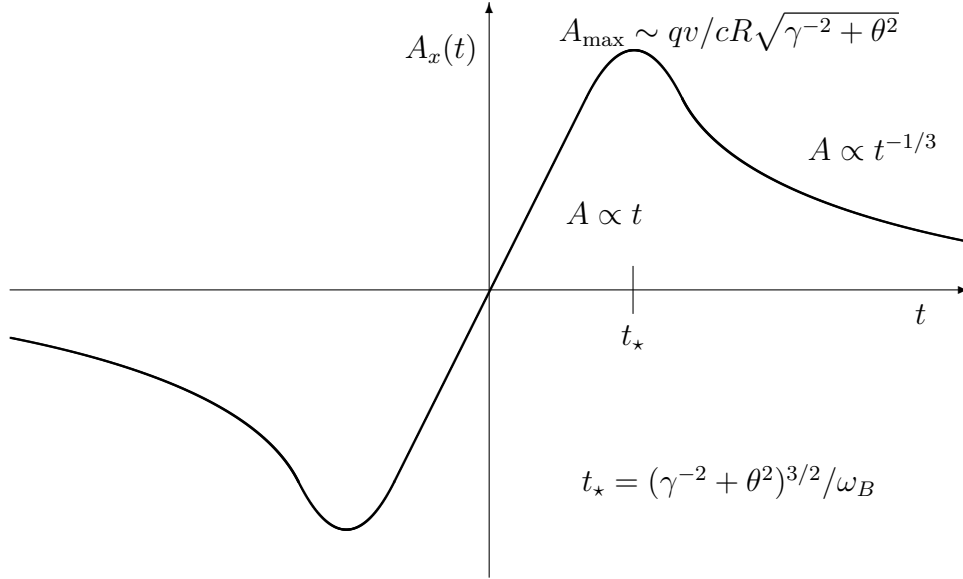


Figure 13.5: Sketch of the form of the potential for a pulse of synchrotron radiation from a highly relativistic electron with Lorentz factor γ as seen by an observer lying an angle θ out of the plane of the orbit.

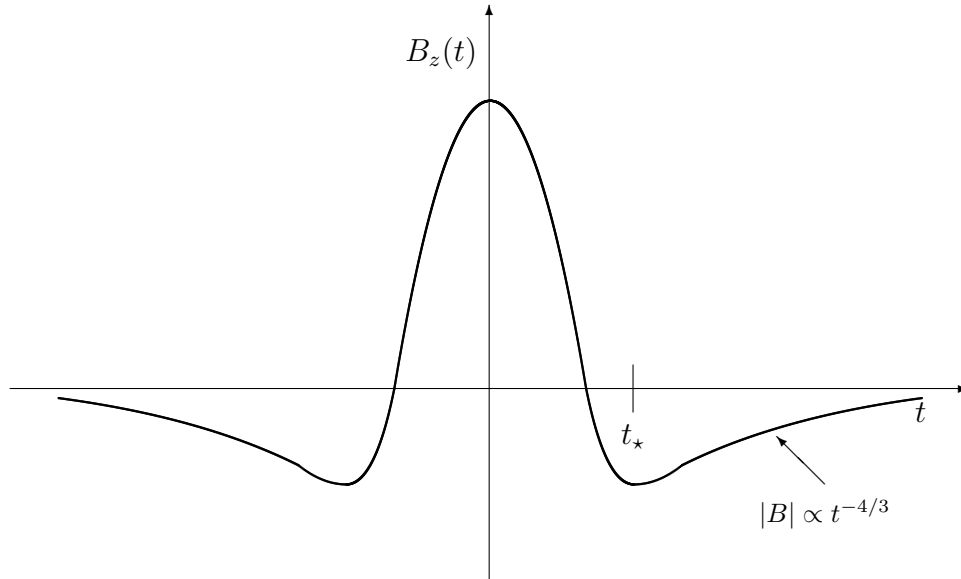


Figure 13.6: Schematic profile of the magnetic (or electric) field for a pulse of synchrotron radiation.

13.4.2 Low-Frequency Power Spectrum

While there is relatively little power at $\omega \ll \omega_*$ it is still of some interest to obtain the form of the power spectrum. The low-frequency power spectrum is dominated by the wings in the pulse. We can model these as

$$B(t, \theta) \simeq \frac{qv}{c^2 R} \omega_B^{-1/3} t^{-4/3} \quad \text{for } t \gtrsim t_* \quad (13.29)$$

where $t_* \simeq \theta^3 / \omega_B$. The transform of the field is then

$$B_\omega(\theta) = \int dt B(t, \theta) e^{i\omega t} \simeq \frac{qv}{c^2 R} \left(\frac{\omega}{\omega_B} \right)^{1/3} \quad \text{for } \omega \lesssim \omega_* \quad (13.30)$$

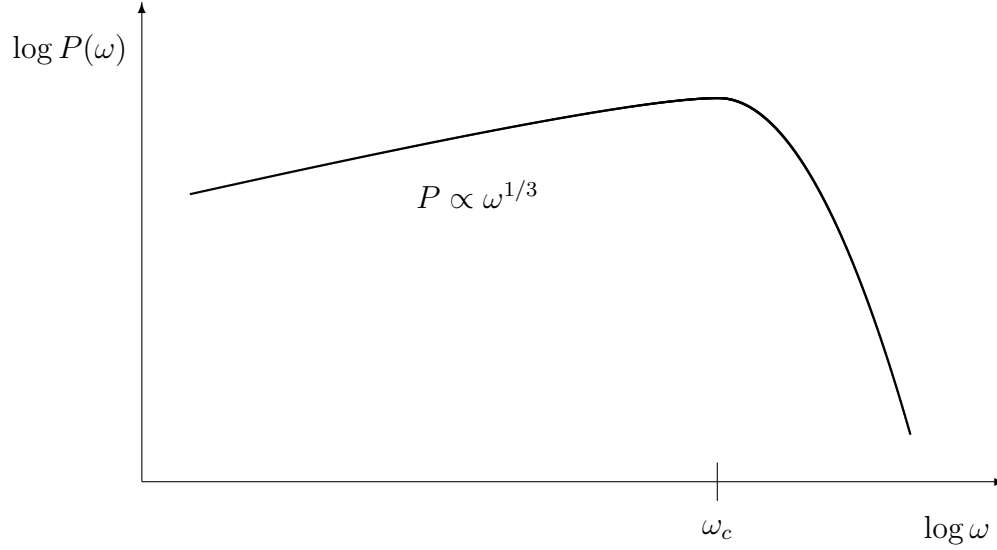


Figure 13.7: Sketch of the form of the power spectrum of synchrotron radiation for mono-energetic electrons.

with rapid attenuation at higher frequencies $\omega \gg \omega_*$. We have used here $\int dt t^{-4/3} e^{i\omega t} = \omega^{1/3} \int dy y^{-4/3} e^{iy}$ where the latter integral is some number of order unity.

It would appear then that the power spectrum — which is proportional to B_ω^2 — is $P(\omega) \propto \omega^{2/3}$. This is true for an observer at a specific angle relative to the orbit plane. However, in general, we have a distribution of pitch angles, so we really want to integrate over the distribution. Equivalently, we want to integrate over possible angles for the observer. Then we need to allow for the fact that the cut-off frequency ω_* is angle dependent. As we will now show, this results in more observed power at low frequencies.

The total energy emitted in a single gyration is obtained by integrating the Poynting flux:

$$\Delta W = R^2 \int d\theta \int dt \frac{cB^2(t, \theta)}{4\pi} = \frac{cR^2}{8\pi^2} \int d\omega \int d\theta B_\omega^2(\theta) \quad (13.31)$$

by Parseval's theorem. The cut-off frequency is $\omega_* = \omega_B / (1/\gamma^2 + \theta^2)^{3/2}$. At some observed frequency ω , an observer will see appreciable radiation only if the angle θ is less than some θ_{\max} such that $\omega_*(\theta_{\max}) = \omega$. If $\omega \ll \omega_c$ this means that $\theta_{\max}(\omega) \simeq (\omega_B/\omega)^{1/3}$. At larger angles, $\omega_* < \omega$ and the radiation is suppressed. To the level of sophistication that we are working, we can then replace the $\int d\theta B_\omega^2(\theta)$ by θ_{\max} times the asymptotic expression (13.30), to give

$$\Delta W \simeq \int d\omega \frac{q^2 v^2}{c^3 R^2} \left(\frac{\omega}{\omega_B} \right)^{1/3}. \quad (13.32)$$

The total power is $P = \omega_B \Delta W / 2\pi$, since the pulses occur with frequency $\omega_B / 2\pi$, so we have $P = \int d\omega P(\omega)$ where the power spectrum is

$$P(\omega) \sim \frac{q^2 v^2 \omega_B^{2/3} \omega^{1/3}}{c^3}. \quad (13.33)$$

The low-frequency power-spectrum is therefore a power-law: $P(\omega) \propto \omega^{1/3}$. The general form for the synchrotron power spectrum for electrons of a single energy is sketched in figure 13.7.

Note that $\omega_B = qB/\gamma mc = qBc/E$. Thus a highly relativistic electron will have the same low-frequency power as a highly relativistic proton of the same frequency; the low-frequency power

depends only on the *energy* of the particle, and not its velocity (assuming $\gamma \gg 1$ at least). The high-frequency cut-off at $\omega_* = \gamma^3 \omega_B \simeq E^2 e B / m^3 c^5$ on the other hand, depends also on the *mass* of the particle. This is because the cut-off, unlike the low- ω power, is critically dependent on how close the particle velocity is to the speed of light.

Finally, a puzzle: Compare the spectrum obtained here with the low-energy spectrum from bremsstrahlung. In that case we argued that a collision will produce a pulse of radiation, and so the low-frequency spectrum should be flat, $P(\omega) \propto \omega^0$. Here we also have a pulse of radiation. Why do we not find a flat, low- ω power spectrum? The answer lies in the shape of the pulse. For bremsstrahlung, the time integral of the field $\int dt B(t)$ is non-zero, and the transform of the field at low frequencies is constant. Here the integral of the pulse vanishes: $\int dt B(t) = 0$ and there is no analogous flat-spectrum component.

13.5 Power-Law Electrons

To obtain a realistic synchrotron spectrum we need to convolve the mono-energetic electron spectrum derived above with the energy distribution function for the electrons.

An interesting model is where the electrons have a power law distribution in energy:

$$n(\gamma) d\gamma = C \gamma^{-p} d\gamma. \quad (13.34)$$

The synchrotron power spectrum will then be a superposition of copies of the spectrum for mono-energetic electrons derived above, with appropriate scaling of amplitude and frequency.

To find the form of the composite power spectrum, we can argue as follows: The number of electrons in a logarithmic interval of γ is

$$dn \sim C \gamma^{1-p} d \ln \gamma \quad (13.35)$$

and the power radiated by a single electron is $P \sim \gamma^2 c \sigma_T N^2 \propto \gamma^2$ and appears at frequency $\omega \sim \gamma^2 \omega_G$. This means that frequency ω corresponds to energy $\gamma = \sqrt{\omega/\omega_G}$ and therefore the power radiated by the electrons with energy $\sim \gamma$ is

$$dP = \omega P(\omega) d \ln \omega \propto \gamma^{3-p} d \ln \gamma \propto \omega^{(3-p)/2} d \ln \omega \quad (13.36)$$

It then follows that the composite power spectrum is a power-law

$$P_{\text{tot}}(\omega) \propto \omega^{-s} \quad (13.37)$$

with *spectral index* $s = (p - 1)/2$.

Such models give a reasonable description of emission from radio-galaxies, which typically have power-law-like spectra extending over a substantial range of frequencies with spectral index $s \sim 0.8$.

Radio galaxies also often display a cut-off at low frequencies due to synchrotron self-absorption. See R+L for further discussion.

Chapter 14

Compton Scattering

Compton scattering is the generalization of Thomson scattering to allow for the recoil effect if the photon energy is not completely negligible compared to the electron rest mass.

For Thomson scattering we found that the scattered radiation had the same frequency as the incident radiation, so the energy of the quanta are unchanged, and we obtained the differential cross-section

$$\frac{d\sigma}{d\Omega} = \frac{1}{2} r_0^2 (1 + \cos^2 \theta) = \frac{3\sigma_T}{16\pi} (1 + \cos^2 \theta) \quad (14.1)$$

where θ is the angle between the directions of the incoming and scattered photons.

The generalization of this involves two modifications

- Recoil of the electron — this can be understood from simple relativistic kinematics.
- The cross-section is modified (the Klein-Nishina formula) if the photon energy in the rest frame of the electron exceeds the electron rest mass energy. This requires quantum electrodynamics.

14.1 Kinematics of Compton Scattering

Suitable null 4-vectors to represent the initial and final photon 4-momenta are

$$\vec{P}_{\gamma i} = \frac{\epsilon}{c} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \vec{P}_{\gamma f} = \frac{\epsilon_1}{c} \begin{bmatrix} 1 \\ \cos \theta \\ \sin \theta \cos \phi \\ \sin \theta \sin \phi \end{bmatrix} \quad (14.2)$$

where ϵ denotes the energy, the subscript 1 denotes the outgoing photon state, (ie after one scattering) and we have chosen the initial photon have momentum parallel to the x -axis.

Similarly, the 4-momenta for the initial and final electron states are

$$\vec{P}_{ei} = \begin{bmatrix} mc \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \vec{P}_{ef} = \begin{bmatrix} E/c \\ P_x \\ P_y \\ P_z \end{bmatrix} \quad (14.3)$$

where we are working in the rest-frame of the initial electron. These 4-momenta are illustrated in figure 14.1.

Conservation of the total 4-momentum is

$$\vec{P}_{\gamma i} + \vec{P}_{ei} = \vec{P}_{\gamma f} + \vec{P}_{ef}. \quad (14.4)$$

If we specify the incoming momenta \vec{P}_{ei} and $\vec{P}_{\gamma i}$ then the outgoing 4-momenta contain six free parameters, ϵ_1 , θ and ϕ for the photon and \mathbf{P}_{ef} for the electron (with the electron energy then fixed

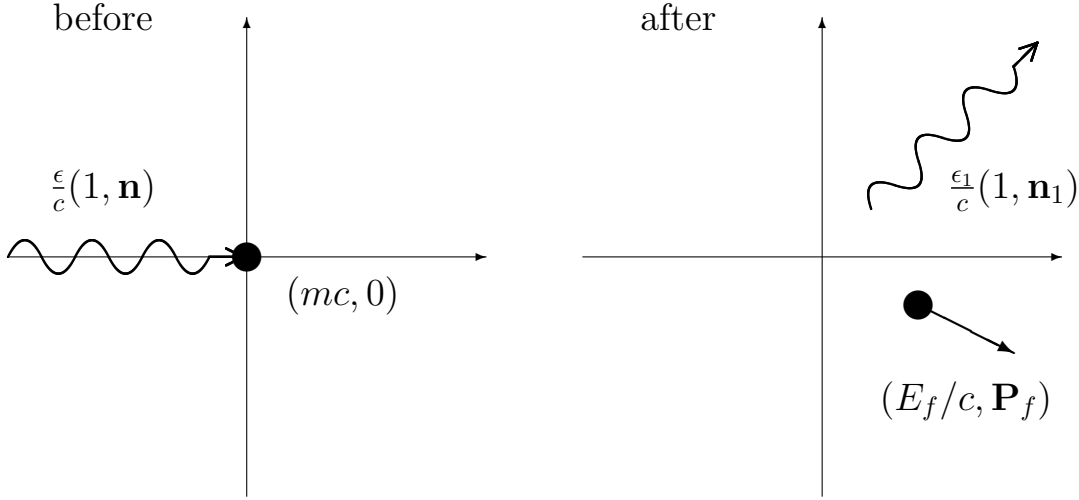


Figure 14.1: Four-momenta of particles involved in a Compton scattering event, working in a frame such that the electron is initially at rest and the initial photon direction is $\mathbf{n} = (1, 0, 0)$. The final photon direction is $\mathbf{n}_1 = (\cos \theta, \sin \theta \cos \phi, \sin \theta \sin \phi)$.

by the mass-shell condition $E^2 = p^2 c^2 + m^2 c^4$). If we specify the direction θ, ϕ of the outgoing photon say, then equation (14.4) provides us with the necessary four constraints to fully determine the collision (ie the energy of the photon and the 3-momentum of the outgoing electron).

If we simply want to determine the energy of the outgoing photon ϵ_1 , then we only need one equation. A convenient way to throw out the unwanted information \mathbf{P}_{ef} is to take the norm of \vec{P}_{ef} . If we orient our spatial coordinate system $\phi = 0$, so the outgoing photon momentum lies in the $x-y$ plane, then \vec{P}_{ef} is

$$\vec{P}_{ef} = P_{\gamma i} + \vec{P}_{ei} - \vec{P}_{\gamma f} = \frac{1}{c} \begin{bmatrix} \epsilon + mc^2 - \epsilon_1 \\ \epsilon - \epsilon_1 \cos \theta \\ \epsilon_1 \sin \theta \\ 0 \end{bmatrix} \quad (14.5)$$

and the mass-shell requirement $E_{ef}^2 = c^2 |\mathbf{P}_{ef}|^2 + m^2 c^4$ becomes

$$(\epsilon + mc^2 - \epsilon_1)^2 = (\epsilon - \epsilon_1 \cos \theta)^2 + (\epsilon_1 \sin \theta)^2 + m^2 c^4. \quad (14.6)$$

Which is a single equation one can solve for ϵ_1 given ϵ and θ . Expanding out the products and reordering gives

$$\epsilon_1 = \frac{\epsilon}{1 + \frac{\epsilon}{mc^2}(1 - \cos \theta)} \quad (14.7)$$

and expressing the photon energies in terms of wavelength $\epsilon = h\nu = hc/\lambda$ gives

$$\lambda_1 - \lambda = \lambda_c(1 - \cos \theta) \quad (14.8)$$

where the parameter

$$\lambda_c \equiv \frac{h}{mc} \quad (14.9)$$

is the *Compton wavelength*.

Equations (14.7,14.8) describe the energy loss for photons scattering off stationary electrons. They show that the collision is effectively elastic (ie $\epsilon_1 \simeq \epsilon$) if $\epsilon \ll mc^2$.

14.2 Inverse Compton Effect

Also of interest is the change in energy of photons scattering off *moving* electrons. This can be obtained by a) Lorentz transforming the photon 4-momentum to obtain its value in the electron rest frame b) computing the energy change of the photon as above and c) Lorentz transforming the outgoing photon 4-momentum back to the ‘laboratory’ or observer frame.

Let us take the electron to be moving in the x -direction and take the initial photon 4-momentum to be $\vec{P}_{\gamma i} = \epsilon(1, \cos \theta, \sin \theta, 0)$ then transforming to the rest frame (primed quantities) gives

$$\begin{bmatrix} \epsilon' \\ \epsilon' \cos \theta' \\ \epsilon' \sin \theta' \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma & & \\ -\beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} \epsilon \\ \epsilon \cos \theta \\ \epsilon \sin \theta \\ 0 \end{bmatrix} = \begin{bmatrix} \epsilon\gamma(1 - \beta \cos \theta) \\ \epsilon\gamma(\beta - \cos \theta) \\ \epsilon \sin \theta \\ 0 \end{bmatrix} \quad (14.10)$$

so the rest frame incoming photon energy is

$$\epsilon' = \epsilon\gamma(1 - \beta \cos \theta). \quad (14.11)$$

This photon will scatter to an outgoing photon with energy $\epsilon'_1 \simeq \epsilon'$ provided $\epsilon' \ll mc^2$, and with some direction θ'_1, ϕ'_1 . Applying the inverse Lorentz boost to get back to the lab-frame gives

$$\epsilon_1 = \epsilon'_1 \gamma (1 + \beta \cos \theta'_1) \quad (14.12)$$

and so the initial and final energies in the lab-frame are related by

$$\epsilon_1 = \epsilon\gamma^2(1 + \beta \cos \theta'_1)(1 - \beta \cos \theta). \quad (14.13)$$

The reason for writing the energy change in this seemingly awkward way — with one angle in the rest-frame system and one in the lab-frame system — is that both θ and θ'_1 have a broad distributions. The angle θ is the distribution of incoming angles in the lab-frame, and the distribution of $\mu = \cos(\theta)$ is flat. The angle θ'_1 is the direction of the scattered photon in the electron-frame. This is not isotropic, but still has a broad distribution. In contrast, for high energy electrons, both θ' and θ_1 have very narrow distributions as illustrated in figure 14.2.

It then follows that for typical collisions, the photon energy is boosted by a factor $\sim \gamma^2$:

$$\epsilon_1 \sim \gamma^2 \epsilon. \quad (14.14)$$

The only exceptions to this rule are for incoming photons with $\theta \simeq 0$ (ie propagating in the same direction as the electron) or if the outgoing photon has $\theta'_1 \simeq \pi$ (ie the scattered photon direction is opposite to the electron velocity), but these are special cases.

This result can also be understood in terms of ‘beaming’ (see figure 14.2). Consider some isotropic or nearly isotropic photon gas and a rapidly moving electron with $\mathbf{v} = \beta\hat{\mathbf{x}}$. Boosting into the electron frame we find that the electron sees a highly anisotropic radiation field, with most of the photons having momenta parallel to the direction $-\hat{\mathbf{x}}$. These photons get scattered approximately isotropically in the electron frame, and boosting back to the laboratory frame we find that the outgoing photons are tightly beamed along the $+\hat{\mathbf{x}}$ direction.

This process is an extremely efficient way to boost low energy photons to high energies, and (since the phrase ‘Compton scattering’ was initially used to describe the *loss* of photon energy in colliding off cold electrons) is called *inverse Compton scattering*.

The results above are valid only if $\epsilon' \ll mc^2$, which means it is restricted to $\epsilon_1 \ll \gamma mc^2$.

14.3 Inverse Compton Power

Imagine a cloud containing hot electrons and radiation being scattered. What is the *rate* at which energy is given to the radiation field by inverse compton scattering off the rapidly moving electrons?

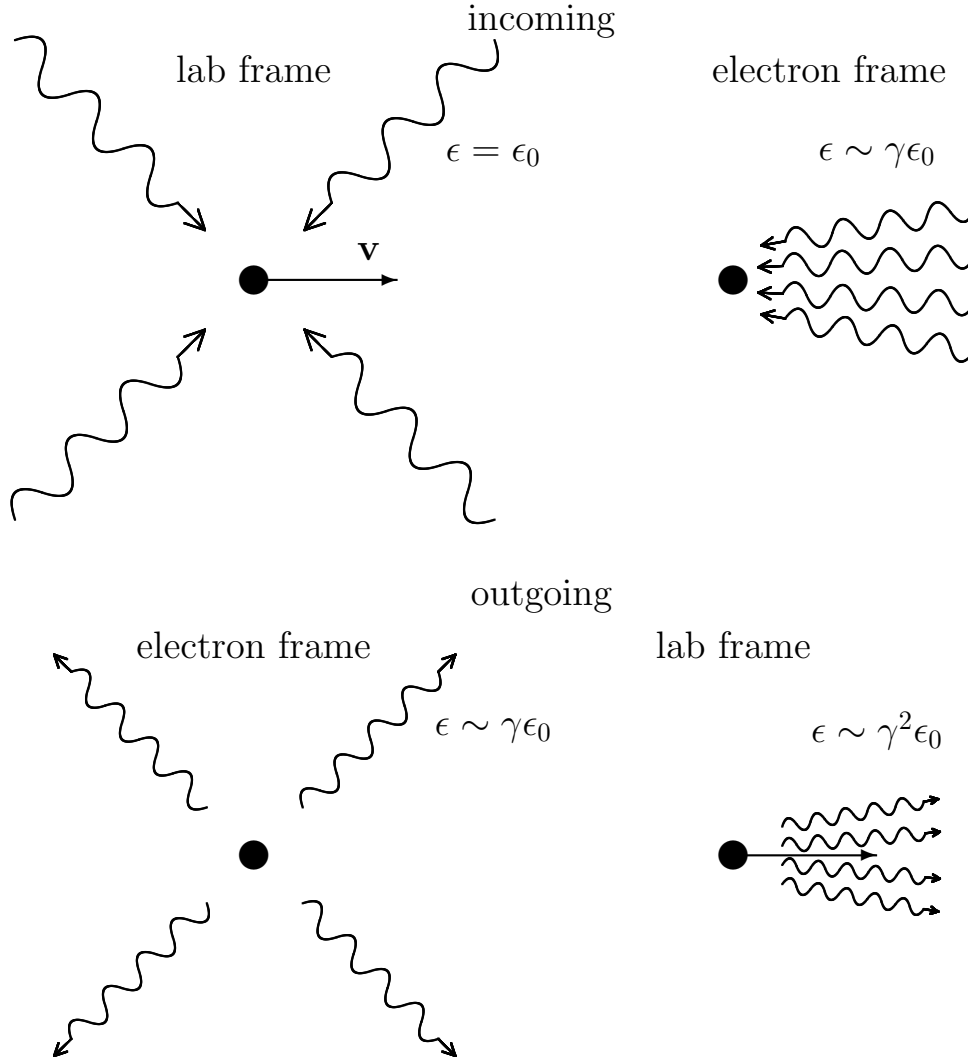


Figure 14.2: The γ^2 energy boost factor can be understood in terms of relativistic beaming. Upper left shows an electron moving with velocity \mathbf{v} in the lab frame and incoming photons which are isotropically distributed. Upper right panel shows the incoming quanta in the rest-frame of the electron. They are now highly anisotropic, so the electron sees them as nearly ‘head-on’ and their typical energies are boosted by a factor $\sim \gamma$. Lower left shows the photons after scattering. They are now approximately isotropic in the electron frame, and have roughly the same energy as they had before being scattered. Lower right panel shows the scattered photons in the lab-frame. They are now again highly collimated, and their typical energies have been boosted by a further factor $\sim \gamma$, so in the lab-frame the overall typical energy is boosted by a factor γ^2 .

To answer this we transform to get the energy density u' in the electron rest frame and then use the Thomson cross-section to give the power scattered as $P' = c\sigma_T u'$, which is Lorentz invariant since the scattered radiation is front-back symmetric in the electron frame, and which therefore also gives the rate at which energy is scattered into the radiation field in the lab-frame.

To make the transformation, recall that one can write the energy density for radiation in range

of frequencies $d\omega$ and range of directions $d\Omega$ in terms of the phase-space density $f(\mathbf{r}, \mathbf{p})$ as

$$u_\nu(\Omega) d\nu d\Omega = f(\mathbf{r}, \mathbf{p}) E d^3p = E^2 f(\mathbf{r}, \mathbf{p}) \frac{d^3p}{E} = h^2 \nu^2 f(\mathbf{r}, \mathbf{p}) \frac{d^3p}{E} \quad (14.15)$$

with $E = h\nu$, $p = h\nu/c$ and $d^3p = p^2 dp d\Omega = (h/c)^3 \nu^2 d\nu d\Omega$. But $f(\mathbf{r}, \mathbf{p})$ (which here is actually independent of \mathbf{r}) and d^3p/E are both Lorentz invariant quantities, so we have

$$u'_{\nu'}(\Omega') d\nu' d\Omega' = (\nu'/\nu)^2 u_\nu(\Omega) d\nu d\Omega. \quad (14.16)$$

The reason for two powers of ν'/ν here is that the energy $E = h\nu$ of each photon gets boosted by one power of ν'/ν and the rate at which photons pass by an observer also increases by the same factor (if photons are passing at a rate of ϵ photons per period of the radiation in one frame then they are passing by at the rate of ϵ photons per period in all frames).

Now the frequency factor is simply given by the Doppler formula $\nu'/\nu = \gamma(1 - \beta \cos \theta)$, which depends only on the speed of the electron and the direction θ of the photon in the lab-frame, so we have for the total energy density in the electron frame

$$u' = \int d\Omega' \int d\nu' u'_{\nu'}(\Omega') = \gamma^2 \int d\Omega (1 - \beta \cos \theta)^2 \int d\nu u_\nu(\Omega) \quad (14.17)$$

and if the radiation is isotropic in the lab-frame then

$$u' = \gamma^2 u \int \frac{d\Omega}{4\pi} (1 - \beta \cos \theta)^2 = \gamma^2 u (1 + \beta^2/3). \quad (14.18)$$

Using $\gamma^2 = 1/(1 - \beta^2)$ we can write this as $u' = (4/3)u(\gamma^2 - 1/4)$ so the rate at which energy is scattered *into* the radiation field is

$$P_+ = P'_+ = c\sigma_T u' = \frac{4}{3}c\sigma_T u(\gamma^2 - 1/4). \quad (14.19)$$

We also need to take into account the rate at which energy is being scattered *out* of the radiation field by these collisions, which we will denote by P_- . In the electron frame this is just equal to P_+ , but this is not very useful since the incoming radiation is highly beamed in the electron frame and so P_- is not Lorentz invariant.

To compute P_- , consider instead of the *energy* density $u_\nu(\Omega)$, the photon *number* density $n_\nu(\Omega) \equiv u_\nu(\Omega)/h\nu$, and which, according to (14.16), transforms as

$$n'_{\nu'}(\Omega') d\nu' d\Omega' = (\nu'/\nu) n_\nu(\Omega) d\nu d\Omega. \quad (14.20)$$

The flux of photons in the electron-frame (number per area per second) in a range of frequency and direction $d\nu' d\Omega'$ is $cn'_{\nu'}(\Omega') d\nu' d\Omega'$ and the rate at which these are scattered out of the beam is just $c\sigma_T n'_{\nu'}(\Omega') d\nu' d\Omega'$ so the total rate of scattering events is

$$\frac{dN}{dt'} = c\sigma_T \int d\Omega' \int d\nu' n'_{\nu'}(\Omega') = c\sigma_T \gamma \int d\Omega (1 - \beta \cos \theta) \int d\nu n_\nu(\Omega) \quad (14.21)$$

which, for isotropic incident radiation, is

$$\frac{dN}{dt'} = 4\pi c\sigma_T \gamma \int d\nu n_\nu(\Omega) = c\sigma_T \gamma n. \quad (14.22)$$

where n is the total number density of photons in the lab-frame.

The proper time interval dt' is related to lab-frame time interval dt by $dt = \gamma dt'$ because of time-dilation, and therefore the rate of scatterings as measured in the lab frame is just

$$\frac{dN}{dt} = \frac{1}{\gamma} \frac{dN}{dt'} = c\sigma_T n. \quad (14.23)$$

Now Thomson scattering is independent of the frequency of the photon, so it follows that the rate at which energy is scattered out of the radiation field is given by multiplying dN/dt by the mean energy per photon, to give

$$P_- = \langle h\nu \rangle \frac{dN}{dt} = c\sigma_T u \quad (14.24)$$

Note that this is precisely the rate at which energy is scattered for a stationary electron. This is something of a coincidence since we can see from (14.21) that the photons removed from the radiation field by the moving electron do *not* have an isotropic distribution in the lab-frame, rather they have a $(1 - \beta \cos \theta)$ distribution, so the photons propagating in the direction opposite to the electron are more likely to be scattered.

Combining P_+ from (14.19) and P_- from (14.24) gives the net *inverse Compton* power for 1 electron of $P = P_+ - P_-$ or

$$P = \frac{4}{3} \beta^2 \gamma^2 c\sigma_T u \quad (14.25)$$

and the total energy transfer rate per unit volume is given by multiplying this by the electron density, or more generally by the distribution function $n(\mathbf{r}, E)$ and integrating over energy.

Equation (14.25) is remarkably simple, and also remarkably similar to the synchrotron power and the bremsstrahlung power, for reasons already discussed.

Interestingly, for low velocities, the Compton power is quadratic in the velocity. There is no first-order effect, since a scatterings may increase or decrease the photon energy.

14.4 Compton vs Inverse Compton Scattering

Equation (14.25) is supposedly valid for all electron energies, and is clearly always positive. However, this does not make sense. For cold electrons, Compton scattering result in a loss of energy for the electrons via the recoil, which was ignored in deriving (14.25).

For low energy electrons (with $v/c = \beta \ll 1$), the radiation in the electron frame is very nearly isotropic ($\delta\nu/\nu \sim \beta \ll 1$, so consequently the variation of the intensity $\delta I/I \sim \beta \ll 1$ and is also small, so we can incorporate the effect of recoil by simply subtracting the mean photon energy loss given by (14.7).

For $v \ll c$ mean rate of energy transfer is given by $dE/dt \simeq (4/3)c\sigma_T uv^2/c^2$ while the rate of scatterings is $dN/dt = c\sigma_T n = c\sigma_T u/\langle \epsilon \rangle$ so the mean photon energy gain per collision (neglecting recoil) is $\langle \Delta\epsilon \rangle/\langle \epsilon \rangle = (4/3)(v/c)^2$, and if $v \ll c$ this is approximately equal to the mean *fractional* energy gain $\langle \Delta\epsilon/\epsilon \rangle = (4/3)(v/c)^2$. For a thermal distribution of electrons this becomes $\langle \Delta\epsilon/\epsilon \rangle = 4kT/mc^2$. The mean fractional energy loss due to recoil is from (14.7) $\Delta\epsilon/\epsilon = \epsilon/Mc^2$, so combining these gives

$$\left\langle \frac{\Delta\epsilon}{\epsilon} \right\rangle = \frac{4kT - h\omega}{mc^2}. \quad (14.26)$$

It $\epsilon > 4kT$ then there is net transfer of energy to the electrons and *vice versa*.

14.5 The Compton y -Parameter

The *Compton y -parameter* is defined as

$$y \equiv \left\langle \frac{\Delta\epsilon}{\epsilon} \right\rangle \times \langle \text{number of scatterings} \rangle. \quad (14.27)$$

- In a system with y much less (greater) than unity the spectrum will be little (strongly) affected by the scattering(s).
- In computing y it is usual to either use the non-relativistic expression (14.26) or the highly relativistic limit $\langle \Delta\epsilon/\epsilon \rangle \simeq 4\gamma^2/3$.
- The mean number of scatterings is given by $\max(\tau, \tau^2)$.

- The y -parameter is generally frequency dependent.

14.6 Repeated Scatterings

14.6.1 Non-Relativistic, High Optical Depth

To a crude but useful approximation we can say that the photon gains energy according to (14.26) in each collision, and if $kT \gg h\nu$ then after N scatterings we have

$$\frac{\langle \epsilon_N \rangle}{\epsilon_0} = \left(1 + \frac{4kT}{mc^2}\right)^N = \left(1 + \frac{1}{N} \frac{4NkT}{mc^2}\right)^N \simeq e^{N \frac{4kT}{mc^2}} = e^y \quad (14.28)$$

so the mean net energy boost is the exponential of the Compton y -parameter.

14.6.2 Highly-Relativistic, Low Optical Depth

In the other extreme, $\gamma \gg 1$, the mean fractional energy gain per collision is $A = \epsilon_1/\epsilon_0 \simeq \gamma^2$ so after k scatterings we expect $\epsilon_k \sim \epsilon_0 A^k$, or equivalently, the number of scatterings required to reach energy ϵ is $k \sim \ln(\epsilon/\epsilon_0)/\ln A$.

For low optical depth $\tau \ll 1$ the probability of k -scatterings is $p(k) \sim \tau^k$, so the intensity in highly boosted photons is

$$I(\epsilon_k) \sim I(\epsilon_0) \tau^k \sim I(\epsilon_0) e^{k \ln \tau} \sim I_0 \exp\left(\frac{\ln \tau \ln(\epsilon_k/\epsilon_0)}{\ln A}\right) \sim I_0 \times \left(\frac{\epsilon_k}{\epsilon_0}\right)^{\ln \tau / \ln A} \quad (14.29)$$

so the result is a power law $I(\epsilon) \propto \epsilon^{-p}$ with $p = -\ln \tau / \ln A$.

14.7 The Sunyaev-Zel'dovich Effect

The *Sunyaev-Zel'dovich effect* is the effect on the microwave background radiation induced by scattering off hot gas.

- Since the input spectrum is nearly thermal and therefore has high occupation number in R-J region, a proper treatment requires consideration of stimulated emission. For non-relativistic gas, the evolution of the spectrum is described by the *Kompaneets* equation (see R+L) which is a *Fokker-Planck* equation describing the diffusion of photon energies. Here we will give only a qualitative discussion of the effect.
- Photons may be scattered out of or into the line of sight, so the number of photons does not change. Photons may be scattered up or down in energy. On average, the increase in energy is $\Delta\epsilon = 4kT\epsilon/mc^2$. The result must be some shift of the energy distribution function $n_\nu(\Omega)$ to generally higher energies, but which preserves the area $\int d\nu n_\nu$. This gives an increase in intensity at high frequencies $h\nu \gtrsim kT_{\text{MBR}}$ and a reduction in the R-J region.
- It has been observed for several clusters of galaxies.
- The R-J *SZ decrement* gives a fluctuation in brightness temperature $\Delta T/T = -2y$. It is on the order of 10^{-4} .
- The SZ effect measures the integral along the line of sight of the electron density times the temperature $\int dz n_e T$, which is proportional to the integral of the *pressure* along the line of sight.
- For a cluster of size R the SZ decrement is proportional to $n_e R$ whereas the X-ray emission is proportional to $n_e^2 R$, so the ratio of the square of the SZ effect to the X-ray emission provides an estimate of the physical size of the cluster.

- Combining the physical size with the redshift and the angular size of the cluster provides a direct estimate of the scale of the universe, or equivalently of the Hubble parameter H_0 .
- There is another effect — the *kinematic SZ effect* — in which clusters give rise to a temperature fluctuation via the Doppler effect. This is on the order of the optical depth to electron scattering times the line of sight velocity and it typically somewhat smaller than the true SZ decrement.

14.8 Compton Cooling and Compton Drag

Compton scattering of microwave background photons off of moving electrons removes energy from electrons. Collisions with ions will replenish the electron energy, so the net result is to cool the gas.

We can estimate the time-scale for this cooling as follows: For non-relativistic electrons, the electrons see radiation which is slightly anisotropic; the temperature is $T(\mu) = (1 + \beta\mu)T_0$, where T_0 is the isotropic temperature seen by a stationary observer. The intensity, which scales as the fourth power of temperature, is then $I(\mu) \sim (1 + 4\beta\mu)I_0$, and computing the radiative force $F = (\sigma_T/c) \int d\Omega \mu I(\mu)$ gives $F \sim (4/3)u\sigma_T\beta$. This gives an equation of motion for a single electron:

$$\frac{dv}{dt} = - \left(\frac{4}{3} \frac{u\sigma_T}{m_e c} \right) v. \quad (14.30)$$

The radiation scattering acts like a viscous drag force, and the velocity decays as $v(t) = v(t_0) \exp(-t/\tau_e)$. The time-scale for the velocities to decay — which is also the time-scale for the gas too cool, provided collisions can replenish the electron energy sufficiently fast is

$$\tau_e \sim \frac{m_e c}{\mu\sigma_T} \quad (14.31)$$

The time-scale for ‘Compton-cooling’ exceeds the age of the universe today, but was effective in the past. It may play an important role in galaxy formation as cooling is necessary to allow the baryonic material to settle within the dark matter potential well, since otherwise the hot gas would simply remain as a hot atmosphere in hydrostatic equilibrium.

Compton drag is a frictional force exerted on ionized gas which is moving relative to the microwave background frame. This is highly analogous to the drag on an electron computed above. The time-scale is much longer, however. This is because for a blob of ionized plasma, the scattering cross-section is supplied almost completely by the electrons while the inertia is provided almost entirely by the ions. For ionized hydrogen, the time-scale for the velocity to decay is then

$$\tau_p \sim \frac{m_p c}{\mu\sigma_T} \quad (14.32)$$

which is longer than τ_e by a factor $m_p/m_e \simeq 2000$. This only becomes effective at very early times, but results in any ionized gas being effectively locked to the frame in which the microwave background appears isotropic.

14.9 Problems

14.9.1 Compton scattering 2

Consider the scattering of a photon off an electron at rest. Write down suitable 4-momentum vectors \vec{p} to describe the incoming/outgoing states (use subscripts i, o for in/out photon states), denote photon energies by e , photon directions by \mathbf{n} and outgoing photon 3-mom by \mathbf{p} .

By considering the squared modulus of the outgoing electron 4-mom (or otherwise) show that

$$e_o = \frac{e_i}{1 + \frac{e_i}{mc^2}(1 - \cos\theta)} \quad (14.33)$$

where θ is the angle between the incoming/outgoing photon directions, and that the photon wavelengths are related by

$$\lambda_o = \lambda_i + \lambda_c(1 - \cos \theta) \quad (14.34)$$

where $\lambda_c = h/mc$ is the Compton wavelength of the electron.

14.9.2 Inverse Compton Effect

Consider a ‘head on’ collision between a photon of energy e and an electron with initial Lorentz factor γ (as measured in the ‘lab-frame’) and in which the photon reverses its direction of motion. (Assume the electron is initially moving left — negative velocity — and the photon is travelling to the right).

What is the energy of the incoming photon in the electron rest frame? What is the energy of the outgoing photon in the electron rest frame? Neglecting the change of photon energy in the rest frame, compute the energy of the outgoing photon in the lab frame. Assuming $\gamma \gg 1$, estimate the fractional error in the lab-frame energy change incurred in neglecting the e -frame energy change.

Now consider a highly relativistic electron propagating through the 3K microwave background. What is the typical incoming photon energy in the electron rest frame? What is the limit on γ such that it is valid to use the Thomson scattering cross-section σ_T to estimate the rate of collisions.

14.9.3 Compton y -parameter

Define the ‘Compton y -parameter’. Consider a galaxy cluster containing gas at temperature 10 keV, scattering microwave background photons. Assuming an electron density $n_e \sim 10^{-3} \text{cm}^{-3}$ and size $\sim 1 \text{Mpc}$, estimate the optical depth τ to electron scattering.

What is the typical change in frequency for a scattered photon?

What is the average change in frequency for a scattered photon?

What is the y -parameter?

What is the fractional change in intensity of the MBR in the Raleigh-Jeans region.

Clusters are thought to move relative to the microwave background with velocities $v \sim (1 - 2) \times 10^{-3}c$. Discuss in general terms what effect would this have on the microwave background temperature? How does this compare with the inverse-compton effect? How might one disentangle these effects?

Part III

Field Theory

Chapter 15

Field Theory Overview

Lagrangian dynamics is ideally suited for analyzing systems with many degrees of freedom. Such systems include lattices of atoms in crystals and, in the ‘continuum limit’, the behavior of continuous fields. For systems with potentials which are quadratic in the coordinates — which is always the case for small amplitude oscillations about a potential minimum — one can always find *normal modes of oscillation*. These are linear combinations of the coordinates q_i in terms of which the system becomes a set of independent simple harmonic oscillators. For the crystal lattice, and for classical fields, these normal modes are just traveling waves. This means that quantum mechanically, the system is simply described by a set of occupation numbers or energy levels for each of the normal modes.

The quantum mechanical states of the system in the *occupation number representation* are

$$|n_1, n_2 \dots n_j \dots\rangle \quad (15.1)$$

with n_j giving the occupation number of the j th normal mode. For a non-interacting field theory these states are exact eigenstates of the total Hamiltonian. The non-interacting or ‘free’ field is rather sterile; it just sits there. It is also highly idealized, in reality the system must interact with the outside world, and there may also be internal interactions. External interactions, such as wiggling an atom in a crystal or wiggling a charge in the electro-magnetic field, can add or remove energy, and so change occupation numbers. Similarly, any degree of anharmonicity of the oscillators will introduce coupling between the normal modes, so the quantum theory should allow for scattering of quanta. The usefulness of the multi-particle states (15.1) in the context of interacting fields is that for many systems the interactions can be considered a small perturbation on the non-interacting theory, so these are approximate eigenstates of the system and one can apply perturbation theory to compute transition amplitudes for the system to go from one state to another.

This program leads to quantum field theory, encompassing quantum electrodynamics — which provides the proper description of the scattering processes which we have treated above in an approximate classical manner for the most part — and also theories of weak and strong interactions. This a huge and formidable subject, combining, as it does, special relativity and quantum mechanics. In the following three chapters I shall try to give a flavor of these theories with a simplified treatment that ignores many important features, but which illustrates some features which are of great relevance for astrophysics. The approach I shall follow is similar to that of Ziman, in his “Elements of Advanced Quantum Theory” in that I shall first consider a ‘solid-state’ model consisting of a lattice of coupled oscillators. This system can be analyzed using regular Lagrangian mechanics, and taking the continuum limit this becomes a classical Lagrangian field theory for ‘scalar-elasticity’ waves. We then consider the quantum mechanics of this system, the quanta of which are phonons. This may seem out of place in a lecture course on astrophysics. Our reason for exploring this model is that if we ‘abstract away’ the underlying physical medium and choose coefficients to make the field equations relativistically covariant we obtain a quantum field theory for the relativistic massive scalar field, whose quanta are massive spin-less bosonic particles (chapter 18). These entirely distinct systems are, in a sense, identical, since they have the same Lagrangian. This means that all the results for

the more concrete, and conceptually less challenging, atomic lattice carry over to the more abstract, but for us much more interesting, scalar field.

This in itself may seem somewhat removed from the kind of interactions we have mostly been considering here which are between electrons (massive but fermionic in nature) and the electro-magnetic field (bosonic but massless and a vector field). One motivation for considering the scalar field is that it is in many ways the simplest type of matter field, and nicely illustrates many features of the quantized electro-magnetic field without the complications of polarization, gauge invariance etc. It shows how mass is introduced into the theory, and the theoretical machinery for calculating transition amplitudes, reaction rates etc carries over to quantum electrodynamics and then to weak-interaction theory and beyond. Another motivation for considering the scalar field is the important role that such fields play in modern cosmology.

Here is a ‘road-map’ to the next three chapters.

In chapter 16 we develop classical non relativistic field theory. In §16.1 we construct a simple mechanical system consisting of a lattice of coupled oscillators; each oscillator consists of a bead on a rod with a spring, and the oscillators are connected to their neighbors with coupling springs. This system is analyzed using regular Lagrangian dynamics, and we obtain the normal modes of oscillations and dispersion relation etc. for such a lattice.

In §16.2 take the *continuum limit* of the Lagrangian for this lattice to obtain the *Lagrangian density* function $\mathcal{L}(\dot{\phi}, \nabla\phi, \phi)$. We then show how the field equations for waves on this lattice can be derived from this Lagrangian density by requiring, as always, that the action S be extremized. In §16.3 we explore the conservation laws which arise from the invariance of the Lagrangian density under time and space translations. The former gives conservation of energy, just as we could have obtained from regular mechanics. The latter gives rise to something quite new, which is not apparent in the usual mechanical analysis; this is the *wave-momentum* which is conserved independently of the ‘microscopic’ momentum embodied in the canonical momenta of the system. We show that the energy and wave-momenta of classical wave packets are related by $\mathbf{P}/E = \mathbf{k}/\omega$ where \mathbf{k} and ω are the spatial and temporal frequencies respectively.

The lattice model we explore (see figure 16.1) is very specific; in addition to the coupling springs, which allow waves to propagate along the lattice, there are internal springs. These have the consequence that the oscillation frequency tends to a finite lower limit as $\mathbf{k} \rightarrow 0$. In fact the dispersion relation is

$$\omega^2 = k^2 c^2 + \frac{m^2 c^4}{\hbar^2} \quad (15.2)$$

where c is the wave-speed for high momentum waves, and m is a parameter with units of mass determined by the spring constants and bead masses. In §16.4 we show how the classical wave-packets for this field have energy-momentum relation $E^2 = P^2 c^2 + M^2 c^4$ which mimics that of relativistic quanta: This is quite deliberate — it allows this non-relativistic system to illustrate many features of relativistic fields — and we show, for instance, that the momentum of a wave packet is given by $\mathbf{p} = \gamma M \mathbf{v}$, where $\gamma = (1 - v^2/c^2)^{-1/2}$ and \mathbf{v} is the group velocity. We also show in §16.5 that the solutions of this system obey a covariance: if $\phi(\mathbf{x}, t)$ is a solution then $\phi'(\mathbf{x}', t') = \phi(\mathbf{x}, t)$ is also a solution, where (\mathbf{x}, t) and (\mathbf{x}', t') are related by a transformation which is formally identical to a Lorentz transformation.

To round off our discussion of classical non-relativistic fields we consider interacting field theories in §16.6. We discuss how interactions — which couple the otherwise independent planar traveling wave solutions — can be introduced either through non-linear springs (a ‘self-interaction’) or by coupling between different fields. We show that interactions between waves are particularly efficient if the waves obey a ‘resonance’ condition; this condition is that the sum of the spatial and temporal frequencies of the incoming and outgoing waves should be equal. This condition arises again later, where in the quantum mechanical analysis the interaction rates include a energy and momentum conserving δ -function which enforces this resonance condition.

In chapter 17 we develop the quantum mechanical description of non-relativistic fields. We start with the quantization of a single simple harmonic oscillator in §17.1 where we introduce the creation and destruction operators a^\dagger and a which are central to all quantum field calculations. In §17.2 we describe the ‘interaction picture’, which is a hybrid of the Schroedinger and Heisenberg pictures,

and show how the time-dependent perturbation theory in this formalism leads to the ‘ S -matrix expansion’ in §17.2.1. This is illustrated with the example of a forced oscillator in §17.2.2.

In §17.3 we apply these concepts to free — i.e. non-interacting — fields, using the example of the scalar elasticity model. We first obtain the creation and destruction operators for phonons of the discrete lattice system in §17.3.1 and show that these have the appropriate commutation relations and are related to the Hamiltonian in the proper way. We generalize to continuous fields in one or more dimensions in §17.3.2.

In §17.4 we introduce perturbative interactions and compute scattering rates for various processes. These include scattering off an impurity in the lattice (§17.4.1); scattering of phonons *via* a $\lambda\phi^4$ self interaction (§17.4.2); and in §17.4.3 we consider ‘second-order’ scattering of a phonon by the exchange of a virtual phonon of a different type. All of these are worked through in detail. In §17.4.4 we briefly discuss the contour integral formalism for computing scattering rates and we discuss the ‘Feynman rules’ for this scalar phonon system in §17.4.5. In §17.4.6 we discuss the ‘kinematic constraints’ placed on scattering and decay processes by the requirements of conservation of energy and momentum.

In chapter 18 we turn to relativistic quantum fields. In §18.1 we develop the theory of the massive scalar field (or Klein-Gordon field). This proves to be very simple since the system is formally identical to the ‘scalar-elasticity’ model we had considered in the previous two chapters. Making this theory relativistically covariant is little more than choosing appropriate coefficients in the Hamiltonian. We then discuss self-interactions, spontaneous symmetry breaking and scattering of particles *via* a coupling between independent fields. We discuss the generalization of the scalar field theory to more complicated fields and discuss quantum electrodynamics in §18.2, though the treatment is rather shallow. In §18.3 we show how the scattering amplitudes computed by perturbation theory lead immediately to kinetic theory in the form of the fully relativistic, fully quantum-mechanical Boltzmann equation.

In §18.4 we return to the scalar field and explore the role of such fields in cosmology. We show how such fields can drive inflation, either in the very early or very late Universe; how they can behave like ‘cold dark matter’ or like relativistic particles and we also show how, with an appropriate potential function, they can lead to domain-walls, cosmic strings and other ‘topological defects’. In §18.5 we explore in a little more detail the evolution of the Klein-Gordon field in the non-relativistic limit (i.e. where the wavelength is much greater than the Compton wavelength). We show that if we factor out a rapid common oscillation factor, the equation of motion for such fields is in fact the time dependent Schrödinger equation. We discuss the correspondence principle, and also show how non-relativistic fields are coupled to gravity.

Chapter 16

Classical Field Theory

Here we develop the classical Lagrangian approach to field theory. We first introduce a model consisting of a discrete lattice of coupled oscillators. While fairly simple, the model proves to be extremely versatile and the properties of the sound waves on this lattice demonstrate many of the properties usually associated with relativistic fields. Taking the ‘continuum limit, we show how the equations of motion for a field can be generated from the Lagrangian density. Next, we derive the conservation laws for momentum and energy; the latter is very similar to that in regular Lagrangian mechanics, but the former is very different and is a uniquely field theoretical construct. We explore the relation between wave energy and momentum for sound waves in our model system — which is formally identical to that for relativistic particles — and also show how the sound-wave equations display an invariance under Lorentz-like boost transformations. We then consider the effect of adding interaction terms to the Lagrangian. These introduce a coupling between the otherwise independent normal modes, and we emphasize the resonance conditions that must be satisfied to allow effective coupling. We then discuss some puzzles and paradoxes concerning the wave momentum. We next show that in the long-wavelength limit (corresponding to the non-relativistic limit) that, in addition to energy and momentum a fifth quantity is conserved. This extra conservation law is the wave-mechanical analog of conservation of particle number or proper mass. It is only approximately conserved, but this approximation becomes exact in the limit of small group velocity. Finally, we develop the equations governing the transport of energy and momentum in a self-interacting field theory. We show that the sound-wave energy density behaves just like a collisional gas, with same equations of motion, and with exactly the same adiabatic indices in the relativistic and non-relativistic limits.

16.1 The BRS Model

Consider a system consisting of a 1-dimensional array of identical beads of mass M constrained to move in the vertical direction by rods and with a spring of spring constant K connected to the particle. The rods are assumed to be infinitely stiff, and the bead slide up and down with no friction. We will refer to this beads, rods and springs system as the ‘BRS’ model.

Let ϕ_j denote the displacement of the j th particle from the rest position. As it stands we simply have a set of independent SHM oscillators with equations of motion $\ddot{\phi}_j = -(K/M)\phi_j$. Now add some additional springs with spring constant K' which link the beads to their neighbors as illustrated in figure 16.1. The Lagrangian is just the kinetic minus the potential energy, which we can readily write down:

$$L(\phi_i, \dot{\phi}_i) = \frac{M}{2} \sum_i \dot{\phi}_i^2 - \frac{K'}{4} \sum_i (\phi_{i-1} - \phi_i)^2 - \frac{K'}{4} \sum_i (\phi_{i+1} - \phi_i)^2 - \frac{K}{2} \sum_i \phi_i^2. \quad (16.1)$$

The reason for the factor $1/4$ in the energy for the connecting springs is that the energy in these springs is shared between two neighbors.

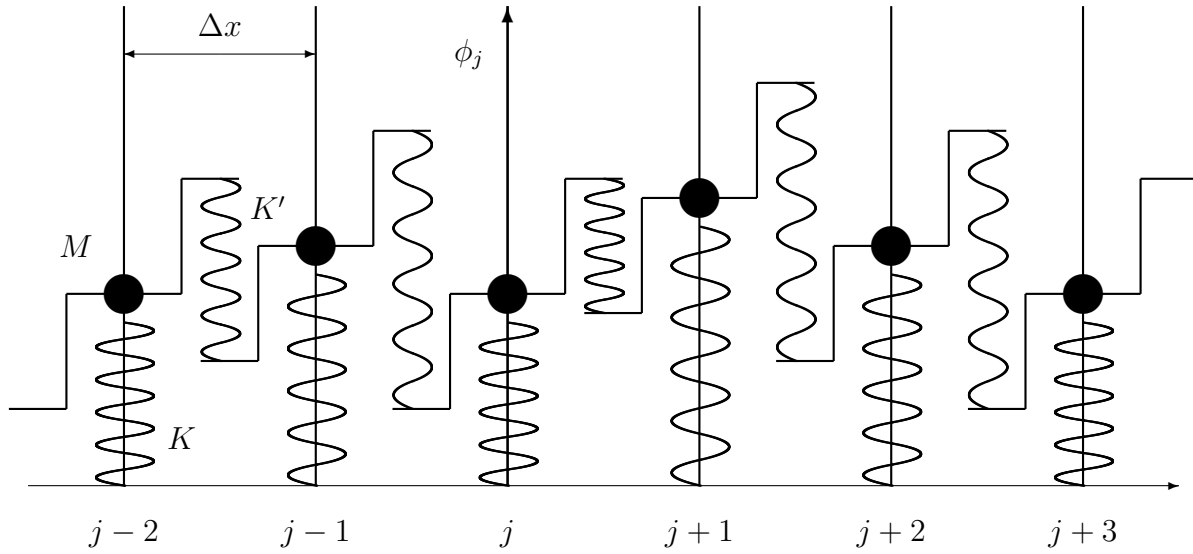


Figure 16.1: A lattice of coupled oscillators. Massive beads are constrained to move in the vertical direction by rods, and tethered to the base by springs with spring constant K . Neighboring particles are also coupled by springs of spring constant K' . The displacement of the j th bead is ϕ_j .

The equations of motion for this system are the Euler-Lagrange equations

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\phi}} \right) = \frac{\partial L}{\partial \phi}. \quad (16.2)$$

The partial derivatives with respect to the velocities appearing on the LHS are

$$\frac{\partial L}{\partial \dot{\phi}_j} = M \dot{\phi}_j \quad (16.3)$$

and the partial derivatives with respect to the displacements on the RHS are

$$\frac{\partial L}{\partial \phi_j} = K'[(\phi_{j-1} - \phi_j) + (\phi_{j+1} - \phi_j)] - K\phi_j \quad (16.4)$$

so the Euler-Lagrange equations are

$$M\ddot{\phi}_j - K'[\phi_{j-1} - 2\phi_j + \phi_{j+1}] + K\phi_j = 0. \quad (16.5)$$

In terms of the ϕ_j we have a set of coupled oscillators. It is easy to find a set of normal modes which are decoupled. Let us imagine we construct a circular loop of N of these systems (which means we don't need to worry about boundary conditions) and define the discrete Fourier transform of the displacements

$$\Phi_k(t) \equiv \sum_j \phi_j(t) e^{2\pi i j k / N} \quad (16.6)$$

The transform of the Euler-Lagrange equations is then

$$\sum_j \left[M\ddot{\phi}_j - K'[\phi_{j-1} - 2\phi_j + \phi_{j+1}] + K\phi_j \right] e^{2\pi i j k / N} = 0. \quad (16.7)$$

Now $\sum \phi_{j-1} e^{2\pi i j k / N} = \Phi_k e^{2\pi i k / N}$ and $\sum \ddot{\phi}_j e^{2\pi i j k / N} = \ddot{\Phi}_k$ so (16.7) says

$$\ddot{\Phi}_k + [2K'(1 - \cos 2\pi k / N) + K]\Phi_k = 0 \quad (16.8)$$

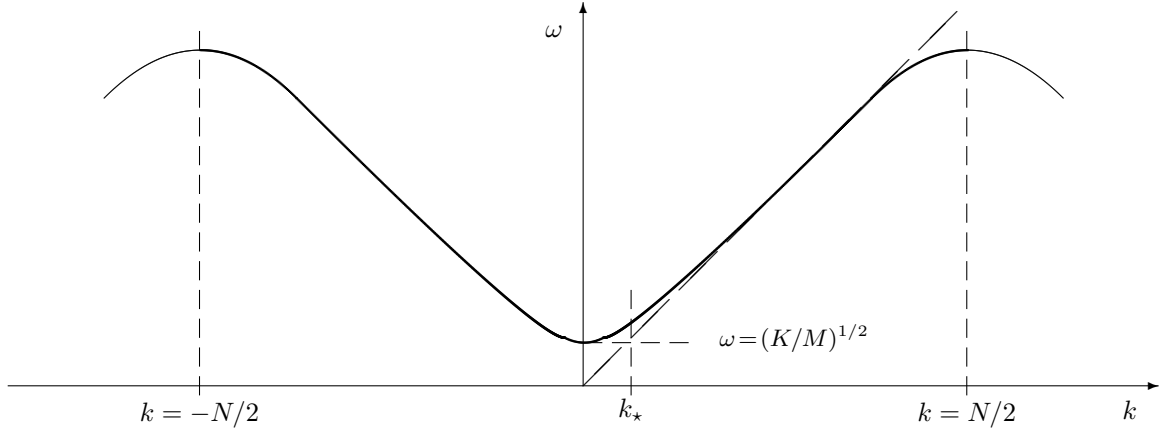


Figure 16.2: Dispersion relation for the discrete lattice model (16.9). The minimum frequency is $\omega_{\min} = \sqrt{K/M}$ at $k = 0$. Here all of the beads oscillate up and down together. The maximum frequency is $\omega_{\max} = \sqrt{(4K' + K)/M}$ at $k = \pm N/2$. Here neighboring beads are 180 degrees out of phase with each other. The frequency is also defined for wave-numbers $|k| > N/2$, but these are just aliased versions of the waves within the range $|k| \leq N/2$. For the case $K \ll K'$, i.e. where the connecting springs are relatively strong, and as assumed here, there is another scale in the problem: $k_* = N\sqrt{K/2K'}$. The significance of this spatial frequency scale is that for $k \ll k_*$ most of the potential energy is in the K -springs, while for $k \gg k_*$ most of the potential energy is in the K' springs which link the particles. For $k \ll N$ (i.e. wavelengths much bigger than the separation of the oscillators) the dispersion relation is $\omega(k) \simeq \omega_{\min} \sqrt{1 + (k/k_*)^2}$. Thus, for $k_* \ll k \ll N$ we have $\omega \propto k$ and the waves are non-dispersive.

so evidently the discrete transforms Φ_k defined in (16.6) are a set of normal modes; i.e. the equation of motion for the mode k is independent of that for all of the other modes $k' \neq k$.

The solutions of (16.8) are $\Phi_k(t) = \Phi_{k0} e^{\pm i\omega t}$ with frequency given by the dispersion relation

$$\omega(k) = \sqrt{\frac{2K'(1 - \cos 2\pi k/N) + K}{M}} \quad (16.9)$$

as sketched in figure 16.2.

For completeness, we note that the momentum conjugate to the displacement ϕ_i is

$$p_i = \frac{\partial L}{\partial \dot{\phi}_i} = M \dot{\phi}_i \quad (16.10)$$

and that the Hamiltonian is

$$H(p_i, \phi_i) = \sum_i p_i \dot{\phi}_i - L = \sum_i p_i^2 / M - L. \quad (16.11)$$

This is identical in form to the Lagrangian (16.1) save that the terms involving the spring constant K , K' have positive sign. We could have inferred this from the fact that $L = T - U$ while the Hamiltonian is the total energy $H = T + U$.

16.2 The Continuum Limit

It is interesting to consider the limiting situation where the wavelength is much greater than the spacing between the units (or equivalently, for finite wavelength, in the limit that the spacing $\Delta x \rightarrow 0$). In this limit, the displacement varies little from unit to unit, so viewed macroscopically ϕ_j behaves

like a continuously varying field $\phi(x)$ with $x = j\Delta x$. We can therefore replace finite differences like $\phi_{j+1} - \phi_j$ in the Lagrangian with $\Delta x \nabla \phi$.

The Lagrangian then becomes

$$L = \sum_j \Delta x \left(\frac{M}{2\Delta x} \dot{\phi}^2 - \frac{K'\Delta x}{2} (\nabla \phi)^2 - \frac{K}{2\Delta x} \phi^2 \right) \quad (16.12)$$

and in the limit $\Delta x \rightarrow 0$ the sum becomes an integral: $\sum \Delta x \dots \rightarrow \int dx \dots$, so the Lagrangian is

$$L = \int dx \mathcal{L}(\dot{\phi}, \nabla \phi, \phi) \quad (16.13)$$

where we have defined the *Lagrangian density*

$$\mathcal{L}(\dot{\phi}, \nabla \phi, \phi) = \frac{1}{2} \rho (\dot{\phi}^2 - c_s^2 (\nabla \phi)^2 - \mu^2 \phi^2). \quad (16.14)$$

This is a quadratic function of the field and its space and time derivatives with constant coefficients

$$\rho = \frac{M}{\Delta x} \quad c_s = \sqrt{\frac{K'}{M}} \Delta x \quad \mu = \sqrt{\frac{K}{M}}. \quad (16.15)$$

Here ρ is the line density; μ is the frequency for the oscillators if decoupled, and c_s , as we shall see, is the asymptotic wave velocity for high spatial frequency waves. Similarly, the total Hamiltonian is a spatial integral

$$H = \int d^3x \mathcal{H}(\dot{\phi}, \nabla \phi, \phi) \quad (16.16)$$

where the *Hamiltonian density* is

$$\mathcal{H}(\dot{\phi}, \nabla \phi, \phi) = \frac{1}{2} \rho (\dot{\phi}^2 + c_s^2 (\nabla \phi)^2 + \mu^2 \phi^2) \quad (16.17)$$

where the three terms represent the kinetic energy density and the densities of energy in the springs K' and K respectively.

The action now becomes a 2-dimensional integral over time and space:

$$S = \int dt L = \int dt \int dx \mathcal{L}(\dot{\phi}, \nabla \phi, \phi). \quad (16.18)$$

For a system with a finite number of degrees of freedom n (the displacements of the n beads in our lattice system) we obtain the Euler-Lagrange equations by varying the n paths $q_i(t) \rightarrow q_i(t) + \delta q_i(t)$ and requiring that the action be stationary. Here the index i has become the continuous variable x and we require that S above be stationary with respect to variations if the *field* $\phi(x, t)$:

$$\phi(x, t) \rightarrow \phi'(x, t) = \phi(x, t) + \delta \phi(x, t). \quad (16.19)$$

The variation of the field velocity and gradient are

$$\dot{\phi}' = \dot{\phi} + \partial(\delta \phi)/\partial t = \dot{\phi} + \delta \dot{\phi} \quad (16.20)$$

and

$$\nabla \phi' = \nabla \phi + \nabla(\delta \phi), \quad (16.21)$$

so the variation of the action $\delta S = S' - S$ corresponding to the field variation (16.19) is

$$\delta S = \int dt \int dx \mathcal{L}(\dot{\phi} + \delta \dot{\phi}, \nabla \phi + \nabla(\delta \phi), \phi + \delta \phi) - \int dt \int dx \mathcal{L}(\dot{\phi}, \nabla \phi, \phi). \quad (16.22)$$

For an infinitesimal variation, we can make a 1st order Taylor expansion to give

$$\delta S = \int dt \int dx \left[\delta \dot{\phi} \frac{\partial \mathcal{L}}{\partial \dot{\phi}} + \nabla(\delta \phi) \frac{\partial \mathcal{L}}{\partial \nabla \phi} + \delta \phi \frac{\partial \mathcal{L}}{\partial \phi} \right]. \quad (16.23)$$

Reversing the order of the integrals, the first term is the integral over position of

$$\int_{t_1}^{t_2} dt \delta\dot{\phi} \frac{\partial \mathcal{L}}{\partial \dot{\phi}}. \quad (16.24)$$

In this integral, x is to be considered constant, so the symbol $\delta\dot{\phi}$ denotes the ordinary time derivative of $\delta\phi(x, t)$ at this position. This means we can integrate by parts to give

$$\int_{t_1}^{t_2} dt \delta\dot{\phi} \frac{\partial \mathcal{L}}{\partial \dot{\phi}} = \left[\delta\phi \frac{\partial \mathcal{L}}{\partial \dot{\phi}} \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} dt \delta\phi \frac{d(\partial \mathcal{L} / \partial \dot{\phi})}{dt}. \quad (16.25)$$

where the operator d/dt denotes the derivative with respect to time at fixed position x . Now the boundary terms vanishes if $\delta\phi(x, t)$ is assumed to vanish at the initial and final times. Alternatively, we can take the time integral from $t = -\infty$ to $t = +\infty$ and require that the fields tend to zero as $t \rightarrow \pm\infty$. This allows us to replace the integral involving $\delta\dot{\phi}$ by one involving $\delta\phi$.

A word on the notation is perhaps in order. The time derivative operator d/dt here is really the partial derivative with respect to time at constant position. Normally we would write this as $\partial/\partial t$ but here that would be ambiguous since it does not tell us what variables are to be held constant. In general, the Lagrangian density may depend explicitly on time (e.g. if any of the coefficients like ρ and μ were time dependent) and we have $\mathcal{L}(\dot{\phi}, \nabla\phi, \phi, t)$. When we write $\partial(\partial \mathcal{L} / \partial \dot{\phi}) / \partial t$ we mean the time derivative at constant value of the field and its derivatives. In our case this vanishes since $\partial \mathcal{L} / \partial t = 0$. However, if we evaluate $\partial \mathcal{L} / \partial \dot{\phi}$ using the actual field $\phi(x, t)$ and its derivatives, this is some specific function of x and t . This has a well defined time derivative at constant x , which is what we mean when we write $d(\partial \mathcal{L} / \partial \dot{\phi}) / dt$.

Similarly, the spatial integral appearing in the second term is

$$\int dx \nabla(\delta\phi) \frac{\partial \mathcal{L}}{\partial \nabla\phi} = \left[\delta\phi \frac{\partial \mathcal{L}}{\partial \nabla\phi} \right] - \int dx \delta\phi \frac{d(\partial \mathcal{L} / \partial (\nabla\phi))}{dx} \quad (16.26)$$

where the operator d/dx denotes the derivative with respect to position at fixed time. Here again we can discard the boundary terms if we assume that the fields tend to zero at spatial infinity, or if we invoke periodic boundary conditions.

With these substitutions, every term in (16.23) now contains a multiplicative factor $\delta\phi$, and the variation of the action is

$$\delta S = - \int dt \int dx \delta\phi \left[\frac{d(\partial \mathcal{L} / \partial \dot{\phi})}{dt} + \frac{d(\partial \mathcal{L} / \partial (\nabla\phi))}{dx} - \frac{\partial \mathcal{L}}{\partial \phi} \right]. \quad (16.27)$$

The actual field $\phi(x, t)$ is such that the action is extremized; i.e. the variation in the action must vanish for an arbitrary perturbation $\delta\phi(x, t)$ about this field. This means that the quantity in brackets must vanish at all points (x, t) . This gives the Euler-Lagrange equation

$$\frac{d(\partial \mathcal{L} / \partial \dot{\phi})}{dt} + \frac{d(\partial \mathcal{L} / \partial (\nabla\phi))}{dx} - \frac{\partial \mathcal{L}}{\partial \phi} = 0, \quad (16.28)$$

which provides the equation of motion for the field.

This was derived here for a Lagrangian density with no explicit time or position dependence, but had we taken instead $\mathcal{L}(\dot{\phi}, \nabla\phi, \phi, x, t)$ we would have obtained precisely the same variation of the action (16.22) and therefore we would also have obtained precisely the same field equations as in (16.28) above.

This is quite general, but also somewhat abstract. If we specialize to the BRS model Lagrangian density (16.14), for example, the Euler-Lagrange equation (16.28) becomes

$$\ddot{\phi} - c_s^2 \nabla^2 \phi + \mu^2 \phi = 0. \quad (16.29)$$

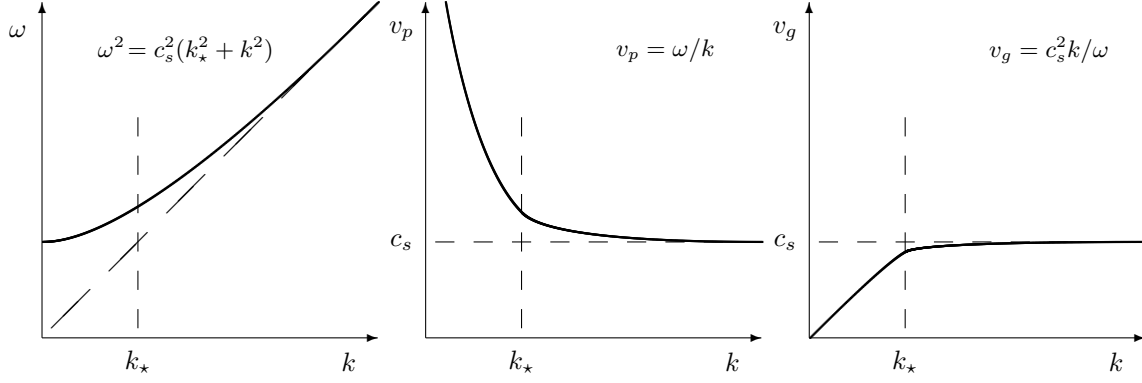


Figure 16.3: The panel on the left shows the continuum limit dispersion relation panel for the BRS model. The center panel shows the phase velocity and the right panel shows the group velocity.

This is a 1-dimensional wave equation, with traveling wave solutions

$$\phi(x, t) = \phi_0 e^{i(\omega t - kx)}, \quad (16.30)$$

and dispersion relation

$$\omega(k)^2 = c_s^2 k^2 + \mu^2. \quad (16.31)$$

This dispersion relation is entirely equivalent to the low-frequency limit of (16.9) with $i = x/\Delta x = Nx/L$ and with the integer mode index k replaced by $k/\Delta k$ with $\Delta k = 2\pi/L$. We could also have obtained the continuum limit wave equation directly from the discrete version in a similar way. Our motivation for taking the more laborious route above is to illustrate how the wave-equation etc. can be derived directly from a Lagrangian density like (16.14) for a continuous field ϕ , regardless of whether this is really the limit of some discrete lattice system.

Just as before, the dispersion relation (16.31) defines a characteristic spatial frequency $k_* = \sqrt{\mu/c_s}$. Referring to the Hamiltonian density (16.17), for $k \gg k_*$ most of the potential energy is in the field gradient term $\rho c_s^2 (\nabla \phi)^2/2$ while for $k \ll k_*$ the potential is mostly $\rho \mu^2 \phi^2/2$. The *phase velocity* is

$$v_{\text{phase}} = \frac{\omega_k}{k} = c_s \sqrt{1 + (k_*/k)^2} \quad (16.32)$$

where $\omega_k \equiv \omega(k)$. This tends to infinity for low wave numbers $k \ll k_*$ and to c_s for $k \gg k_*$. The *group velocity* is

$$v_{\text{group}} = \frac{d\omega_k}{dk} = c_s^2 \frac{k}{\omega_k} = \frac{c_s}{\sqrt{1 + (k_*/k)^2}}. \quad (16.33)$$

The group velocity gives the speed of motion of a *wave-packet*, and therefore also the speed at which information can be propagated. Here the group velocity tends to c_s as $k \rightarrow \infty$. For small wave numbers $k \ll k_*$, the group velocity is proportional to k . The dispersion relation and the phase and group velocities are illustrated in figure 16.3.

The generalization to two or more spatial dimensions is quite straightforward. We could, for example, cover a table with a grid of vertical rods with beads bouncing up and down on springs, and couple each bead to its four nearest neighbors in much the same way as we did in one dimension. The 1-dimensional discrete index i could then be replaced by a 2-vector \mathbf{i} with integer components i^1, i^2 . Taking the continuum limit then involves replacing \mathbf{i} with the continuous variable $\mathbf{x} = \mathbf{i}/\Delta x$. The Lagrangian density would now be

$$\mathcal{L}(\dot{\phi}, \partial\phi/\partial x^1, \partial\phi/\partial x^2, \phi) = \mathcal{L}(\dot{\phi}, \phi_{,j}, \phi) \quad (16.34)$$

where $\phi_{,j} = \partial\phi/\partial x^j$, and the action becomes a three dimensional integral $S = \int dt \int d^2x \mathcal{L}$. Following the same reasoning as above we are led to an Euler-Lagrange equation like (16.28), but

with multiple spatial gradient terms. Using the Einstein summation convention, this can be written succinctly as

$$\frac{d(\partial\mathcal{L}/\partial\dot{\phi})}{dt} + \frac{d(\partial\mathcal{L}/\partial\phi_{,j})}{dx^j} - \frac{\partial\mathcal{L}}{\partial\phi} = 0. \quad (16.35)$$

The continuum limit Lagrangian density for the multi-dimensional lattice system is identical to (16.14), but where now the ∇ operator is the multi-dimensional spatial gradient operator, so $(\nabla\phi)^2 = \phi_{,i}\phi_{,i}$. Taking the derivative with respect to the field gradients gives $\partial\mathcal{L}/\partial\phi_{,j} = c_s^2\rho\phi_{,j}$, so $d(\partial\mathcal{L}/\partial\phi_{,j})/dx^j = c_s^2\rho\phi_{,jj} = c_s^2\rho\nabla^2\phi$. The continuum limit equation of motion for the multi-dimensional system is then identical to (16.29), but where now ∇^2 denotes the Laplacian operator.

16.3 Conservation of Wave-Momentum

In ordinary classical mechanics, for a closed system, i.e. one for which the Lagrangian has no explicit time dependence, the total energy $E = \sum p_i\dot{q}_i - L$ is conserved. In classical Lagrangian field theory, the time is replaced by space-time coordinates $t \rightarrow x^\mu = (t, \mathbf{x})$. The action is $S = \int dt L \rightarrow \int dt \int d^3x \mathcal{L}$ for example, and the 2nd time derivative operator in the equation of motion becomes the wave operator $d^2/dt^2 \rightarrow d^2/dt^2 - c_s^2\nabla^2$. If the Lagrangian density does not depend explicitly on the spatial coordinates x^i — as in the continuous field theories considered here — then there are additional conserved quantities, whose properties we will now elucidate.

To start with, let the Lagrangian density be $\mathcal{L}(\dot{\phi}, \phi_{,i}, \phi, x^i, t)$, where, as usual $\phi_{,i} = d\phi/dx^i$ (remember d/dx^i here means space derivative at fixed time t). We have allowed here for an *explicit* dependence on time or position. In the underlying discrete lattice model, the former could describe a time variation of the spring constants, for example, and the latter could describe a system where the masses and/or springs are not all identical.

Now let's calculate $d\mathcal{L}/dx^i$, the total derivative of \mathcal{L} with respect to spatial coordinate x^i at constant time. This is

$$\frac{d\mathcal{L}}{dx^i} = \frac{\partial\mathcal{L}}{\partial\dot{\phi}} \frac{d\dot{\phi}}{dx^i} + \frac{\partial\mathcal{L}}{\partial\phi_{,j}} \frac{d\phi_{,j}}{dx^i} + \frac{\partial\mathcal{L}}{\partial\phi} \frac{d\phi}{dx^i} + \frac{\partial\mathcal{L}}{\partial x^i} \quad (16.36)$$

where summation over the repeated index j is implied. Now the penultimate term here is $(\partial\mathcal{L}/\partial\phi)\phi_{,i}$, but we can eliminate $\partial\mathcal{L}/\partial\phi$ using the Euler-Lagrange equation (16.28) to obtain

$$\frac{d\mathcal{L}}{dx^i} = \frac{\partial\mathcal{L}}{\partial\dot{\phi}} \frac{d\dot{\phi}}{dx^i} + \frac{d(\partial\mathcal{L}/\partial\dot{\phi})}{dt} \phi_{,i} + \frac{\partial\mathcal{L}}{\partial\phi_{,j}} \frac{d\phi_{,j}}{dx^i} + \frac{d(\partial\mathcal{L}/\partial\phi_{,j})}{dx^j} \phi_{,i} + \frac{\partial\mathcal{L}}{\partial x^i} \quad (16.37)$$

where we have used $d\dot{\phi}/dx^i = d\phi_{,i}/dt$ and $d\phi_{,j}/dx^i = \phi_{,ij} = d\phi_{,i}/dx^j$. Written this way, we see that the first pair of terms on the RHS are the time derivative of $(\partial\mathcal{L}/\partial\dot{\phi})\phi_{,i}$ at fixed position, and the second pair of terms are the derivative with respect to x^j of $(\partial\mathcal{L}/\partial\phi_{,j})\phi_{,i}$ at fixed time. The LHS of this equation can also be written as $d\mathcal{L}/dx^i = \delta_{ij}d\mathcal{L}/dx^j$, so, rearranging terms, we have

$$\frac{d}{dt} \left(-\frac{\partial\mathcal{L}}{\partial\dot{\phi}} \phi_{,i} \right) = -\frac{d}{dx^j} \left(\delta_{ij}\mathcal{L} - \frac{\partial\mathcal{L}}{\partial\phi_{,j}} \phi_{,i} \right) + \frac{\partial\mathcal{L}}{\partial x^i}. \quad (16.38)$$

There are three equations here, one for each of $i = 1, 2, 3$. If we now stipulate that $\partial\mathcal{L}/\partial x^i = 0$; i.e. there be no explicit dependence of the Lagrangian density on position x^i (which, in the underlying lattice model, says that all of the beads and springs are in fact identical), then each of these says that the partial time derivative of some quantity $p_i(\mathbf{x}, t)$ is minus the divergence of a vector $\mathbf{w}_i(\mathbf{x}, t)$:

$$\dot{p}_i = -\nabla \cdot \mathbf{w}_i \quad (16.39)$$

where, for example, for $i = 1$ we have

$$p_x = -(\partial\mathcal{L}/\partial\dot{\phi})\phi_{,x} \quad \text{and} \quad \mathbf{w}_x = \mathcal{L} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \phi_{,x} \begin{bmatrix} \partial\mathcal{L}/\partial\phi_{,x} \\ \partial\mathcal{L}/\partial\phi_{,y} \\ \partial\mathcal{L}/\partial\phi_{,z} \end{bmatrix}. \quad (16.40)$$

Each of equations (16.39) is just like the equation expressing conservation of electric charge: $\dot{\rho} = -\nabla \cdot \mathbf{j}$. Integrating over all space, the right hand sides of (16.39) disappear and we find

$$\frac{d\mathbf{P}}{dt} = 0 \quad (16.41)$$

where

$$\mathbf{P}(t) = \int d^3x \mathbf{p}(\mathbf{x}, t) = - \int d^3x \frac{\partial \mathcal{L}}{\partial \dot{\phi}} \nabla \phi. \quad (16.42)$$

Equation (16.41) expresses conservation of the three components of the vector \mathbf{P} .

Clearly, it was critical here to assume that the final term $\partial \mathcal{L} / \partial x^i$ in (16.38) vanish since, in general, $\int d^3x \partial \mathcal{L} / \partial x^i \neq 0$ (this can be seen most easily from (16.36) which shows that $\partial \mathcal{L} / \partial x^i \neq d\mathcal{L} / dx^i$, whose spatial integral does vanish, rather $\partial \mathcal{L} / \partial x^i$ contains three other terms which are not, in general expressible as a spatial gradient). The conservation of \mathbf{P} is a direct consequence of the invariance of the Lagrangian under shifts of position.

This is quite general. We have made no assumptions about the form of the Lagrangian density, save that it is a function only of $\dot{\phi}$, $\nabla \phi$, ϕ and possibly t . For a great many systems, however, the velocity $\dot{\phi}$ appears in the Lagrangian density only in a kinetic energy term like $\rho \dot{\phi}^2 / 2$. This would encompass systems like the one we have constructed, but with arbitrary potential energy. The conserved vector is then

$$\mathbf{P} = -\rho \int d^3x \dot{\phi} \nabla \phi. \quad (16.43)$$

What is this vector? The Lagrangian density has the dimensions of energy density, while ϕ is a displacement with units of length, so it is evident from (16.42) that \mathbf{P} has units of *momentum*. However, this quantity is quite distinct from the sum of the ‘microscopic’ conjugate momenta which, for our beads and springs model is $\sum p_i = M \sum \dot{\phi}_i$ and which, in the continuum limit becomes $\rho \int d^3x \dot{\phi}$. This is very different from the quantity appearing in (16.43) which is of second order in the field, while the summed microscopic momentum is first order. If this were not enough to convince us that these are fundamentally different entities, we might also note that the microscopic momentum vanishes for a wave, since opposite half-cycles cancel, and that, for our 1-dimensional beads and rods system, the bead momentum is perpendicular to the wave propagation direction while (16.42) is parallel.

The vector \mathbf{P} is not included in the microscopic momentum. Both \mathbf{P} and the microscopic momentum are conserved *independently*. They arise from quite different symmetries; in the discrete lattice model, conservation of the microscopic momentum follows from invariance of the Lagrangian if we shift the entire system in space; this symmetry is the homogeneity of space itself. The microscopic momentum is conserved even if the beads and springs are heterogeneous. The conservation of the wave-momentum \mathbf{P} follows from invariance of the Lagrangian under shifts in position *along* the lattice. This is a much more restrictive condition. We will call \mathbf{P} the *field-momentum* or the *wave-momentum*. The vector \mathbf{p} is the *density* of wave-momentum, and the tensor w_{ij} is the momentum flux density.

Now the other conserved quantity is the total energy E . Following the same line of argument as above, expanding the time derivative of the Lagrangian density $d\mathcal{L}/dt$ we are led to

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}} \dot{\phi} - \mathcal{L} \right) = - \frac{d}{dx^j} \left(\frac{\partial \mathcal{L}}{\partial \phi_{,j}} \dot{\phi} \right) - \frac{\partial \mathcal{L}}{\partial t}. \quad (16.44)$$

Now, if the last term $\partial \mathcal{L} / \partial t = 0$, this is

$$\dot{\epsilon} = -\nabla \cdot \mathbf{F} \quad (16.45)$$

with energy density ϵ being the Hamiltonian density as obtained before, and with energy current density $F_j = \dot{\phi} \partial \mathcal{L} / \partial \phi_{,j}$. For the BRS model, and indeed for any model where the field gradient $\nabla \phi$ enters the Lagrangian density only in a term $-\rho c_s^2 (\nabla \phi)^2 / 2$, the energy current is

$$\mathbf{F} = -c_s^2 \rho \dot{\phi} \nabla \phi = c_s^2 \mathbf{p} \quad (16.46)$$

so the energy flux is proportional to the momentum density.

Consider a wave-packet with mean wave-number \mathbf{k} and overall size $\gg 1/k$, so the packet is essentially monochromatic. If we perform the spatial integrals for the energy and wave momentum, then these are clearly localized in the region of the packet. As this packet moves, the energy and momentum are transported with it. Now the energy density is $\epsilon = \rho(\dot{\phi}^2 + (\nabla\phi)^2 + \mu^2\phi^2)/2 = \rho(\omega_{\mathbf{k}}^2 + k^2 + m^2)\phi^2/2 = \rho\omega_{\mathbf{k}}^2\phi^2$. Integrating this over space gives the total energy $E = \rho\omega_{\mathbf{k}}^2 \int d^3x \phi^2$. Similarly, the momentum is $\mathbf{P} = \rho\omega_{\mathbf{k}}\mathbf{k} \int d^3x \phi^2$. The ratio of momentum to energy for a wave packet is therefore

$$\frac{\mathbf{P}}{E} = \frac{\mathbf{k}}{\omega_{\mathbf{k}}} \quad (16.47)$$

which is compatible with the quantum mechanical relations $\mathbf{P} = \hbar\mathbf{k}$ and $E = \hbar\omega_{\mathbf{k}}$.

The energy and momentum for a large classical wave-packet can be written as

$$E = \omega_{\mathbf{k}} \left[\omega_{\mathbf{k}} \rho \int d^3x \phi^2 \right] \quad \text{and} \quad \mathbf{P} = \mathbf{k} \left[\omega_{\mathbf{k}} \rho \int d^3x \phi^2 \right]. \quad (16.48)$$

The quantity in brackets is constant for a nearly-monochromatic wave packet. It can be written as $N\hbar$, where N can be thought of as the number of fundamental quanta comprising the wave packet.

16.4 Energy and Momentum in the BRS Model

Specializing to the BRS dispersion relation $\omega_{\mathbf{k}}^2 = c_s^2 k^2 + \mu^2$ (16.31), the energy-momentum relation for a wave-packet is

$$E^2 = P^2 c_s^2 + m^2 c_s^4 \quad (16.49)$$

where m , which has dimensions of mass, is defined as

$$m = \frac{\mu \omega_{\mathbf{k}} \rho \int d^3x \phi^2}{c_s^2}. \quad (16.50)$$

Equation (16.49) is identical in form to the energy-momentum relation for a relativistic particle of mass m , but with the speed of light replaced by c_s ; the asymptotic wave-velocity for a highly energetic wave packets.

The group velocity is related to the wave-vector by $\mathbf{v}_g = c_s^2 \mathbf{k} / \omega_{\mathbf{k}}$, or equivalently by $\mathbf{v}_g = c_s^2 \mathbf{k} / \sqrt{c_s^2 k^2 + \mu^2}$ by (16.33). Solving for \mathbf{k} gives

$$\mathbf{k} = \frac{1}{\sqrt{1 - v_g^2/c_s^2}} \frac{\mu}{c_s^2} \mathbf{v}_g \quad (16.51)$$

or from (16.48), (16.50)

$$\mathbf{P} = \gamma m \mathbf{v}_g \quad \text{and} \quad E = \gamma m c_s^2 \quad (16.52)$$

with

$$\gamma \equiv (1 - v_g^2/c_s^2)^{-1/2}. \quad (16.53)$$

These results are exactly analogous to the momentum-velocity relation for a relativistic particle.

16.5 Covariance of the BRS Model

Waves in the 1-dimensional BRS model are not relativistically covariant. There is a specific inertial frame — that in which the system as a whole is stationary — in which the waves obey the dispersion relation (16.31). If we have a wave of a certain spatial frequency traveling along, then if we run alongside it it will have the same spatial frequency, but now $\omega = 0 \neq \omega_{\mathbf{k}}$. However, consider the combined transformation on the space and time coordinates

$$\begin{bmatrix} c_s t' \\ x' \end{bmatrix} = \Lambda \begin{bmatrix} c_s t \\ x \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} c_s t \\ x \end{bmatrix} = \begin{bmatrix} \gamma(c_s t - \beta x) \\ \gamma(x - \beta c_s t) \end{bmatrix}. \quad (16.54)$$

This is very similar to a Lorentz transformation. If we consider the origin in the un-primed coordinate system $x = 0$, this moves along the path $x' = -\beta\gamma c_s t$, $t' = \gamma t$, i.e. with speed $v' = dx'/dt' = -\beta c_s$. This transformation therefore corresponds to a boost with the origin in x', t' coordinates moving at velocity $v = \beta c_s$ in the un-primed coordinate space.

It is easy to show that the Lagrangian density, equation of motion for the field etc., are all invariant under this transformation. This is very useful. It tells us that if we have one solution of the field equation $\phi(\vec{x}) \equiv \phi(x, c_s t)$ then the function

$$\phi'(\vec{x}) = \phi(\Lambda\vec{x}) \quad (16.55)$$

is also a solution. Here Λ can be the transformation matrix for an arbitrary boost.

As an illustration, consider a plane wave $\phi(x, t) = \phi_0 e^{i\psi(x, t)}$, with phase $\psi(x, t) = \omega t - kx$ with $\omega = \omega_k$. Applying the transformation for a boost gives $\psi = \omega' t' - k' x'$ with

$$\begin{bmatrix} \omega'/c_s \\ k' \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} \omega/c_s \\ k \end{bmatrix}. \quad (16.56)$$

One can readily show that $\omega'^2 = c_s^2 k'^2 + \mu^2$, so ω' and k' still satisfy the dispersion relation (16.31): $\omega' = \omega_{k'}$. Thus a wave-like solution in transformed coordinates is also a solution of the field equations, but with frequency and wave-number (i.e. energy and momentum) transformed appropriately. An arbitrary solution can be written as a sum of plane-waves, so the result (16.55) is quite general.

The generalization to 2 or 3 dimensions is quite straightforward. Just as for the relativistic Lorentz transformation, distances perpendicular to the boost are unaffected.

16.6 Interactions in Classical Field Theory

The Lagrangian density (16.14) is somewhat idealized in that the spring forces are assumed to be perfectly linear in the field displacements (so the spring energies are perfectly quadratic). Real springs are quadratic only in the limit of vanishingly small displacements, and a more realistic model for the spring potential energy would be to have a potential energy density

$$V(\phi) = \rho\mu^2\phi^2/2 + \lambda\phi^4 + \dots \quad (16.57)$$

This kind of non-quadratic potential can also be realized with an ideal spring as illustrated in figure 16.4. With this modification, the pure non-interacting plane-wave solutions are no longer exact solutions of the wave equation. The non-quadratic contribution to the potential energy will introduce a *coupling*, or interaction, between the waves, and the contribution to the Hamiltonian density is called the *interaction Hamiltonian*.

This specific type of interaction is called a *self interaction* since the ϕ field interacts with itself. Another type of interaction is to couple two or more different fields. One way to introduce an interaction between two fields ϕ and χ of the type we have been discussing is shown in figure 16.5. If we assume the field displacements are small and perform an expansion we find that the stretching of the K_χ spring is

$$\Delta l = \chi \left(1 - \frac{1}{2} \frac{\phi^2}{a^2} \right) + \frac{1}{2} \frac{\phi^2}{a} \quad (16.58)$$

plus terms which are fourth order or higher in the fields. The potential energy is then

$$V(\phi, \chi) = \frac{1}{2} K_\phi \phi^2 + \frac{1}{2} K_\chi \left[\chi^2 + \frac{1}{2} \frac{\phi^2 \chi}{a} - \frac{\phi^2 \chi^2}{a^2} + \frac{1}{4} \frac{\phi^4}{a^2} \right] \quad (16.59)$$

plus terms of fifth order or higher in the fields. We see that this modification to our model has introduced two types of interaction between the fields — one of the form $\alpha\phi^2\chi$ and another proportional to $\phi^2\chi^2$ — and a ϕ^4 type self-interaction of the ϕ field.

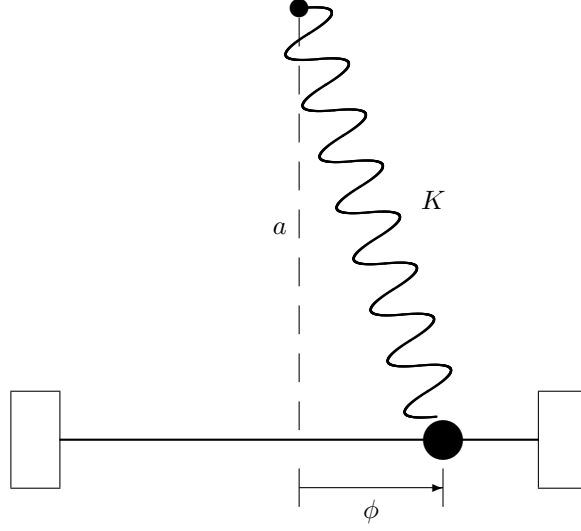


Figure 16.4: A self-interacting field can be realized with a slight modification to the original model. Here the bead slides along the horizontal rod, but is tethered by an orthogonally connected spring. The length of the spring is $l = \sqrt{a^2 + \phi^2}$. If the relaxed spring has length l_0 , the energy in the spring is $V(\phi) = K(l - l_0)^2/2$. This gives a potential which, for small displacement ϕ can be expanded as $V(\phi) = \text{constant} + \mu^2\phi^2/2 + \lambda\phi^4 + \dots$. The ‘mass term’ coefficient is $\mu = K(1 - l_0/a)/2$. If $a > l_0$ (i.e. the spring is in tension when $\phi = 0$) this is positive. The coupling coefficient is $\lambda = Kl_0/4a^3$ which is always positive. For $a < l_0$ (i.e. the spring is in compression for $\phi = 0$) the mass term μ is negative. In this case the potential has a ‘w’ shape, $\phi = 0$ is a point of unstable equilibrium, and there are two asymmetric minima $\phi = \pm\phi_0$. This is an ‘end on’ view of a chain of these units — think of multiple replicas stacked perpendicular to the page — and the connecting arms and springs are not shown.

Consider the $\mathcal{H}_{\text{int}} = \alpha\phi^2\chi$ interaction term. This will introduce a term in the Lagrangian density $\mathcal{L}_{\text{int}} = -\mathcal{H}_{\text{int}}$ and will therefore introduce a term $-\partial\mathcal{L}_{\text{int}}/\partial\chi = -\alpha\phi^2$ in the equation of motion for the χ field, which becomes

$$\ddot{\chi} - c_s^2\nabla^2\chi + \mu_\chi^2\chi + \alpha\phi^2 = 0. \quad (16.60)$$

From the point of view of the χ field, the interaction looks like a force proportional to ϕ^2 , and consequently oscillations of the ϕ field can excite oscillations of the χ field. Now the most efficient way to excite an oscillation mode of the χ field is to drive it with a force which looks like the velocity $\dot{\chi}$ for that mode (which is also a traveling wave). Consider the situation where (in the absence of the interaction) the ϕ field consists of two plane waves:

$$\phi(\vec{x}) = \cos(\vec{k}_1 \cdot \vec{x}) + \cos(\vec{k}_2 \cdot \vec{x}) \quad (16.61)$$

where $\vec{k} = (\omega/c_s, \mathbf{k})$ and where ω and \mathbf{k} satisfy the ϕ -field dispersion relation $\omega^2 = c_s^2k^2 + \mu_\phi^2$. Squaring this, we see that the force in the χ -field equation of motion is proportional to

$$\begin{aligned} \phi^2 &= \cos^2(\vec{k}_1 \cdot \vec{x}) + 2\cos(\vec{k}_1 \cdot \vec{x})\cos(\vec{k}_2 \cdot \vec{x}) + \cos^2(\vec{k}_2 \cdot \vec{x}) \\ &= 1 + \frac{1}{2}\cos(2\vec{k}_1 \cdot \vec{x}) + \cos((\vec{k}_1 + \vec{k}_2) \cdot \vec{x}) + \cos((\vec{k}_1 - \vec{k}_2) \cdot \vec{x}) + \frac{1}{2}\cos(2\vec{k}_2 \cdot \vec{x}) \end{aligned} \quad (16.62)$$

The force therefore contains four contributions which look like plane waves with various temporal and spatial frequencies. The second term here, for example is $\cos((\vec{k}_1 + \vec{k}_2) \cdot \vec{x}) = \cos((\omega_1 + \omega_2)t - (\mathbf{k}_1 + \mathbf{k}_2) \cdot \mathbf{x})$. This will be efficient at exciting oscillations of the χ field if $\Omega = \omega_1 + \omega_2$ and $\mathbf{q} = \mathbf{k}_1 + \mathbf{k}_2$ satisfy the χ -field dispersion relation $\Omega^2 = c_s^2q^2 + \mu_\chi^2$. For example, let’s assume that $\mu_\chi > \mu_\phi$. We could then collide two plane ϕ -field waves with $\omega_1 = \omega_2 = \mu_\chi/2$ and $\mathbf{k}_1 = -\mathbf{k}_2$. The result is a

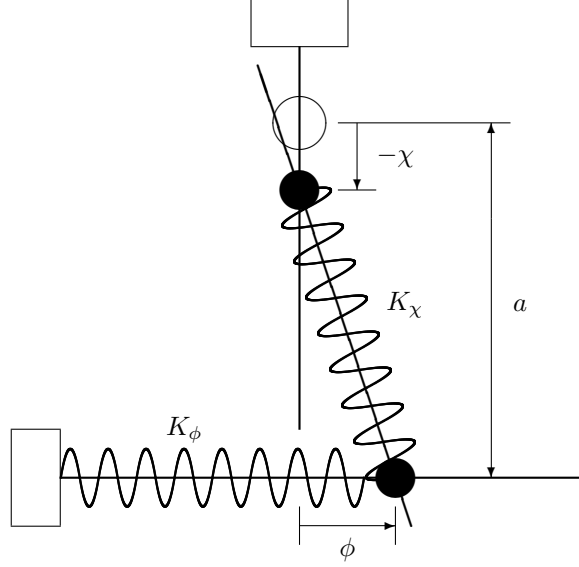


Figure 16.5: We can modify our beads and springs model to introduce an interaction term in the total Hamiltonian. To do this, we need two separate sets of coupled oscillators to represent the ϕ and χ fields. We will assume that the beads all have the same mass and that the coupling springs are identical, so the asymptotic high-frequency group velocity is the same for both types of waves. The base springs — which determined the oscillation frequency for low- k waves — will generally be different for the two fields. We now lay the ϕ oscillators on their side, and we connect the χ -bead base spring to the ϕ -bead as shown. The displacements of the beads from their rest positions (where the springs are relaxed) are indicated. This is an ‘end on’ view of the chain of these units — think of multiple replicas stacked perpendicular to the page — and the connecting arms and springs are not shown.

component in the force with temporal frequency $\Omega = \mu_\chi$ and $\mathbf{q} = \mathbf{k}_1 + \mathbf{k}_2 = 0$, which is just what is needed to excite the $\mathbf{k} = 0$ mode of the χ -field.

More generally, we expect strong coupling between a pair of ϕ -modes with spatial frequencies \mathbf{k}_1 and \mathbf{k}_2 and a χ -mode with wave-number \mathbf{q} provided the *resonance conditions*

$$\begin{aligned} \mathbf{q} &= \mathbf{k}_1 + \mathbf{k}_2 \\ \Omega_{\mathbf{q}} &= \omega_1 + \omega_2 \end{aligned} \quad (16.63)$$

are satisfied.

If we consider instead the interaction $\mathcal{H}_{\text{int}} = \lambda\phi^2\chi^2/2$ we find that the equations of motion become

$$\begin{aligned} \ddot{\phi} - c_s^2 \nabla^2 \phi + \mu_\phi^2 \phi + \lambda\chi^2 \phi &= 0 \\ \ddot{\chi} - c_s^2 \nabla^2 \chi + \mu_\chi^2 \chi + \lambda\phi^2 \chi &= 0. \end{aligned} \quad (16.64)$$

Using the same line of argument above, one can readily show that the interaction can efficiently transfer energy from a pair of ϕ -modes $\mathbf{k}_1, \mathbf{k}_2$ to a pair of χ -modes $\mathbf{q}_1, \mathbf{q}_2$ provided

$$\begin{aligned} \mathbf{q}_1 + \mathbf{q}_2 &= \mathbf{k}_1 + \mathbf{k}_2 \\ \Omega_1 + \Omega_2 &= \omega_1 + \omega_2. \end{aligned} \quad (16.65)$$

In the equation of motion for the amplitude for the mode \mathbf{q}_1 , the combination of modes $\mathbf{k}_1, \mathbf{k}_2$ and \mathbf{q}_2 produces a resonant force. Things are a little different from the $\phi\phi \rightarrow \chi$ process above, since here the interaction is not effective if, for instance, the χ -field vanishes initially. Similarly, with $H_{\text{int}} = \alpha\phi^2\chi$, the process $\chi \rightarrow \phi\phi$ is not effective classically unless there is some initial power in the ϕ field at the appropriate frequencies. We can see this from the model in figure 16.5. If $\phi = 0$ initially, then no amount of wiggling of the χ -field can excite the ϕ oscillations.

We will see that the quantum mechanical transition amplitudes for these processes contain a 4-dimensional δ -function which enforces these resonance conditions, and at the same time enforces conservation of total energy and wave-momentum.

We obtained the conservation of wave-momentum in §16.3 for a single field. However, the derivation can easily be extended to interacting fields, and it is easy to show that for any coupling of the form $H_{\text{int}} = V(\phi, \chi)$ the sum of the wave-momenta for the various fields is conserved.

16.7 Wave-Momentum Puzzles

Conservation of wave-momentum in the ‘scalar elasticity’ model considered here raises some interesting puzzles.

- Consider a ring of coupled oscillators (with beads sliding on vertical rods) mounted on a stationary turntable, and let there be a wave propagating around in a particular direction, say clockwise. This wave carries momentum around the ring, yet there is clearly no motion of any material around the ring. Now imagine there is a small amount of friction which causes the wave to damp — does the turntable absorb the momentum and start to spin? If not, where did the wave momentum go?
- For the same system, consider an external agent who applies forces to the particles in this system in order to excite a clockwise propagating wave. Does this agent experience a recoil?
- Consider a linear chain of oscillators mounted on a stationary skate-board with a momentum carrying wave propagating along it. Let the end bead be fixed to its rod, so that the wave is reflected and its momentum is reversed. Does the skate-board start moving?

What is amusing about these puzzles is that a student with only an elementary knowledge of dynamics would have no hesitation in answering (correctly) that the turntable does not spin up; that the forcing agent feels no recoil (there being no force applied in the direction of the wave-momentum) and that the skate-board does not suddenly accelerate. It is only with the benefit of the mathematically sophisticated Lagrangian treatment that we are prepared to contemplate such preposterous notions.

The resolution of the apparent conflict between the naive and sophisticated treatment in most of the above examples is that we have broken the symmetry of the system which was required in order that wave-momentum be conserved. When we apply a force, or when we pin one of the beads, the system is no longer symmetrical and wave momentum is no longer conserved. In the second example, for instance, momentum is not conserved while the force acts — so wave-momentum is created with no recoil — but once the force stops acting the wave momentum is conserved.

The first example is a little different. Here we need to think more carefully about what it means to add friction. Fundamentally, this can be thought of as exciting phonons in another lattice; that of the material comprising the turntable. We could model this, in principle, by adding some coupling term to our Lagrangian; this would allow the momentum to be transferred from the waves on our lattice to sound waves in the turntable. If the Lagrangian for the latter were spatially homogeneous then the wave-momentum would still be preserved, but would not reveal itself as net rotation; to see the momentum we would need to look carefully at the sound waves to see that they are in fact anisotropic. Realistically, the fact that the turntable is finite means that its Lagrangian density is *not* translationally invariant and therefore the wave-momentum would not be conserved. The same is true for impurities in the material which would scatter and isotropize the wave-energy.

If wave-momentum conservation is so easily broken then what use is this concept? The answer is that we believe that in reality the Lagrangian is perfectly symmetrical under spatial translations. A real impurity in a lattice isn’t really an asymmetric term in the Lagrangian, as we have pretended, rather it is an asymmetric configuration of some other field, which is coupled to our lattice waves *via* an interaction term which is symmetric under translations.

16.8 Conservation of ‘Charge’

Consider two fields $a(\mathbf{x}, t)$, $b(\mathbf{x}, t)$ obeying the free-field ‘scalar elasticity’ field equations

$$\begin{aligned}\ddot{a} - c_s^2 \nabla^2 a + \mu^2 a &= 0 \\ \ddot{b} - c_s^2 \nabla^2 b + \mu^2 b &= 0.\end{aligned}\tag{16.66}$$

These fields have the identical parameters c_s , μ but are completely independent of one another. Now multiply the first by b and the second by a and subtract. The terms involving μ cancel and we have

$$b\ddot{a} - a\ddot{b} = c_s^2 (b\nabla^2 a - a\nabla^2 b).\tag{16.67}$$

Now first term on the LHS can be written as $b\ddot{a} = \partial(b\dot{a})/\partial t - \dot{b}\dot{a}$ and similarly for the second term. The first term in parentheses on the RHS is likewise $b\nabla^2 a = \nabla \cdot (b\nabla a) - \nabla b \cdot \nabla a$ and similarly for the second. With these substitutions, the terms involving $\dot{a}\dot{b}$ on the LHS cancel, as do those involving $\nabla a \cdot \nabla b$ on the RHS, and we obtain

$$\frac{\partial}{\partial t}(b\dot{a} - a\dot{b}) = -c_s^2 \nabla \cdot (a\nabla b - b\nabla a).\tag{16.68}$$

This is clearly a *conservation law*, of the familiar form $\partial n/\partial t = -\nabla \cdot \mathbf{j}$ with ‘density’ $n = b\dot{a} - a\dot{b}$ and ‘current’ $\mathbf{j} = c_s^2 (a\nabla b - b\nabla a)$. An immediate consequence of this is that the integral of the density $Q = \int d^3x n(\mathbf{x}, t)$ is a constant.

This is very peculiar. Why should two independent fields, when combined in this way, have such a conservation law? What is the physical meaning of the conserved ‘charge’ Q here.

Before trying to answer these questions, let’s explore this in a slightly more general and also more formal manner. First, let’s consider a and b to be the real and imaginary parts of a *complex scalar field* ϕ :

$$\phi = a + ib \quad \text{and} \quad \phi^* = a - ib\tag{16.69}$$

from which a and b can be recovered as

$$a = (\phi + \phi^*)/2 \quad \text{and} \quad b = (\phi - \phi^*)/2i.\tag{16.70}$$

The Lagrangian density is

$$\mathcal{L} = \frac{\rho}{2} (\dot{a}^2 - c_s^2 (\nabla a)^2 - \mu^2 a^2 + \dot{b}^2 - c_s^2 (\nabla b)^2 - \mu^2 b^2)\tag{16.71}$$

which, in terms of ϕ , ϕ^* is

$$\mathcal{L} = \frac{\rho}{2} (\dot{\phi}\dot{\phi}^* - c_s^2 \nabla\phi \cdot \nabla\phi^* - \mu^2 \phi\phi^*).\tag{16.72}$$

In (16.72), $\dot{\phi}$, $\dot{\phi}^*$, $\nabla\phi$, $\nabla\phi^*$, ϕ and ϕ^* are all considered to be *independent* fields. (This is just like how the real scalar field Lagrangian $\mathcal{L}(\dot{\phi}, \nabla\phi, \phi)$ is considered to be a function of three independent fields ϕ , $\dot{\phi}$ and $\nabla\phi$). The equations of motion are obtained, as usual, by requiring that the action $S = \int dt \int d^3x \mathcal{L}$ be stationary with respect to variations $\phi \rightarrow \phi + \delta\phi$ and $\phi^* \rightarrow \phi^* + \delta\phi^*$. The former variation gives

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}} \right) + \frac{d}{dx^i} \left(\frac{\partial \mathcal{L}}{\partial \phi_{,i}} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0\tag{16.73}$$

which, in our case, yields

$$\ddot{\phi}^* - c_s^2 \nabla^2 \phi^* + \mu^2 \phi^* = 0\tag{16.74}$$

and the latter gives an identical equation for ϕ . These equations of motion, with (16.69) are equivalent to equations (16.66).

Now the Lagrangian density (16.72) is rather symmetrical. In particular, it is invariant under the transformation

$$\phi \rightarrow \phi' = e^{i\theta} \phi \quad \text{and} \quad \phi^* \rightarrow \phi'^* = e^{-i\theta} \phi^*\tag{16.75}$$

where θ is an arbitrary constant. This provides a nice way to obtain the conservation law (16.68). We can think of the transformed field ϕ' as a function of \mathbf{x} , t and θ , but the invariance of \mathcal{L} implies that total derivative of \mathcal{L} with respect to the parameter θ vanishes:

$$0 = \frac{d\mathcal{L}}{d\theta} = \frac{\partial\mathcal{L}}{\partial\dot{\phi}} \frac{\partial\dot{\phi}}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\phi_{,i}} \frac{\partial\phi_{,i}}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\phi} \frac{\partial\phi}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\dot{\phi}^*} \frac{\partial\dot{\phi}^*}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\phi_{,i}^*} \frac{\partial\phi_{,i}^*}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\phi^*} \frac{\partial\phi^*}{\partial\theta}. \quad (16.76)$$

Using the equations of motion (16.73) and its complex conjugate to eliminate $\partial\mathcal{L}/\partial\phi$ and $\partial\mathcal{L}/\partial\phi^*$ and using $\partial\phi/\partial\theta = \partial(\partial\phi/\partial\theta)/\partial t = i\partial\phi/\partial t$ etc. this becomes

$$\frac{\partial n}{\partial t} = -\nabla \cdot \mathbf{j} \quad (16.77)$$

with density

$$n = i \left(\frac{\partial\mathcal{L}}{\partial\dot{\phi}} \phi - \frac{\partial\mathcal{L}}{\partial\dot{\phi}^*} \phi^* \right) \quad (16.78)$$

and current

$$\mathbf{j} = i \left(\frac{\partial\mathcal{L}}{\partial\nabla\phi} \phi - \frac{\partial\mathcal{L}}{\partial\nabla\phi^*} \phi^* \right). \quad (16.79)$$

This is an example of *Noether's theorem*, which says that if the Lagrangian density is invariant under some continuous transformation such as (16.75) there is a corresponding conservation law. This specific type of transformation is known as a *global gauge transformation*.

For the free field Lagrangian (16.72) the density and current are

$$n = i(\phi\dot{\phi}^* - \dot{\phi}\phi^*) \quad (16.80)$$

and

$$\mathbf{j} = ic_s^2(\phi^*\nabla\phi - \phi\nabla\phi^*) \quad (16.81)$$

which we could have readily obtained from (16.68) using (16.70).

The current density \mathbf{j} is reminiscent of the momentum density \mathbf{p} . Recall that conservation of the wave-momentum for a single real field was obtained by considering the total derivative of the Lagrangian density with respect to position: $d\mathcal{L}/d\mathbf{x}$, which yielded $\mathbf{p} = \nabla\phi\partial\mathcal{L}/\partial\dot{\phi}$. Applying the same line of reasoning for the complex field yields

$$\mathbf{p} = \nabla\phi \frac{\partial\mathcal{L}}{\partial\dot{\phi}} + \nabla\phi^* \frac{\partial\mathcal{L}}{\partial\dot{\phi}^*} \quad (16.82)$$

or, for the Lagrangian (16.72),

$$\mathbf{p} = -\frac{\rho}{2}(\dot{\phi}\nabla\phi^* + \dot{\phi}^*\nabla\phi). \quad (16.83)$$

With (16.69) this is easily shown to be equivalent to the sum of the momentum densities for the fields a and b :

$$\mathbf{p} = -\rho(\dot{a}\nabla a + \dot{b}\nabla b). \quad (16.84)$$

Now consider a plane-wave $\phi = e^{i(\omega_{\mathbf{k}}t - \mathbf{k}\cdot\mathbf{x})}$. This has momentum density $\mathbf{p} = \rho\omega_{\mathbf{k}}\mathbf{k}\phi\phi^*$ which is parallel to $\hat{\mathbf{k}}$, as befits a wave propagating in the direction $\hat{\mathbf{k}}$, and the same is true for the wave $\phi' = e^{-i(\omega_{\mathbf{k}}t - \mathbf{k}\cdot\mathbf{x})}$ which also propagates in the direction $\hat{\mathbf{k}}$. The *current* for the field ϕ is $i(\phi\nabla\phi^* - \phi^*\nabla\phi)$, which is also parallel to $\hat{\mathbf{k}}$, but the current for the field ϕ' has opposite sign. The same is true for wave packets; a packet with positive frequency $\phi \propto e^{+i\omega_{\mathbf{k}}t}$ has positive charge Q , which it transports in the direction $\hat{\mathbf{k}}$, giving a positive current. A negative frequency wave packet, with $\phi \propto e^{-i\omega_{\mathbf{k}}t}$, has a negative charge and has a negative current in the direction $\hat{\mathbf{k}}$.

The 'charge' for these quasi-monochromatic wave packets here is very similar to circular polarization for an electro-magnetic field. What we have called a positively charged field is one in which the field a lags b by 90 degrees and *vice versa*. If the two fields are in phase ($a = b$) we have a linearly polarized wave and the charge density and current vanish. There is a difference, however,

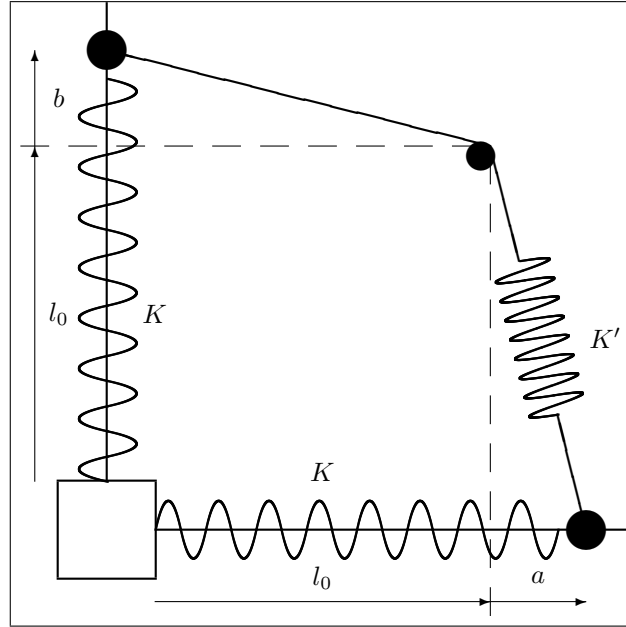


Figure 16.6: This modification to the BRS model provides a interaction between two fields $a(\mathbf{x}, t)$ and $b(\mathbf{x}, t)$. What we have here is two beads on rods with springs, much as in the original model, but with a cord attached to bead b with passes over the frictionless pulley at the upper right and then connects to bead b through the spring with spring constant K' . The length of the K springs, when relaxed, is l_0 . If the relaxed length of the cord plus spring connecting the beads is l'_0 , then, for small a, b the potential energy is $V(a, b) = K(a^2 + b^2)/2 + K'((2l_0 - l'_0) + (a^2 + b^2)/2l_0 + \dots)^2$. This is of the form $V(a, b) = \mu a^2/2 + \mu b^2/2 + \lambda(a^2 + b^2)^2 + \dots$. The first two terms are the usual free-field mass terms while the last term provides a self-interaction. If we represent the fields as $\phi = a + ib$ then the potential term is $V(\phi) = \lambda|\phi|^4$. This interaction respects the conservation law (16.68).

in that the electro-magnetic field vector \mathbf{E} ‘lives’ in the real 3-dimensional space, whereas here the vector $\phi = (a, b)$ exists in a relatively abstract internal 2-dimensional space.

For free fields, the physical meaning of all this is somewhat questionable, since, as emphasized at the outset, the fields a, b are completely independent of one another. The mysterious conservation law (16.68) is nothing more than the statement that

$$\frac{\ddot{a} - c_s^2 \nabla^2 a}{a} = \frac{\ddot{b} - c_s^2 \nabla^2 b}{b} \quad (16.85)$$

i.e. the parameter μ is the same in the two equations of motion. The *physical* implication of this kind of conservation law only really emerges we consider a pair of interacting fields. We can introduce an interaction between the a and b fields by adding a term $-V(a, b)$ to the Lagrangian density \mathcal{L} . This in turn adds a term $-\partial V/\partial a$ to the RHS of the equation of motion for a in (16.66) and similarly for b . In general, this will violate the conservation law (16.68), since we add to the RHS a term $a\partial V/\partial b - b\partial V/\partial a$. However, under certain circumstances this term will also vanish. It is easy to see that this will be the case if the interaction potential V is a function of $a^2 + b^2$: i.e. $V(a, b) = F(a^2 + b^2)$, in which case $a\partial V/\partial b = 2abF' = b\partial V/\partial a$. This result is neatly expressed in the complex field formalism; this particular potential is equivalently $V(\phi) = F(\phi\phi^*)$, which is clearly invariant under the global gauge transformation (16.75), so the Lagrangian density for this interacting field theory is also invariant. A realization of such an interaction is illustrated in figure 16.6.

What this analysis shows is that a field theory for a pair of fields with identical mass parameter μ can naturally accommodate two types of waves, which transport a conserved ‘charge’. These two types of waves are the positive and negative frequency components of the complex field ϕ . This is

rather different from the single real field, where the negative frequency Fourier components are just the mirror image, with conjugation, of the positive frequency components. Here the positive and negative frequency components are quite independent, and positive (negative) frequency components carry positive (negative) ‘charge’. In the quantum theory of such a field, the corresponding quanta are charged particles and anti-particles. Precisely what this ‘charge’ is depends, essentially, on the interactions with other fields. It turns one can add an interaction with the electro-magnetic field, for instance, so that the charge here is electric charge. However, this is not the only possibility, and the conserved quantity could be some other entity. Other possibilities are to have a field with more than two components, and in this way one can construct theories with more complicated conservation laws.

16.9 Conservation of Particle Number

In addition to energy and the three components of the wave-momentum — which are precisely conserved in interacting theories — a non-interacting field possesses a fifth conserved quantity, which corresponds to conservation of number of particles. We will first obtain this directly from the form of the solutions. We then show how the conservation law can be understood as arising from a symmetry of the free-field Lagrangian, when written in an appropriate manner.

The general solution of the free field equation of motion

$$\ddot{\phi} - c_s^2 \nabla^2 \phi + \mu^2 \phi = 0 \quad (16.86)$$

can be written as

$$\phi(\mathbf{x}, t) = \phi^+ + \phi^- = \sum_{\mathbf{k}} \phi_{\mathbf{k}} e^{i(\omega_{\mathbf{k}} t - \mathbf{k} \cdot \mathbf{x})} + \sum_{\mathbf{k}} \phi_{\mathbf{k}}^* e^{-i(\omega_{\mathbf{k}} t - \mathbf{k} \cdot \mathbf{x})}. \quad (16.87)$$

The complex Fourier amplitudes can be related to the transform of the field and its time derivative on some initial time-slice $t = t_0$ as

$$\phi_{\mathbf{k}} = \frac{1}{2} \left(\tilde{\phi}(\mathbf{k}, t_0) + \tilde{\dot{\phi}}(\mathbf{k}, 0)/i\omega_{\mathbf{k}} \right) e^{-i\omega_{\mathbf{k}} t_0} \quad (16.88)$$

where

$$\tilde{\phi}(\mathbf{k}, t) = \int d^3x \phi(\mathbf{x}, t) e^{i\mathbf{k} \cdot \mathbf{x}} \quad \text{and} \quad \tilde{\dot{\phi}}(\mathbf{k}, t) = \int d^3x \dot{\phi}(\mathbf{x}, t) e^{i\mathbf{k} \cdot \mathbf{x}}. \quad (16.89)$$

Now consider the quantities

$$n(\mathbf{x}, t) = i(\phi^+ \dot{\phi}^- - \phi^- \dot{\phi}^+) \quad (16.90)$$

and

$$\mathbf{j}(\mathbf{x}, t) = -ic_s^2(\phi^+ \nabla \phi^- - \phi^- \nabla \phi^+). \quad (16.91)$$

Since ϕ^+ and ϕ^- are complex conjugates of each other both n and \mathbf{j} are real. In terms of the Fourier coefficients $\phi_{\mathbf{k}}$ they are

$$n(\mathbf{x}, t) = \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* (\omega_{\mathbf{k}} + \omega_{\mathbf{k}'}) e^{i((\omega_{\mathbf{k}} - \omega_{\mathbf{k}'})t - (\mathbf{k} - \mathbf{k}') \cdot \mathbf{x})} \quad (16.92)$$

and

$$\mathbf{j}(\mathbf{x}, t) = c_s^2 \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* (\mathbf{k} + \mathbf{k}') e^{i((\omega_{\mathbf{k}} - \omega_{\mathbf{k}'})t - (\mathbf{k} - \mathbf{k}') \cdot \mathbf{x})}. \quad (16.93)$$

The time derivative of $n(\mathbf{x}, t)$ is

$$\frac{\partial n}{\partial t} = i \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* (\omega_{\mathbf{k}}^2 - \omega_{\mathbf{k}'}^2) e^{i((\omega_{\mathbf{k}} - \omega_{\mathbf{k}'})t - (\mathbf{k} - \mathbf{k}') \cdot \mathbf{x})} \quad (16.94)$$

whereas the divergence of $\mathbf{j}(\mathbf{x}, t)$ is

$$\nabla \cdot \mathbf{j} = -ic_s^2 \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* (\mathbf{k}^2 - \mathbf{k}'^2) e^{i((\omega_{\mathbf{k}} - \omega_{\mathbf{k}'})t - (\mathbf{k} - \mathbf{k}') \cdot \mathbf{x})}. \quad (16.95)$$

However, the frequency is defined to be $\omega_{\mathbf{k}}^2 = c^2 k^2 + \mu^2$, so we have

$$\frac{\partial n}{\partial t} = -\nabla \cdot \mathbf{j}. \quad (16.96)$$

Thus we have a conservation law, relating the density (16.90) and corresponding current (16.91).

For a single plane wave, with amplitude ϕ_0 , the density and current are

$$\begin{aligned} n &= \phi_0^2 \omega_{\mathbf{k}} \\ \mathbf{j} &= c_s^2 \phi_0^2 \mathbf{k} \end{aligned} \quad (16.97)$$

so $(c_s n, \mathbf{j})$ form a four-vector (the amplitude of a wave being an invariant also). Note that the current in this case is just equal to the density times the group velocity: $\mathbf{j} = \mathbf{v}_g n$.

The same is true for a large wave-packet, for which the global conserved quantity is

$$N = \int d^3x n(\mathbf{x}, t) = \omega_{\mathbf{k}} \int d^3x \phi^2. \quad (16.98)$$

As already discussed, this quantity corresponds to the *particle number*.

Another interesting model is a statistically homogeneous random field, for which $\langle \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* \rangle = P_{\phi}(\mathbf{k}) \delta_{\mathbf{k}\mathbf{k}'}$ where P_{ϕ} is the power spectrum. This is working in a periodic box, where the Fourier modes are discrete. If we think of the modes as being continuous, this becomes $\langle \phi_{\mathbf{k}} \phi_{\mathbf{k}'}^* \rangle = (2\pi)^3 P_{\phi}(\mathbf{k}) \delta(\mathbf{k} - \mathbf{k}')$. The expectation values of the density and current are then

$$\begin{aligned} \langle n \rangle &= \sum_{\mathbf{k}} \langle |\phi_{\mathbf{k}}|^2 \rangle \omega_{\mathbf{k}} \rightarrow \int \frac{d^3k}{(2\pi)^3} P_{\phi}(\mathbf{k}) \omega_{\mathbf{k}} \\ \langle \mathbf{j} \rangle &= c_s^2 \sum_{\mathbf{k}} \langle |\phi_{\mathbf{k}}|^2 \rangle \mathbf{k} \rightarrow \int \frac{d^3k}{(2\pi)^3} P_{\phi}(\mathbf{k}) \mathbf{k} \end{aligned} \quad (16.99)$$

The quantity $\omega_{\mathbf{k}} P_{\phi}(\mathbf{k})$ plays the role, in wave mechanics, of the phase space density $f(\mathbf{p})$ in particle dynamics. Like $f(\mathbf{p})$ it is invariant under boosts (see below). Integrating over all momenta — i.e. all values of wave-number \mathbf{k} — gives the particle number density; multiplying by \mathbf{k} and integrating gives the momentum density and multiplying by $\omega_{\mathbf{k}}$ and integrating gives the energy density.

The conservation law (16.96) was obtained directly from the properties of the solution of the field equations. It is also interesting to see how this emerges as a consequence of a symmetry of the Lagrangian. With $\phi = \phi^+ + \phi^-$ the Lagrangian density is

$$\begin{aligned} \mathcal{L} &= \frac{\rho}{2} ((\dot{\phi}^+)^2 - c_s^2 (\nabla \phi^+)^2 - \mu^2 \phi^{+2}) + \\ &\quad \rho (\phi^+ \dot{\phi}^- - c_s^2 \nabla \phi^+ \cdot \nabla \phi^- - \mu^2 \phi^+ \phi^-) + \\ &\quad \frac{\rho}{2} ((\dot{\phi}^-)^2 - c_s^2 (\nabla \phi^-)^2 - \mu^2 \phi^{-2}). \end{aligned} \quad (16.100)$$

If we vary either ϕ^+ or ϕ^- we obtain the same equation of motion:

$$\ddot{\phi}^+ - c_s^2 \nabla^2 \phi^+ - \mu^2 \phi^+ + \ddot{\phi}^- - c_s^2 \nabla^2 \phi^- - \mu^2 \phi^- = 0. \quad (16.101)$$

This is just the equation of motion for ϕ , with $\phi \rightarrow \phi^+ + \phi^-$. However, consider the Lagrangian density

$$\mathcal{L}' = \rho (\dot{\phi}^+ \dot{\phi}^- - c_s^2 \nabla \phi^+ \cdot \nabla \phi^- - \mu^2 \phi^+ \phi^-). \quad (16.102)$$

Varying ϕ^+ and ϕ^- now yield the field equations

$$\begin{aligned} \ddot{\phi}^+ - c_s^2 \nabla^2 \phi^+ - \mu^2 \phi^+ &= 0 \\ \text{and} \\ \ddot{\phi}^- - c_s^2 \nabla^2 \phi^- - \mu^2 \phi^- &= 0. \end{aligned} \quad (16.103)$$

But ϕ^+ and ϕ^- are complex conjugates, so any field $\phi^+(\mathbf{x}, t)$ which is a solution of the field equations derived from \mathcal{L}' is automatically a solution of the field equations derived from \mathcal{L} . Thus, we can take the Lagrangian density for the system to be \mathcal{L}' given by (16.102), since it generates equivalent solutions. Now this Lagrangian, like that for ϕ has no explicit time or position dependence, so energy and the three components of momentum are all conserved. However, (16.102) has an additional symmetry: it is invariant under the global gauge transformation $\phi^+ \rightarrow \phi^+ e^{i\theta}$ and $\phi^- \rightarrow \phi^- e^{-i\theta}$. This transformation corresponds to variation of the choice of initial time slice ($t = 0$) on which we determine ϕ^+ and ϕ^- . The conservation law (16.96) can readily be obtained by requiring that the total derivative of the Lagrangian density with respect to θ vanish, just as we did to obtain charge conservation for the complex field.

It should not come as a surprise that the free field has more conserved quantities than just the energy and momentum density. The Fourier amplitudes $\phi_{\mathbf{k}}$ can be determined from the displacement and velocity data $\phi(\mathbf{x}, t)$, $\dot{\phi}(\mathbf{x}, t)$ on any time slice, so there are in effect a triply infinite number of conserved quantities. However, this behaviour is specific to the free field. For the free field, the 4-dimensional Fourier modes are confined to the 3-surface $\omega_{\mathbf{k}}^2 = c_s^2 k^2 + \mu^2$, so the information content in the 4-dimensional field $\phi(\mathbf{x}, t)$ is really only 3-dimensional. If we admit interactions, this will no longer be the case. A necessary consequence of any interactions is that the energy shell will ‘fuzz-out’ somewhat, and it is no longer possible to determine the entire space-time behaviour from measurements on one time slice.

Consider adding an interaction term $\mathcal{L}_{\text{int}} = \rho\lambda\phi^4$ to the free-field Lagrangian. In terms of ϕ^+ , ϕ^- this is

$$\mathcal{L}_{\text{int}} = \rho\lambda^4((\phi^+)^4 + 4(\phi^+)^3\phi^- + 6(\phi^+)^2(\phi^-)^2 + 4\phi^+(\phi^-)^3 + (\phi^-)^4). \quad (16.104)$$

However, only the central term here respects the global gauge transformation symmetry, so adding interactions therefore violates conservation of particle number.

16.10 Particle Number Conservation at Low Energies

We saw in the previous section that in the wave mechanics of a continuous medium there are four exactly conserved quantities — the energy and the three components of the wave-momentum — but that particle number is only conserved for non-interacting fields. Now particle number is also violated in interactions of high energy particles; if we collide very energetic (i.e. highly relativistic) particles then there is ample energy to create new particles that were not present in the initial state. However, if we collide particles with kinetic energy $E \ll m_0 c^2$, then there is not enough energy to overcome the threshold to produce new particles, so particle number is conserved at low energies. We will now see how particle conservation is conserved in a classical interacting field theory, in the limit that the wavelength is large compared to $\lambda_* = 2\pi/k_* = 2\pi c_s/\mu$. For the BRS model this length scale plays the role of the *Compton wavelength*, so the condition $\lambda \gg \lambda_*$ corresponds to the highly non-relativistic limit. In essence, what happens is that in the limit $k \ll k_*$ the field ϕ has rapid temporal oscillation at frequency $\omega \simeq \mu$ with a relatively slowly varying ‘envelope’. The evolution of this envelope, it turns out, is governed by an equation which is very similar to the Schroedinger equation. The *correspondence principle*, which states that non-relativistic particle dynamics phenomena can be equally well described by Schroedinger’s wave mechanics — then carries over, with a little modification, to the scalar elasticity waves.

Consider first non-interacting waves ($\lambda = 0$), for which the wave equation is

$$\ddot{\phi} - c_s^2 \nabla^2 \phi + \mu^2 \phi = 0. \quad (16.105)$$

In the limit we are considering, i.e. $c_s k \ll \mu$, the second term is much smaller in magnitude than the third, so the first and third terms must be nearly equal. If we neglect the second term entirely, the equation of motion is $\ddot{\phi} + \mu^2 \phi = 0$, the solutions of which are $\phi(\mathbf{x}, t) = \phi_0(\mathbf{x}) \exp(\pm i\mu t)$; i.e. the field just sits there without moving, aside from wiggling up and down at frequency $\omega = \mu$. What we would like to do is to divide out the rapid time oscillation and develop an equation of motion for

the relatively slowly varying modulating function. This is a little tricky since a general ϕ field will contain both positive and negative frequencies: $\phi = \phi^+ + \phi^-$ as in (16.87). If the spatial frequency of the waves is $k \leq k_{\max}$, then the positive energy part of the wave is limited to a narrow band of spatial frequencies $\mu \leq \omega \leq \mu\sqrt{1 + c_s^2 k_{\max}^2}/\mu^2$, and similarly for the negative frequency modes. Thus, we can let

$$\phi(\mathbf{x}, t) = \psi(\mathbf{x}, t)e^{i\mu t} + \psi^*(\mathbf{x}, t)e^{-i\mu t} \quad (16.106)$$

where $\psi(\mathbf{x}, t)$ and its conjugate are band-limited with temporal frequencies in the range $0 \leq \omega \leq c_s^2 k_{\max}^2/2\mu$. The free field Lagrangian density is then

$$\begin{aligned} \mathcal{L} &= \frac{\rho}{2}(\dot{\phi}^2 - c_s^2(\nabla\phi)^2 - \mu^2\phi^2) \\ &= \frac{\rho}{2}e^{2i\mu t}[(\psi + i\mu\psi)^2 - c_s^2(\nabla\psi)^2 - \mu^2\psi^2] \\ &\quad + \rho[(\psi + i\mu\psi)(\dot{\psi}^* - i\mu\psi^*) - c_s^2\nabla\psi \cdot \nabla\psi^* - \mu^2\psi\psi^*] \\ &\quad + \frac{\rho}{2}e^{-2i\mu t}[(\dot{\psi}^* - i\mu\psi^*)^2 - c_s^2(\nabla\psi^*)^2 - \mu^2\psi^{*2}], \end{aligned} \quad (16.107)$$

and for a $\lambda\phi^4$ interaction term this is augmented with

$$\mathcal{L}_{\text{int}} = -\rho\lambda\phi^4 = -\rho\lambda(\psi^4 e^{4i\mu t} + 4\psi^3\psi^* e^{2i\mu t} + 6\psi^2\psi^{*2} + 4\psi\psi^{*3} e^{-2i\mu t} + \psi^{*4} e^{-4i\mu t}). \quad (16.108)$$

Thus the transformation (16.106) has resulted in an explicit time dependence in the Lagrangian, with terms containing rapidly oscillating factors $e^{\pm 2i\mu t}$ and $e^{\pm 4i\mu t}$. However — and this is a key point — if we are dealing with low spatial frequency waves $k \leq k_{\max}$, so the temporal frequency for the ψ field is $\omega \leq k_{\max}^2/2\mu \ll \mu$, the contribution to the action $S = \int dt \int d^3x \mathcal{L}$ from these terms is very small since all the other factors are relatively very slowly varying. Thus, for $k \ll \mu/c_s$ we can ignore most of the terms above and use the effective Lagrangian

$$\mathcal{L} = \rho[\dot{\psi}\dot{\psi}^* + i\mu(\psi\dot{\psi}^* - \psi^*\dot{\psi}) - c_s^2\nabla\psi \cdot \nabla\psi^* - 6\lambda\psi^2\psi^{*2}]. \quad (16.109)$$

This can be further simplified, since, in the regime we are considering, $\dot{\psi} \ll \mu\psi$ so we can neglect the first term as compared to the second. We can therefore take the Lagrangian to be

$$\mathcal{L} = \rho[i\mu(\psi\dot{\psi}^* - \psi^*\dot{\psi}) - c_s^2\nabla\psi \cdot \nabla\psi^* - 6\lambda\psi^2\psi^{*2}]. \quad (16.110)$$

The equation of motion is obtained, as always, by requiring that the action be stationary with respect to variation of the fields ψ, ψ^* . For the variation $\psi^* \rightarrow \psi^* + \delta\psi^*$ this yields

$$\frac{d(\partial\mathcal{L}/\partial\dot{\psi}^*)}{dt} + \frac{d(\partial\mathcal{L}/\partial\psi_{,j}^*)}{dx^j} - \partial\mathcal{L}/\partial\psi^* = 0 \quad (16.111)$$

or equivalently

$$i\dot{\psi} - \frac{c_s^2}{2\mu}\nabla^2\psi + \frac{6\lambda}{\mu}\psi^2\psi^* = 0 \quad (16.112)$$

and the variation $\psi \rightarrow \psi + \delta\psi$ yields the complex conjugate of (16.112). This is very similar in form to the *time dependent Schroedinger equation* for the wave-function $\psi(\mathbf{x}, t)$ for a particle in a potential, but with the dynamical quantity $6\lambda\psi^2\psi^*/\mu$ in place of the potential $V(\mathbf{x})$. We will discuss this connection further below.

The Lagrangian (16.110), like (16.102), has no explicit dependence on t or on \mathbf{x} , so energy and momentum are conserved, and it is also symmetric under the global gauge transformation $\psi \rightarrow \psi e^{i\theta}$ and $\psi^* \rightarrow \psi^* e^{-i\theta}$, so it therefore also has a conserved particle number.

Conservation of particle number follows from the vanishing of the total derivative of the Lagrangian with respect to the global gauge transformation parameter:

$$0 = \frac{d\mathcal{L}}{d\theta} = \frac{\partial\mathcal{L}}{\partial\psi} \frac{\partial\psi}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\dot{\psi}} \frac{\partial\dot{\psi}}{\partial\theta} + \frac{\partial\mathcal{L}}{\partial\psi_{,j}} \frac{\partial\psi_{,j}}{\partial\theta} + \dots \psi \rightarrow \psi^* \dots \quad (16.113)$$

Using the Euler-Lagrange equation to replace $\partial\mathcal{L}/\partial\psi$ by $d(\partial\mathcal{L}/\partial\dot{\psi})/dt + d(\partial\mathcal{L}/\partial\psi_{,i})/dx^i$ and with $\partial\psi/\partial\theta = i\psi$, $\partial\dot{\psi}/\partial\theta = i\dot{\psi}$ etc this becomes

$$i\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\psi}}\psi - \frac{\partial\mathcal{L}}{\partial\dot{\psi}^*}\psi^*\right) = -i\frac{d}{dx^i}\left(\frac{\partial\mathcal{L}}{\partial\psi_{,i}}\psi - \frac{\partial\mathcal{L}}{\partial\psi_{,i}^*}\psi^*\right) \quad (16.114)$$

or

$$\frac{\partial n}{\partial t} = -\nabla \cdot \mathbf{j} \quad (16.115)$$

where the particle number density is

$$n = 2\rho\mu\psi\psi^* \quad (16.116)$$

and the particle flux density is

$$\mathbf{j} = i\rho c_s^2(\psi^*\nabla\psi - \psi\nabla\psi^*). \quad (16.117)$$

Conservation of momentum is obtained from the total spatial derivative of the Lagrangian density:

$$\frac{d\mathcal{L}}{dx^i} = \frac{\partial\mathcal{L}}{\partial\psi}\frac{\partial\psi}{\partial x^i} + \frac{\partial\mathcal{L}}{\partial\dot{\psi}}\frac{\partial\dot{\psi}}{\partial x^i} + \frac{\partial\mathcal{L}}{\partial\psi_{,j}}\frac{\partial\psi_{,j}}{\partial x^i} + \dots \psi \rightarrow \psi^* \dots \quad (16.118)$$

Using the Euler-Lagrange equation as above and using the commutation property of partial derivatives this becomes

$$\frac{d}{dt}\left(-\frac{\partial\mathcal{L}}{\partial\dot{\psi}}\psi_{,i} - \frac{\partial\mathcal{L}}{\partial\dot{\psi}^*}\psi_{,i}^*\right) = -\frac{d}{dx^j}\left(-\frac{\partial\mathcal{L}}{\partial\psi_{,j}}\psi_{,i} - \frac{\partial\mathcal{L}}{\partial\psi_{,j}^*}\psi_{,i}^* + \delta_{ij}\mathcal{L}\right) \quad (16.119)$$

or

$$\frac{\partial p^i}{\partial t} = -\frac{\partial w^{ij}}{\partial x^j} \quad (16.120)$$

where the momentum density is

$$\mathbf{p} = i\rho\mu(\psi^*\nabla\psi - \psi\nabla\psi^*) \quad (16.121)$$

and the momentum flux tensor is

$$w^{ij} = \rho c_s^2(\psi_{,i}\psi_{,j}^* + \psi_{,i}^*\psi_{,j} - \delta_{ij}[\nabla^2(\psi\psi^*) - \frac{6\lambda}{c_s^2}\psi^2\psi^{*2}]). \quad (16.122)$$

Conservation of energy is obtained from the total time derivative of the Lagrangian density:

$$\frac{d\mathcal{L}}{dt} = \frac{\partial\mathcal{L}}{\partial\psi}\frac{\partial\psi}{\partial t} + \frac{\partial\mathcal{L}}{\partial\dot{\psi}}\frac{\partial\dot{\psi}}{\partial t} + \frac{\partial\mathcal{L}}{\partial\psi_{,j}}\frac{\partial\psi_{,j}}{\partial t} + \dots \psi \rightarrow \psi^* \dots \quad (16.123)$$

Again, using the Euler-Lagrange equations and the commutation property of partial derivatives, this becomes

$$\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\psi}}\dot{\psi} + \frac{\partial\mathcal{L}}{\partial\dot{\psi}^*}\dot{\psi}^* - \mathcal{L}\right) = -\frac{d}{dx^j}\left(\frac{\partial\mathcal{L}}{\partial\psi_{,j}}\dot{\psi} + \frac{\partial\mathcal{L}}{\partial\psi_{,j}^*}\dot{\psi}^*\right). \quad (16.124)$$

Using the equations of motion to replace $i\dot{\psi}$ by $c_s^2\nabla^2\psi/2\mu - 6\lambda\psi^2\psi^*/\mu$ and similarly for $i\dot{\psi}^*$ this becomes

$$\frac{\partial\epsilon}{\partial t} = -\nabla \cdot \mathbf{F} \quad (16.125)$$

where the energy density is

$$\epsilon = \rho(c_s^2\nabla\psi \cdot \nabla\psi^* + 6\lambda\psi^2\psi^{*2}) \quad (16.126)$$

and the energy flux vector is

$$\mathbf{F} = i\rho c_s^2\left[\frac{c_s^2}{2\mu}(\nabla\psi^*\nabla^2\psi - \nabla\psi\nabla^2\psi^*) + \frac{6\lambda}{\mu}\psi\psi^*(\psi^*\nabla\psi - \psi\nabla\psi^*)\right] \quad (16.127)$$

To summarize, the conservation laws are

$$\begin{aligned} \text{number :} & \quad \partial n / \partial t = -\nabla \cdot \mathbf{j} \\ \text{momentum :} & \quad \partial p^i / \partial t = -\partial w^{ij} / \partial x^j \\ \text{energy :} & \quad \partial \epsilon / \partial t = -\nabla \cdot \mathbf{F} \end{aligned} \quad (16.128)$$

where

$$\begin{aligned} n &= 2\rho\mu\psi\psi^* \\ \mathbf{j} &= i\rho c_s^2(\psi^*\nabla\psi - \psi\nabla\psi^*) \\ \mathbf{p} &= i\rho\mu(\psi^*\nabla\psi - \psi\nabla\psi^*) \\ w^{ij} &= \rho c_s^2(\psi_{,i}\psi_{,j}^* + \psi_{,i}^*\psi_{,j} - \delta_{ij}[\nabla^2(\psi\psi^*) - \frac{6\lambda}{c_s^2}\psi^2\psi^{*2}]) \\ \epsilon &= \rho(c_s^2\nabla\psi \cdot \nabla\psi^* + 6\lambda\psi^2\psi^{*2}) \\ \mathbf{F} &= i\rho c_s^2[\frac{c_s^2}{2\mu}(\nabla\psi^*\nabla^2\psi - \nabla\psi\nabla^2\psi^*) + \frac{6\lambda}{\mu}\psi\psi^*(\psi^*\nabla\psi - \psi\nabla\psi^*)] \end{aligned} \quad (16.129)$$

There is a close parallel here with Schroedinger's development of quantum mechanics, where the state for a particle is represented by a 'wave-function' $\psi(\mathbf{x}, t)$. The density n is, aside from a constant, the probability density. The form of the particle current is also familiar from non-relativistic quantum theory. The globally conserved quantities are

$$\begin{aligned} N &= \int d^3x n = 2\rho\mu \int d^3x \psi^*\psi \\ \mathbf{P} &= \int d^3x \mathbf{p} = i\rho\mu \int d^3x (\psi^*\nabla\psi - \psi\nabla\psi^*) = 2\rho\mu \int d^3x \psi^*i\nabla\psi \\ E &= \int d^3x \epsilon = \rho c_s^2 \int d^3x \nabla\psi^* \cdot \nabla\psi = 2\rho\mu \int d^3x \psi^* \frac{-c_s^2\nabla^2}{2\mu} \psi \end{aligned} \quad (16.130)$$

where in the last equation we have neglected the contribution to the energy density from the interaction. If ψ were a wave-function, then second line above would be the expectation value of the momentum operator $\mathbf{p} \propto i\nabla$ in the state ψ . The last line, similarly, is proportional to the expectation value of the Hamiltonian operator $H = p^2/2m$. The low-energy limit of the free-field scalar-elasticity equation of motion is exactly equivalent to the Schroedinger equation for a free-particle. For the interacting field, the equation of motion is formally similar to the Schroedinger equation for a particle in a potential $V = 6\lambda\psi^*\psi/\mu$. However, this analogy should not be taken too far, since this potential is both time and position dependent, which properties generally destroy energy and momentum conservation.

16.11 Ideal Fluid Limit of Wave Mechanics

As an example of the power of the energy and momentum conservation laws, consider a 3-dimensional field $\phi(\mathbf{x}, t)$ of the type we have been discussing. Let us assume that the field has a self-interaction, let's say a $\lambda\phi^4$ interaction term, but which is sufficiently weak that the waves can be locally approximated as a sum of free-field traveling waves. Now imagine we inject some wave energy into this system in some arbitrary and inhomogeneous manner. This is analogous to throwing a brick into a swimming pool; we may initially have an organized disturbance, but after a little time interactions between the waves and the walls of the pool will randomly distribute the energy among the various modes, and the result is a *Gaussian random wave field*. Such a field can be realized by summing waves with random complex amplitudes $\tilde{\phi}(\mathbf{k})$ (or random real amplitudes and random phases) and is fully characterized by the *power spectrum* $P_\phi(\mathbf{k}) = \langle |\tilde{\phi}_\mathbf{k}|^2 \rangle$. Here the interactions between wave modes will similarly effectively randomize the phases and amplitudes of the waves, and the system will relax to a locally homogeneous state (the situation here differs slightly from the swimming pool example in that there the interactions with the wall do not respect momentum conservation, so the power spectrum for waves in the pool becomes isotropic).

Now while wave interactions can efficiently locally homogenize the wave energy spectrum, large scale inhomogeneities in the energy and momentum distribution will take a very long time to be erased. This is very similar to the situation in gas dynamics, where collisions can rapidly render the velocity distribution locally Maxwellian, but where large scale inhomogeneities — sound waves for instance — take a very long time to damp out. What we will develop here is a system of equations which describe the spatio-temporal evolution of the wave energy and momentum. These turn out to be identical to the ideal fluid equations for a collisional gas of relativistic particles.

16.11.1 Local Average Stress-Energy Tensor

We have already obtained the energy and momentum densities $\epsilon = \dot{\phi} \partial \mathcal{L} / \partial \dot{\phi} - \mathcal{L}$ and $\mathbf{p} = -\dot{\phi} \partial \mathcal{L} / \partial \nabla \phi$, which for the system considered here become

$$\epsilon = \frac{\rho}{2} (\dot{\phi}^2 + (\nabla \phi)^2 + \mu^2 \phi^2) \quad (16.131)$$

and

$$\mathbf{p} = -\rho \dot{\phi} \nabla \phi. \quad (16.132)$$

These are fluctuating quantities, with coherence length- and time-scales $\sim 1/k$ and $1/\omega_{\mathbf{k}}$ respectively. Let's average the energy and momentum densities and flux densities in some region of space-time of size $L \gg 1/k$ and duration $T \gg 1/\omega_{\mathbf{k}}$. The average energy density is

$$\langle \epsilon \rangle = \frac{\rho}{2} \left(\langle \dot{\phi}^2 \rangle + c_s^2 \langle (\nabla \phi)^2 \rangle + \mu^2 \langle \phi^2 \rangle \right). \quad (16.133)$$

Each of these expectation values can be written as an integral over wave-number involving the power spectrum. For example, $\langle \dot{\phi}^2 \rangle = (2\pi)^{-3} \int d^3 k \omega_{\mathbf{k}}^2 P_{\phi}(\mathbf{k})$, and therefore

$$\langle \epsilon \rangle = \frac{\rho}{2} \int \frac{d^3 k}{(2\pi)^3} (\omega_{\mathbf{k}}^2 + c_s^2 k^2 + \mu^2) P_{\phi}(\mathbf{k}) \quad (16.134)$$

but the waves are assumed to obey the free-field dispersion relation to a very good approximation, so the term in parentheses is just twice $\omega_{\mathbf{k}}^2$ and the mean energy density is

$$\langle \epsilon \rangle = \rho \int \frac{d^3 k}{(2\pi)^3} \omega_{\mathbf{k}}^2 P_{\phi}(\mathbf{k}). \quad (16.135)$$

Similarly, the mean momentum density is

$$\langle \mathbf{p} \rangle = \left\langle -\frac{\partial \mathcal{L}}{\partial \dot{\phi}} \nabla \phi \right\rangle = -\rho \langle \dot{\phi} \nabla \phi \rangle = \rho \int \frac{d^3 k}{(2\pi)^3} \omega_{\mathbf{k}} \mathbf{k} P_{\phi}(\mathbf{k}). \quad (16.136)$$

The energy flux is $F_i = \dot{\phi} \partial \mathcal{L} / \partial \phi_{,i}$, so the mean energy flux is

$$\langle \mathbf{F} \rangle = \left\langle \dot{\phi} \frac{\partial \mathcal{L}}{\partial \nabla \phi} \right\rangle = -\rho c_s^2 \langle \dot{\phi} \nabla \phi \rangle = \rho c_s^2 \int \frac{d^3 k}{(2\pi)^3} \omega_{\mathbf{k}} \mathbf{k} P_{\phi}(\mathbf{k}). \quad (16.137)$$

Finally, the average momentum flux tensor is

$$\langle w_{ij} \rangle = \langle \delta_{ij} \mathcal{L} - \frac{\partial \mathcal{L}}{\partial \phi_{,j}} \phi_{,i} \rangle = \frac{\delta_{ij}}{2} \langle \dot{\phi}^2 - c_s^2 (\nabla \phi)^2 - \mu^2 \phi^2 \rangle + \rho c_s^2 \langle \phi_{,j} \phi_{,i} \rangle. \quad (16.138)$$

The first term vanishes by virtue of the dispersion relation, so

$$\langle w_{ij} \rangle = \rho c_s^2 \langle \phi_{,j} \phi_{,i} \rangle = c_s^2 \rho \int \frac{d^3 k}{(2\pi)^3} k_i k_j P_{\phi}(\mathbf{k}). \quad (16.139)$$

The contribution to the energy density from wave modes in $d^3 k$ is $d^3 \epsilon = \rho d^3 k \omega_{\mathbf{k}}^2 P_{\phi}(\mathbf{k}) / (2\pi)^3$, so the quantities here are energy weighted averages. The energy flux density, for example, is the average of $c_s^2 \mathbf{k} / \omega_{\mathbf{k}}$. But $c_s^2 \mathbf{k} / \omega_{\mathbf{k}} = \mathbf{v}_g(\mathbf{k})$, the group velocity for waves with wave-number \mathbf{k} , so $\langle \mathbf{F} \rangle$ is $\langle \epsilon \rangle$ times the energy weighted average group velocity. Similarly, the average momentum flux tensor $\langle w_{ij} \rangle$ is $\langle \epsilon \rangle$ times the energy weighted group velocity dispersion tensor.

We can combine the various factors in the 4×4 matrix

$$T^{\mu\nu} = \begin{bmatrix} \epsilon & \vdots & c_s \mathbf{p} \\ \dots & & \dots \\ c_s \mathbf{p} & \vdots & w_{ij} \end{bmatrix} = c_s^2 \int \frac{d^3 k}{(2\pi)^3} P_{\phi}(\mathbf{k}) k^{\mu} k^{\nu} \quad (16.140)$$

where the four-momentum (or four-wave-number) is $\vec{k} = (\omega_{\mathbf{k}}/c_s, \mathbf{k})$.

Now recall that solutions of the wave equations satisfy a covariance under Lorentz-like boost transformations. If we have a solution $\phi(\vec{x}) = \phi(\mathbf{x}, t)$ then we can generate a 3-dimensional family of solutions $\phi'(\vec{x}) = \phi(\Lambda\vec{x})$ where Λ is the transformation matrix for a boost, parameterized by the boost velocity $\mathbf{v} = \beta c_s$. The transformation for a boost along the x -axis is a shearing in the $x - t$ plane. Now $d^3k P_\phi(\mathbf{k})/(2\pi)^3$ is the contribution to the variance $\langle \phi^2 \rangle$ from the region d^3k about \mathbf{k} . Since the variance is invariant under boosts (since ϕ is invariant) so also is $d^3k P_\phi(\mathbf{k})$. This means that the matrix $T^{\mu\nu}$ is in fact a 4-tensor; i.e. its components transform like $k^\mu k^\nu$ under a boost. It is called the *stress energy tensor*.

This tensor takes a particularly simple form if we apply a boost such that the momentum flux density for the transformed field vanishes. If there is no net momentum flux, the effect of collisions is to produce a spherically symmetric, or isotropic, power spectrum $P_\phi(\mathbf{k}) = P_\phi(k)$, and the stress-energy tensor is therefore

$$T^{\mu\nu} = \begin{bmatrix} \epsilon_0 & & & \\ & P & & \\ & & P & \\ & & & P \end{bmatrix} \quad (16.141)$$

where ϵ_0 is the energy density of the transformed field, and we have used

$$\langle w_{ij} \rangle = c_s^2 \int \frac{d^3k}{(2\pi)^3} k^i k^j P_\phi(k) = \epsilon_0 \frac{\int \frac{d^3k}{(2\pi)^3} v_g^i v_g^j \omega_{\mathbf{k}}^2 P_\phi(k)}{c_s^2 \int \frac{d^3k}{(2\pi)^3} \omega_{\mathbf{k}}^2 P_\phi(k)} = \epsilon_0 \langle |\mathbf{v}_g|^2 \rangle \delta_{ij} / 3c_s^2 \quad (16.142)$$

to define the *pressure*

$$P \equiv \epsilon_0 \langle |\mathbf{v}_g|^2 \rangle / 3c_s^2. \quad (16.143)$$

The pressure is therefore essentially the energy weighted mean square group velocity.

The stress-energy tensor for the actual field can then be obtained by applying the boost transformation matrix to (16.141). For a boost along the x -axis, for instance,

$$T^{\mu\nu} = \begin{bmatrix} \gamma^2(\epsilon_0 + \beta^2 P) & \beta\gamma^2(\epsilon_0 + P) & & \\ \beta\gamma^2(\epsilon_0 + P) & \gamma^2(\beta^2 \epsilon_0 + P) & & \\ & & P & \\ & & & P \end{bmatrix} \quad (16.144)$$

and we can then read off the components of the energy density, momentum density, momentum flux etc. The actual energy density is $\epsilon = \gamma^2(\epsilon_0 + \beta^2 P)$ and the actual momentum density is $c_s \mathbf{p} = \beta\gamma^2(\epsilon_0 + P)$. The latter can also be written as $c_s \mathbf{p} = \beta(\epsilon + P)$. Thus given some actual $T^{\mu\nu}$ we can determine the parameters β , ϵ_0 and P as follows: 1) Find the 3-rotation matrix that aligns the momentum \mathbf{p} with the x -axis. 2) Read off the pressure P from either T^{22} or T^{33} . 3) Compute the dimensionless velocity $\beta = c_s \mathbf{p} / (\epsilon + P)$. This also provides $\gamma = 1/\sqrt{1 + \beta^2}$. 4) Solve $\epsilon = \gamma^2(\epsilon_0 + \beta^2 P)$ for ϵ_0 .

An alternatively, but particularly useful form for the stress-energy tensor can be obtained if we define the 4-velocity $\vec{u} = (\gamma c_s, \gamma \mathbf{v})$. This is exactly analogous to the 4-velocity in special relativity, but with the asymptotic sound speed c_s in place of the speed of light, and has all the familiar properties: e.g. $\vec{u} \cdot \vec{u} = u^\mu u_\mu = -c_s^2$, where $u_\mu = \eta_{\mu\nu} u^\nu$ with $\eta_{\mu\nu}$ the usual Minkowski metric. In terms of \vec{u} , we can write

$$T^{\mu\nu} = (\epsilon_0 + P) u^\mu u^\nu / c_s^2 + \eta^{\mu\nu} P. \quad (16.145)$$

It is easy to check that this agrees with (16.144) for the particular case $\mathbf{v} = (\beta c_s, 0, 0)$. However, ϵ_0 and P are the energy density and pressure as calculated under the specific boost transformation that annuls the momentum density, and they therefore transform as scalars under boosts, so consequently (16.145) is a 4-tensor equation and is therefore valid for arbitrary boost velocity \mathbf{u} .

To summarize, we first expressed the average stress energy tensor in terms of the power spectrum of the waves (16.140). We then obtained an alternative, but equivalent, expression (16.145) in terms of the five parameters ϵ_0 , P and the three components of the velocity boost \mathbf{u} needed to transform from the zero momentum field to the actual field.

16.11.2 Evolution Equations

So far we have imagined this sea of waves to be perfectly homogeneous. Now imagine that the sea has been rendered *locally* statistically homogeneous by the wave interactions, but that there is some persistent large scale inhomogeneity. We can compute the evolution of such macroscopic inhomogeneities by taking the local spatial average of the conservation laws $\dot{\epsilon} = -\nabla \cdot \mathbf{F}$ and $\dot{\mathbf{p}} = -\nabla \cdot \mathbf{w}$. These can be succinctly combined as the time and space components of the 4-vector equation

$$T^{\mu\nu}{}_{,\nu} = 0. \quad (16.146)$$

While succinct, this rather obscures the physical content. To reveal this, note that the time component of this set of equations (obtained by setting $\mu = 0$) is, from (16.145)

$$0 = c_s^2 T^{0\nu}{}_{,\nu} = \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^0 u^\nu + c_s^2 P \eta^{0\nu}] = \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^0 u^\nu] - c_s^2 \frac{\partial P}{\partial t} \quad (16.147)$$

where we have used $\eta^{\mu\nu} = \text{diag}(-1, 1, 1, 1)$ and $\vec{x} = (c_s t, \mathbf{x})$. The space components are similarly obtained by setting $\mu = i$:

$$\begin{aligned} 0 &= c_s^2 T^{i\nu}{}_{,\nu} = \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^i u^\nu + c_s^2 P \eta^{i\nu}] \\ &= \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^0 u^\nu \beta^i] + c_s^2 \frac{\partial P}{\partial x^i} \\ &= \beta^i \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^0 u^\nu] + (\epsilon_0 + P)u^0 u^\nu \frac{\partial \beta^i}{\partial x^\nu} + c_s^2 \frac{\partial P}{\partial x^i} \end{aligned} \quad (16.148)$$

where, to obtain the second line we have used $u^i = u^0 \beta^i$ and $\eta^{i\nu} = \delta^{i\nu}$, and in the last step we have simply used the rule for differentiating a product. Comparing with (16.147) we see that the first term on the right hand side can be written as $\beta^i c_s \partial P / \partial t = v^i \partial P / \partial t$ and (16.148) becomes

$$(\epsilon_0 + P) \gamma u^\nu \frac{\partial v^i}{\partial x^\nu} + c_s^2 \frac{\partial P}{\partial x^i} + v^i \frac{\partial P}{\partial t}. \quad (16.149)$$

But $u^\nu \partial / \partial x^\nu = \gamma(\partial / \partial t + (\mathbf{v} \cdot \nabla))$ and therefore this becomes the vector equation

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{c_s^2}{\gamma^2(\epsilon_0 + P)} \left[\nabla P + \mathbf{v} \frac{\partial P}{\partial t} \right] \quad (16.150)$$

where we recognize, on the left hand side, the convective derivative of $\dot{\mathbf{v}}$, the local rate of change of \mathbf{v} as seen by an observer moving at velocity \mathbf{v} (i.e. an observer in who's frame the momentum density vanishes). Equation (16.150) is the relativistic form of the Euler equation.

A more useful form for the fourth component of the conservation laws is obtained if we dot $T^{\mu\nu}{}_{,\nu}$ with u^μ . Multiplying by c_s^2 , this gives

$$\begin{aligned} 0 &= u_\mu T^{\mu\nu}{}_{,\nu} = \frac{1}{c_s^2} u_\mu \frac{\partial}{\partial x^\nu} [(\epsilon_0 + P)u^\mu u^\nu] + u^\nu \frac{\partial P}{\partial x^\nu} \\ &= \frac{u_\mu u^\mu}{c_s^2} \frac{\partial(\epsilon_0 + P)u^\nu}{\partial x^\nu} + (\epsilon_0 + P) \frac{u^\nu u_\mu}{c_s^2} \frac{\partial u^\mu}{\partial x^\nu} + u^\nu \frac{\partial P}{\partial x^\nu} \\ &= -\frac{\partial(\epsilon_0 + P)u^\nu}{\partial x^\nu} + 0 + u^\nu \frac{\partial P}{\partial x^\nu} \\ &= -(\epsilon_0 + P) \frac{\partial u^\nu}{\partial x^\nu} - u^\nu \frac{\partial \epsilon_0}{\partial x^\nu} \end{aligned} \quad (16.151)$$

where we have used $u_\mu \partial u^\mu / \partial x^\nu = (1/2) \partial(\vec{u} \cdot \vec{u}) / \partial x^\nu = 0$. This then provides an expression for the convective derivative of the energy density:

$$\frac{\partial \epsilon_0}{\partial t} + (\mathbf{v} \cdot \nabla) \epsilon_0 = -\frac{\epsilon_0 + P}{\gamma} \left[\frac{\partial \gamma}{\partial t} + \nabla \cdot (\gamma \mathbf{v}) \right]. \quad (16.152)$$

Equations (16.152) and (16.150) take a particularly simple form in the vicinity of a point where the momentum density, and therefore also \mathbf{v} , vanishes, since we can then take $\gamma = 1$ to obtain

$$\frac{d\epsilon_0}{dt} = -(\epsilon_0 + P)\nabla \cdot \mathbf{v} \quad (16.153)$$

and

$$\frac{d\mathbf{v}}{dt} = -\frac{\nabla P}{\epsilon_0 + P}. \quad (16.154)$$

These are in fact identical to the relativistically correct energy equation and Euler equation for a collisional gas. They provide four equations for five unknowns ϵ_0 , P and \mathbf{u} . To close this system of equations we need an *equation of state*; a rule giving the pressure as a function of ϵ_0 . In the following sections we find this relation for the two limiting cases of high spatial frequency waves with $k \gg k_*$, corresponding to the highly relativistic gas, and the opposite case $k \ll k_*$, which corresponds to the non-relativistic gas.

High Frequency Waves

It is very easy to find the equation of state in the limit that the sound waves have very high spatial frequency $|\mathbf{k}| \gg \mu/c_s$. In this case the field is effectively massless and the group velocity is $|\mathbf{v}_g| = c_s$ for essentially all \mathbf{k} and therefore $\langle |\mathbf{v}_g|^2 \rangle = c_s^2$ and therefore $P = \epsilon_0/3$. This is the same as the equation of state for a gas in which the particles are highly relativistic.

The energy equation, in the vicinity of a point where $\mathbf{v} = 0$, in this limit says

$$\frac{d\epsilon_0}{dt} = -\frac{4}{3}\epsilon_0\nabla \cdot \mathbf{v}. \quad (16.155)$$

Now $\nabla \cdot \mathbf{v} = V^{-1}dV/dt$, with V the volume, so this tells us that the energy, and hence also the pressure, changes inversely as the 4/3 power of the volume: $PV^{4/3} = \text{constant}$.

The general energy and Euler equations are then

$$\dot{\epsilon}_0 = -\frac{4}{3}\frac{\epsilon_0}{\gamma} \left[\frac{\partial \gamma}{\partial t} + \nabla \cdot (\gamma \mathbf{v}) \right] \quad (16.156)$$

and

$$\dot{\mathbf{v}} = -\frac{c_s}{4\gamma^2\epsilon_0} \left[\nabla \epsilon_0 + \mathbf{v} \frac{\partial \epsilon_0}{\partial t} \right]. \quad (16.157)$$

These provide a coupled, but closed, system of 4 equations for the four functions, or fields, $\epsilon_0(\mathbf{x}, t)$, $\mathbf{v}(\mathbf{x}, t)$ (note that $\gamma = 1/\sqrt{1 - v^2/c_s^2}$ is not an independent variable). If these fields, and their derivatives, are known on one time slice, then (16.156) and (16.157) allow one to advance the fields to the next time slice and in this wave determine the future evolution from arbitrary initial conditions.

We started with a field equation for $\phi(\mathbf{x}, t)$, and a rather complicated one at that, with a self-interacting field. We finished with a system of 4 deterministic equations for the scalar $\epsilon_0(\mathbf{x}, t)$ and the vector field $\mathbf{v}(\mathbf{x}, t)$. The significance of the latter is that \mathbf{v} is the boost transformation which locally annuls the momentum density, and ϵ_0 is the energy density of the transformed field. Once we have the solution, we can then determine the stress energy tensor $T^{\mu\nu}(\mathbf{x}, t)$ using (16.145).

These equations, like the equations of ideal fluid dynamics, are non-linear. For small perturbations about a uniform energy density state, we can linearize and let $\epsilon_0 = \bar{\epsilon} + \epsilon_1$, and $\mathbf{v} = 0 + \mathbf{v}_1$, where quantities with subscript 1 are assumed to be small perturbations: $\epsilon_1 \ll \bar{\epsilon}$ and $\mathbf{v}_1 \ll c_s$. The latter means we can set $\gamma = 1$ and ignore $\mathbf{v} \cdot \nabla$ compared to $\partial/\partial t$ in the convective derivative. In this limit the energy and Euler equations become

$$\bar{\epsilon} \frac{\partial \mathbf{v}_1}{\partial t} = -\frac{c_s^2}{4} \nabla \epsilon_1 \quad (16.158)$$

and

$$\frac{\partial \epsilon_1}{\partial t} = -\frac{4}{3}\bar{\epsilon} \nabla \cdot \mathbf{v}_1 \quad (16.159)$$

or, on taking the divergence of (16.158) and the time derivative of (16.159) and eliminating the velocity,

$$\frac{\partial^2 \epsilon_1}{\partial t^2} - \frac{c_s^2}{3} \nabla^2 \epsilon_1 = 0. \quad (16.160)$$

This is a wave equation, with wave speed $c_s/\sqrt{3}$. These are very different from the waves of the underlying ϕ field; these are waves of wave energy density. This equation, for the evolution of self-interacting sound waves in our solid state lattice system is exactly analogous to the equation governing acoustic waves in a radiation dominated plasma.

Low-Frequency Waves

Now let's consider the opposite limit where $k \ll k_*$. As discussed in §16.10 we then have, in addition to energy and momentum, a fifth approximate conservation law. These five conservation laws were given in (16.128), along with the constituent relations (16.129). The latter can be simplified in the case that the interaction coupling constant is large enough that the interactions render the power-spectrum isotropic, but small enough that the interaction energy $\lambda\phi^4$ is negligible compared to the other contributions to the energy. One can then drop the terms in (16.129) which involve λ .

Proceeding much as before, we can assume that the interactions will render the wave-field as a locally homogeneous Gaussian random field, and we can compute quantities such as the energy density, particle density in terms of the power spectrum. One can then show that the conservation laws (16.128) reduce to the continuity, Euler and energy equations.

16.12 Discussion

We have shown here how one can treat continuous media using the Lagrangian formalism. We started with a simple 'solid state' system consisting of a lattice of coupled oscillators and determined the discrete set of normal modes, which are the discrete Fourier transform of the displacements ϕ_j . We then took the continuum limit, and showed that the action is the 2-dimensional space-time integral the *Lagrangian density* $\mathcal{L}(\dot{\phi}, \nabla\phi, \phi)$, which is a quadratic function of the field and its derivatives. The condition $\delta S = 0$ generated the equation of motion of the system in the form of a 2nd order differential equation in space and time. The normal modes are again Fourier modes (continuous Fourier modes now). The generalization to 2 or 3 spatial dimensions is straightforward.

We then showed that the symmetry of the Lagrangian density under spatial translations gave rise to conservation of *wave momentum*. Applying this to non-interacting wave packets we found that energy and momentum are related by $\mathbf{P}/E = \mathbf{k}/\omega_{\mathbf{k}}$. We also investigated multiple, possibly coupled, fields. We saw that a two component field can have additionally conservation laws which look suspiciously like conservation of electric charge, or some other conserved quantity, and we showed that low-energy (i.e. non-relativistic) waves display a further conservation analogous to particle conservation. Finally, we explored in some detail, the behaviour of interacting fields with interactions sufficiently weak that they do not substantially modify the energy of the system, but sufficiently strong that they frustrate the propagation of wave packets over large distances. We found that the energy and momentum density in such self-interacting wave systems obey a closed system of equations which are precisely like those obeyed by relativistic collisional gas of pointlike particles, in the ideal fluid limit (see e.g. Weinberg's book). What we have seen here is a *classical wave-particle correspondence principle*; we can think of an ideal fluid as a gas of classical particles which collide with each other or as a field of self-interacting waves. The two models for the 'underlying real system' give identical results for the laws governing the macroscopic propagation of energy and momentum.

The underlying discrete lattice model was chosen so that the field equation, dispersion relation, wave-packet energy momentum relation etc. mimic the properties of the relativistic massive scalar field (which we will introduce in chapter 18). The mechanical system introduced here, and explored quantum mechanically in the following chapter, therefore provides a useful, concrete and conceptually less challenging analog of the more abstract relativistic scalar field. In employing this analogy one

should be careful how to interpret the field. In the mechanical model, the field happens to be a spatial displacement. However, the direction of the field was arbitrary; In the 1-dimensional model we drew the beads oscillating vertically, but they could equally well have been drawn oscillating horizontally and either transverse or parallel to the lattice. In any of these cases the Lagrangian density is the same. Similarly, in 2- or 3-dimensions, the direction of motion was quite arbitrary. This is quite different from waves in real elastic solids, where the displacement is a vector. The model we have constructed is for a *scalar* disturbance; the displacement is best thought of as lying in an abstract 1-dimensional internal space. What we have called the microscopic momentum — the sum of the conjugate momenta — ‘lives’ in this abstract space. The wave-momentum, in contrast, is a true vector quantity in the real space, and which transforms in the usual way under spatial rotations, for instance.

The invariance of such classical field theories under what are formally identical to Lorentz transformations is rather interesting, and encourages a rather non-standard view of relativistic invariance. As already emphasised, the system here is *not* truly relativistically covariant; if we imagine a two-dimensional system, then if we observe the propagation of waves on this surface using photons (or other particles with a propagation speed much greater than c_s , the asymptotic velocity for high frequency elasticity waves) then we can readily infer the frame of rest of the underlying lattice; it is that frame for which the dispersion relation is independent of wave-momentum direction. Using photons, one can, of you like, measure the *aether drift* for these elasticity waves. However, the wave-systems considered here allow quite rich behaviour, encompassing analogs of relativistic and non-relativistic particles, and interactions between waves allow quite complex solutions to develop. What if we imagine some being, or perhaps something like a cellular automaton, which lives in this 2-dimensional ‘plani-verse’ and is constructed out of these fields (just as we imagine ourselves to be, in reality, complex solutions of a rather simple set of fundamental field equations) — can such an entity measure the aether drift? The answer is that there is no observation that this entity can make which can reveal the underlying lattice rest-frame, *provided all the fields in the system have identical asymptotic sound speed c_s* . The fields may have quite different ‘base-spring’ strengths (i.e. different masses) and the base-springs may be non-linear or may have pointlike interactions with other fields, but provided the bead-mass and connecting-spring constants give the same asymptotic sound speed, the system is, internally, fully covariant and any aether drift measurements are doomed to failure. If on the other hand, the connecting spring constants and bead masses give different c_s values for the different fields, then the Lorentz-like transformations which preserve the field equations depend on the particular field; the transformation under which one field remains a solution, does not preserve the field equations for other fields.

Thus, it would seem to be quite consistent to believe that the world we inhabit is ‘in reality’ a Galilean system, with absolute time etc., but that all of the fields in our universe happen to have the same asymptotic propagation speed, so the world ‘appears’ to be Lorentz invariant. If the world is just fields, and if all of the fields have identical c_s , then this is meta-physics; if we cannot detect the aether, then we should not introduce it in our physical description of the Universe. However, if we were ever to be able to demonstrate the existence of say ‘tachyonic’ behaviour; e.g. apparently causal correlations between the field values at points with space-like separations, then this could be readily be incorporated in classical field theory, simply by introducing fields with different c_s values.

Finally, there is an important inadequacy in the interacting field theories we have developed. We were only able to establish the wave-particle correspondence principle in the limits of highly relativistic or highly non-relativistic waves. In the particle description one can treat collisional gases at all energies, if we introduce the additional requirement that the gas be locally in thermal equilibrium. This extra ingredient provides the distribution function for particle momenta, and provides a general equation of state. Only in the limits we have considered are the evolutionary equations for the wave energy independent of the details of the momentum distribution. If we try to introduce thermal equilibrium in a classical wave system we immediately run into problems; the equilibrium state should have equal energy per oscillation mode, and this is counter to observations. The resolution, of course, is that the world is not classical, but quantum mechanical, and the energy states of the modes are not continuous but are discretized in units of $\hbar\omega$. In the next chapter we will develop the proper quantum mechanical description of wave systems.

Chapter 17

Quantum Fields

Hamiltonian dynamics provides a direct path from the classical to the quantum description of a system. The *canonical quantization procedure* consists of replacing the position and momentum variables q and p with non-commuting operators which act on quantum mechanical states. For the discrete lattice model these operators are the displacements ϕ_j and momenta $p_j = \partial L / \partial \dot{\phi}_j = M \dot{\phi}_j$.

17.1 The Simple Harmonic Oscillator

Consider a single simple harmonic oscillator, with Hamiltonian

$$H = p^2/2m + kx^2/2. \quad (17.1)$$

One approach to this is to solve Schroedinger's equation $E\psi = H\psi$ (with the usual substitutions $E \rightarrow i\hbar\partial/\partial t$ and $p \rightarrow -i\hbar\partial/\partial x$) to obtain the wave-function ψ . This gives the well known result that the energy levels are quantized with

$$E_n = (n + 1/2)\hbar\omega \quad (17.2)$$

with $\omega = \sqrt{k/m}$. There is, however, a different approach using creation and destruction operators, which proves to be of great value in field theory.

These operators are defined to be

$$\begin{aligned} a &= \frac{1}{\sqrt{2\hbar m\omega}}(m\omega x + ip) \\ a^\dagger &= \frac{1}{\sqrt{2\hbar m\omega}}(m\omega x - ip) \end{aligned} \quad (17.3)$$

If x and p were classical variables then their product aa^\dagger would be $H/\hbar\omega$. But x and p are really non-commuting operators obeying the commutation relation

$$[x, p] \equiv xp - px = i\hbar. \quad (17.4)$$

Now the two possible products of a and a^\dagger are

$$\begin{aligned} aa^\dagger &= \frac{1}{2m\hbar\omega}(m^2\omega^2x^2 - im\omega(xp - px) + p^2) \\ a^\dagger a &= \frac{1}{2m\hbar\omega}(m^2\omega^2x^2 + im\omega(xp - px) + p^2) \end{aligned} \quad (17.5)$$

from which follow two important results

- Subtracting these and using (17.4) gives the commutator of a and a^\dagger

$$[a, a^\dagger] = 1. \quad (17.6)$$

- Adding these gets rid of the cross terms and we find that the Hamiltonian is proportional to the *symmetrized product* of a and a^\dagger :

$$H = \frac{\hbar\omega}{2}(aa^\dagger + a^\dagger a) \quad (17.7)$$

or as $H = \hbar\omega\{a, a^\dagger\}/2$ where the ‘anti-commutator’ is defined as $\{a, a^\dagger\} \equiv aa^\dagger + a^\dagger a$.

Why are these called creation and destruction operators (sometimes ‘ladder operators’)? To see this let’s apply the operator H to the state $|aE\rangle = a|E\rangle$ where $|E\rangle$ is the energy eigenstate with energy E (i.e. $H|E\rangle = E|E\rangle$)

$$\begin{aligned} H|aE\rangle &= Ha|E\rangle = \frac{\hbar\omega}{2}(aa^\dagger a + a^\dagger aa)|E\rangle \\ &= \frac{\hbar\omega}{2}(a(aa^\dagger - 1) + (aa^\dagger - 1)a)|E\rangle \\ &= a(H - \hbar\omega)|E\rangle = (E - \hbar\omega)a|E\rangle = (E - \hbar\omega)|aE\rangle \end{aligned} \quad (17.8)$$

where we have used the commutation relation (17.6) to replace $a^\dagger a$ by $aa^\dagger - 1$. Evidently the state $|aE\rangle$ is also an energy eigenstate, but with energy level one quantum lower than the state $|E\rangle$, so effectively the operator a has destroyed one quantum of energy. Identical reasoning shows that $H|a^\dagger E\rangle = (E + \hbar\omega)|a^\dagger E\rangle$ so a^\dagger has the effect of increasing the energy by one quantum.

These results imply that the energy eigenstates are also eigenstates of the operator $N = a^\dagger a$ since this first lowers the energy but then raises it again. What are the eigenvalues n of this operator? The eigenvalue equation is $N|n\rangle = n|n\rangle$. Now using (17.6) we find that

$$Na|n\rangle = a^\dagger aa|n\rangle = (aa^\dagger a - a)|n\rangle = a(N - 1)|n\rangle = (n - 1)a|n\rangle \quad (17.9)$$

so the state $a|n\rangle$ has eigenvalue one lower than the state $|n\rangle$ and similarly $N|a^\dagger n\rangle = (n + 1)|a^\dagger n\rangle$. Now a applied to the lowest energy state $|n_{\min}\rangle$ must vanish, as must $a^\dagger a|n_{\min}\rangle$ and therefore the eigenvalue of the lowest energy state $\langle n_{\min}|a^\dagger a|n_{\min}\rangle$ must vanish $n_{\min} = 0$, and the eigenvalues are just the integers $n = 0, 1, 2, \dots$. The operator N is called the *number operator*.

What about the normalization of the states $|an\rangle$? These are not normalized, since the bra corresponding to the ket $|an\rangle$ is $\langle an| = \langle n|a^\dagger$ and so $\langle an|an\rangle = \langle n|a^\dagger a|n\rangle = n$. If we require that these eigenstates be normalized so $\langle n|n\rangle = 1$ for all n , then $\langle an|an\rangle = n = n\langle n - 1|n - 1\rangle$ and similarly $\langle a^\dagger n|a^\dagger n\rangle = n + 1 = (n + 1)\langle n + 1|n + 1\rangle$ and therefore

$$\begin{aligned} a|n\rangle &= n^{1/2}|n - 1\rangle \\ a^\dagger|n\rangle &= (n + 1)^{1/2}|n + 1\rangle. \end{aligned} \quad (17.10)$$

We can now work out the energy eigenvalues E_n . Writing out all the steps in gory detail, these are

$$\begin{aligned} E_n &= \langle n|H|n\rangle = \frac{\hbar\omega}{2}\langle n|aa^\dagger + a^\dagger a|n\rangle \\ &= \frac{\hbar\omega}{2}[\sqrt{n+1}\langle n|a|n+1\rangle + \sqrt{n}\langle n|a^\dagger|n-1\rangle] \\ &= \frac{\hbar\omega}{2}[(n+1)\langle n|n\rangle + n\langle n|n\rangle] \\ &= (n + 1/2)\hbar\omega. \end{aligned} \quad (17.11)$$

Finally, note that one can express the physical operators p , x in terms of the creation and destruction operators as

$$\begin{aligned} x &= \sqrt{\hbar/2\omega m}(a^\dagger + a) \\ p &= i\sqrt{\hbar\omega m/2}(a^\dagger - a) \end{aligned} \quad (17.12)$$

These expressions allow one to compute expectation values such as the mean square displacement $\langle n|x^2|n\rangle$ in some state $|n\rangle$ using operator algebra, without requiring any definite prescription for the wave-function of the state. More important for us, physical operators appearing in the *interaction Hamiltonian* can be expressed in terms of the creation and destruction operators, which is very useful if we want to compute the rate at which interactions inject energy or scatter excitations.

17.2 The Interaction Picture

Quantum mechanics was initially developed in two quite independent ways by Schroedinger and by Heisenberg. These two approaches give identical results for the expectation values of physical operators $\langle \psi | Y | \psi \rangle$, but this still leaves considerable freedom in the definition of the states and operators.

In the *Schroedinger picture* the operators (such as x , p for the simple harmonic oscillator) are considered fixed in time while the state $|\psi_S\rangle$ with wave function $\psi(x, t)$ evolves in time according to the time dependent Schroedinger equation

$$i\hbar \frac{d}{dt} |\psi_S\rangle = H |\psi_S\rangle \quad (17.13)$$

the formal solution to which is

$$|\psi_S(t)\rangle = e^{-iH(t-t_0)/\hbar} |\psi_S(t_0)\rangle. \quad (17.14)$$

In the *Heisenberg picture* the states $|\psi\rangle$ are considered fixed in time, so we can take $|\psi_H\rangle = |\psi_S(t_0)\rangle$ say, and the equality $\langle \psi_S | Y_S | \psi_S \rangle = \langle \psi_H | Y_H | \psi_H \rangle$ implies that the operators in the two pictures are related by

$$Y_H = e^{iH(t-t_0)/\hbar} Y_S e^{-iH(t-t_0)/\hbar} \quad (17.15)$$

and differentiating this with respect to t (with Y_S constant) gives the ‘equation of motion’ for operators in the Heisenberg picture

$$i\hbar \frac{dY_H}{dt} = [Y_H, H]. \quad (17.16)$$

In what follows we shall be interested in systems in which the total Hamiltonian is the sum of the free field Hamiltonian H_0 , whose eigenstates have fixed occupation numbers, and an ‘interaction term’:

$$H = H_0 + H_{\text{int}}. \quad (17.17)$$

In this context it proves useful to work in the *interaction picture*, which is a hybrid of the Heisenberg and Schroedinger pictures in which the operators evolve in time as

$$Y_I = e^{iH_0(t-t_0)/\hbar} Y_S e^{-iH_0(t-t_0)/\hbar} \quad (17.18)$$

i.e. like the Heisenberg picture operators of the free field theory, and the equation of motion for the operators is

$$i\hbar \frac{dY_I}{dt} = [Y_I, H_0], \quad (17.19)$$

while the states evolve as

$$i\hbar \frac{d}{dt} |\psi_I\rangle = H_{\text{int}} |\psi_I\rangle. \quad (17.20)$$

For example, for the simple harmonic oscillator $H_0 = \hbar\omega\{aa^\dagger\}/2$, and, using the commutation relation (17.6), the equations of motion for a , a^\dagger become

$$\begin{aligned} da(t)/dt &= -i\omega a(t) \\ da^\dagger(t)/dt &= i\omega a^\dagger(t) \end{aligned} \quad (17.21)$$

the solutions to which are

$$\begin{aligned} a(t) &= ae^{-i\omega t} \\ a^\dagger(t) &= a^\dagger e^{i\omega t} \end{aligned} \quad (17.22)$$

where a , a^\dagger are the initial values of the operators. Note that the commutation relation (17.6) applies to $a(t)$, $a^\dagger(t)$ also: $[a(t), a^\dagger(t)] = [a, a^\dagger] = 1$.

In the interaction picture, the operators evolve rapidly, on timescale $\tau \sim \hbar/H_0$ while the states evolve slowly on timescale $\tau \sim \hbar/H_{\text{int}}$.

17.2.1 The S -Matrix Expansion

If we start at t_1 with the system in some state $|i\rangle$ then on integrating (17.20) we have for the state at some later time t

$$|\psi(t)\rangle = |i\rangle + \frac{1}{i\hbar} \int_{t_1}^t dt' H_{\text{int}}(t') |\psi(t')\rangle. \quad (17.23)$$

This is an *implicit* equation for $|\psi(t)\rangle$. We can obtain an explicit solution as a power law expansion in the strength of the interaction as follows. To zeroth order in the interaction nothing happens; the zeroth order solution is $|\psi^{(0)}(t)\rangle = |i\rangle$. We can obtain a better approximation by inserting $|\psi(t')\rangle = |i\rangle$ on the RHS of (17.23). We can then put this improved solution in the RHS and so on. Successive approximations are

$$\begin{aligned} |\psi^{(0)}(t)\rangle &= |i\rangle \\ |\psi^{(1)}(t)\rangle &= |i\rangle + \frac{1}{i\hbar} \int_{t_1}^t dt' H_{\text{int}}(t') |i\rangle \\ |\psi^{(2)}(t)\rangle &= |i\rangle + \frac{1}{i\hbar} \int_{t_1}^t dt' H_{\text{int}}(t') |i\rangle + \left(\frac{1}{i\hbar}\right)^2 \int_{t_1}^t dt' H_{\text{int}}(t') \int_{t_1}^{t'} dt'' H_{\text{int}}(t'') |i\rangle \\ &\dots \end{aligned} \quad (17.24)$$

The development of the various terms in this expansion can be expressed as a recursion relation

$$|\psi^{(n)}(t)\rangle = |i\rangle + \frac{1}{i\hbar} \int_{t_1}^t dt' H_{\text{int}}(t') |\psi^{(n-1)}(t')\rangle. \quad (17.25)$$

This expansion is the basis for essentially all calculations in perturbative field theory. It is known as the ‘ S -matrix expansion’. It gives the time evolution of the states as $|\psi(t)\rangle = U|i\rangle$ — where U is the time-evolution operator — from which we can compute the amplitude for the system to be in some state $|f\rangle$ as $\langle f|U|i\rangle$.

17.2.2 Example: A Forced Oscillator

As an illustration, let us add a term $H_{\text{int}} = F(t)x$ to the simple harmonic oscillator. This adds a term $L_{\text{int}} = -H_{\text{int}}$ to the Lagrangian, and the equation of motion is then $m\ddot{x} = -\partial L/\partial x = -kx + F(t)$, so $F(t)$ represents an external force which we treat classically. The operator corresponding to this classical interaction is

$$H_{\text{int}}(t) = \sqrt{\frac{\hbar}{2\omega m}} F(t) (a^\dagger(t) + a(t)) \quad (17.26)$$

and the state will evolve to

$$|\psi(t)\rangle = |\psi(t_0)\rangle - \frac{i}{\sqrt{2\omega m\hbar}} \int_{t_0}^t dt F(t) [e^{i\omega t} a^\dagger + e^{-i\omega t} a] |\psi(t_0)\rangle. \quad (17.27)$$

If we assume the state is initially in the vacuum, so $|\psi(t_0)\rangle = |0\rangle$ then the destruction operator has no effect, since $a|0\rangle = 0$, but $a^\dagger|0\rangle = |1\rangle$ so

$$|\psi(t)\rangle = |0\rangle - \frac{i \int dt F(t) e^{i\omega t}}{\sqrt{2m\omega\hbar}} |1\rangle \quad (17.28)$$

and the amplitude for the system to be in the excited state $|1\rangle$ at time t_f given that it was in the ground state $|0\rangle$ at time t_i is

$$\langle 1|\psi(t)\rangle = \frac{i}{\sqrt{2m\omega\hbar}} \int_{t_i}^{t_f} dt F(t) e^{i\omega t} \quad (17.29)$$

which is proportional to the transform of the force at the oscillator frequency.

Squaring the amplitude gives the probability for the transition as

$$p(0 \rightarrow 1) \equiv |\langle 1, t_f | 0, t_i \rangle|^2 = \frac{1}{2m\omega\hbar} \int_{t_i}^{t_f} dt \int_{t_i}^{t_f} dt' F(t) F(t') e^{i\omega(t-t')}. \quad (17.30)$$

For example, if the force is some random function of time, and acts for total time T , this is

$$p(0 \rightarrow 1) = \frac{T}{2m\omega\hbar} \int d\tau \xi_F(\tau) e^{i\omega\tau} \quad (17.31)$$

where

$$\xi_F(\tau) = \frac{1}{T} \int_0^T dt F(t) F(t + \tau) \quad (17.32)$$

is the two-point function for the force. Invoking the Wiener-Khinchin theorem gives

$$p(0 \rightarrow 1) = \frac{TP_F(\omega)}{2m\omega\hbar}, \quad (17.33)$$

so the probability that the system gets excited is proportional to the duration that the perturbation is switched on and to the power in the force at the frequency of the oscillator, an eminently reasonable result.

17.3 Free Fields

We can apply these concepts to the non-interacting field theory, since, as we have seen, this is a set of non-interacting simple harmonic oscillators. Thus we should be able to construct a pair of creation and destruction operators for each mode of the field. We will first do this in detail for the case of the 1-dimensional discrete lattice. The transition to a 3-dimensional continuous system is then straightforward.

17.3.1 Discrete 1-Dimensional Lattice Model

Let us assume a discrete lattice model consisting of a ring of N coupled oscillators. It turns out that one can write the displacement operator $\phi_j(t)$ as a sum over normal modes:

$$\phi_j(t) = \sum_k \sqrt{\frac{\hbar}{2MN\omega_k}} \left(a_k^\dagger e^{i(\omega_k t - 2\pi j k/N)} + a_k e^{-i(\omega_k t - 2\pi j k/N)} \right). \quad (17.34)$$

where a_k, a_k^\dagger are operators which respectively destroy and create excitations in the mode with wave number k .

To justify this we need to show first that these operators have the appropriate commutation relations, and second that they have the appropriate relationship to the Hamiltonian.

To do this we shall need the analogous expression for the velocities $\dot{\phi}_j$, which are just the time derivatives of the displacements ϕ_j given by (17.34):

$$\dot{\phi}_j(t) = i \sum_k \sqrt{\frac{\hbar\omega_k}{2MN}} \left(a_k^\dagger e^{i(\omega_k t - 2\pi j k/N)} - a_k e^{-i(\omega_k t - 2\pi j k/N)} \right). \quad (17.35)$$

Next, we relate a_k and a_k^\dagger to the (spatial) discrete transforms of ϕ_j and $\dot{\phi}_j$ defined as

$$\Phi_k(t) \equiv \sum_j \phi_j(t) e^{i2\pi j k/N} \quad \text{and} \quad \dot{\Phi}_k(t) \equiv \sum_j \dot{\phi}_j(t) e^{i2\pi j k/N}. \quad (17.36)$$

With $a_k(t) = a_k e^{-i\omega_k t}$ and $a_k^\dagger(t) = a_k^\dagger e^{i\omega_k t}$, the first of these is

$$\Phi_k(t) = \sum_j \sum_{k'} \sqrt{\frac{\hbar}{2MN\omega_{k'}}} \left(a_{k'}^\dagger(t) e^{i2\pi j(k-k')/N} + a_{k'}(t) e^{i2\pi j(k+k')/N} \right) \quad (17.37)$$

and similarly for $\dot{\Phi}_k$, but $\sum_j e^{i2\pi j(k\pm k')/N} = N\delta_{k\pm k'}$ and so

$$\begin{aligned} \Phi_k(t) &= \sqrt{\frac{\hbar N}{2M\omega_k}} (a_k^\dagger(t) + a_{-k}(t)) \\ \dot{\Phi}_k(t) &= i\omega_k \sqrt{\frac{\hbar N}{2M\omega_k}} (a_k^\dagger(t) - a_{-k}(t)). \end{aligned} \quad (17.38)$$

Solving for $a_k^\dagger(t)$ and $a_k(t)$, we have

$$\begin{aligned} a_k^\dagger(t) &= \sqrt{\frac{M\omega_k}{2\hbar N}} \left(\Phi_k(t) + \dot{\Phi}_k(t)/i\omega_k \right) \\ a_k(t) &= \sqrt{\frac{M\omega_k}{2\hbar N}} \left(\Phi_{-k}(t) - \dot{\Phi}_{-k}(t)/i\omega_k \right). \end{aligned} \quad (17.39)$$

We are now ready to compute the commutator $[a_k, a_{k'}^\dagger]$. To obtain this we first need the commutators for the $\Phi_k, \dot{\Phi}_k$ operators. Since Φ_k contains only displacement operators ϕ_j it commutes with itself and similarly for $\dot{\Phi}_k$ which contains only momentum operators. The only non-zero contributions to the commutator come from terms like $[\Phi_k, \dot{\Phi}_{-k'}]$, which is

$$[\Phi_k, \dot{\Phi}_{-k'}] = \left[\sum_l \phi_l e^{i2\pi kl/N}, \sum_m \dot{\phi}_m e^{-i2\pi k'm/N} \right] = \sum_l \sum_m e^{i2\pi(kl-k'm)/N} [\phi_l, \dot{\phi}_m] \quad (17.40)$$

but $\dot{\phi}_m = p_m/M$ so $[\phi_l, \dot{\phi}_m] = [\phi_l, p_m]/M = i\hbar\delta_{lm}/M$ (the δ_{lm} here expressing the fact that the operators for displacement and momentum at different sites on the lattice commute), and therefore

$$[\Phi_k, \dot{\Phi}_{-k'}] = \frac{i\hbar}{M} \sum_l e^{i2\pi l(k-k')/N} = \frac{i\hbar N}{M} \delta_{kk'}. \quad (17.41)$$

The commutator of the a, a^\dagger operators is, from (17.39)

$$[a_k(t), a_{k'}^\dagger(t)] = \frac{M}{2\hbar N} \sqrt{\omega_k \omega_{k'}} \left[\Phi_{-k}(t) - \frac{\dot{\Phi}_{-k}(t)}{i\omega_k}, \Phi_k(t) + \frac{\dot{\Phi}_k(t)}{i\omega_k} \right] \quad (17.42)$$

or, using (17.41),

$$[a_k(t), a_{k'}^\dagger(t)] = \delta_{kk'}. \quad (17.43)$$

For $k = k'$ this commutator is identical to the commutator $[a, a^\dagger]$ for a single oscillator and the commutator for different modes $k \neq k'$ vanishes.

We can also express the Hamiltonian in terms of a_k, a_k^\dagger as follows. If we go back to the expression for the Lagrangian (16.1), convert this to the Hamiltonian by changing the sign of the potential energy terms, and substitute $\phi_j \rightarrow N^{-1} \sum_k \Phi_k e^{-2\pi ijk/N}$ we find that the classical Hamiltonian becomes

$$H = \frac{M}{2N} \sum_k \dot{\Phi}_k \dot{\Phi}_k^* + \omega_k^2 \Phi_k \Phi_k^* = \frac{M}{2N} \sum_k \dot{\Phi}_k \dot{\Phi}_{-k} + \omega_k^2 \Phi_k \Phi_{-k} \quad (17.44)$$

and replacing the classical variables by operators we find

$$H = \sum_k \frac{\hbar\omega_k}{2} \{a_k, a_k^\dagger\} \quad (17.45)$$

which, as expected, is just a sum over the modes of the Hamiltonian for each oscillator mode taken separately.

From here one can argue exactly as before that the operator a_k has the effect of reducing the energy in mode k by one quantum and a_k^\dagger increases it by the same amount. Applying the creation operators to the vacuum generates the multi-particle eigenstates (15.1) etc.

This much we might have reasonably anticipated from the results for a single oscillator. The reason for going through the laborious analysis above is to obtain (17.34) and (17.35) for the displacement operator and its conjugate momentum in terms of the ladder operators. This means that if we can express the interaction Hamiltonian H_{int} in terms of the displacement, or the velocity, then we can also express it in terms of the $a_{\mathbf{k}}, a_{\mathbf{k}}^\dagger$ operators. We can then calculate transition rates with relative ease.

17.3.2 Continuous 3-Dimensional Field

The transition from the discrete model to a continuous field is mathematically straightforward. In one dimension we simply replace the position index j by $x = j\Delta x$ and the wave-number index k by the physical wave number $k_{\text{phys}} = 2\pi k/L$, so the factor $2\pi jk/N = k_{\text{phys}}x$ and similarly in higher dimensions $2\pi \mathbf{j} \cdot \mathbf{k}/N \rightarrow \mathbf{k}_{\text{phys}} \cdot \mathbf{x}$. The factor $MN = M_{\text{tot}}$ is just the total mass of all the beads and the (now continuous) displacement and velocity fields become

$$\begin{aligned}\phi(\mathbf{x}, t) &= \sum_{\mathbf{k}} \sqrt{\frac{\hbar}{2M_{\text{tot}}\omega_{\mathbf{k}}}} \left(a_{\mathbf{k}}^\dagger e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} + a_{\mathbf{k}} e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} \right) \\ \dot{\phi}(\mathbf{x}, t) &= i \sum_{\mathbf{k}} \sqrt{\frac{\hbar\omega_{\mathbf{k}}}{2M_{\text{tot}}}} \left(a_{\mathbf{k}}^\dagger e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} - a_{\mathbf{k}} e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} \right)\end{aligned}\tag{17.46}$$

Note that we are working here in a box of size L so that we still have a discrete set of modes. Since $L \rightarrow \infty$ we could replace the discrete sums here by continuous integrals, but the discrete form proves to be more convenient for our purposes.

The operators appearing here still satisfy the commutation relation

$$[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'}.\tag{17.47}$$

and the Hamiltonian can be written as

$$H = \frac{1}{2} \sum_{\mathbf{k}} \hbar\omega_{\mathbf{k}} (a_{\mathbf{k}} a_{\mathbf{k}}^\dagger + a_{\mathbf{k}}^\dagger a_{\mathbf{k}})\tag{17.48}$$

exactly as before. These operators can be used to create and destroy quanta, and to construct the energy eigenstates of the system systematically. The energy eigenstates of the free-field are simply the multi-particle eigenstates (15.1) which are the product of the states for the individual modes.

One might worry whether we are really allowed to carry over the commutation relations $[\phi_l, p_m] = i\hbar\delta_{lm}$ for the discrete position and momentum operators; here this says that the field and momentum operators at two different points commute, regardless of how close they are. Ultimately, the justification for this *physical* assumption is the extent to which the resulting field theory adequately describes nature.

17.4 Interactions

Consider a 2- or 3-dimensional discrete lattice model. Imagine the system is initially in some eigenstate of the free field Hamiltonian and we then ‘switch on’ some small additional interaction term (which will be specified presently) for some period of time. The interaction will cause the state to deviate from the fixed occupation number state, and we will have generally non-zero amplitudes, and hence probabilities, for the system to be found with different occupation numbers after the interaction.

We will illustrate this with several examples which we work through in detail. We will compute the scattering of phonons off an impurity in the lattice and then scattering of phonons with each other *via* non-harmonic (i.e. non-quadratic) behavior of the spring potential energy. Both of these

phenomena can be described using first order perturbation theory. We then explore the scattering of phonons *via* exchange of a virtual particle; an example of a second order perturbation theory calculation.

17.4.1 Scattering off an Impurity

What happens if we introduce an ‘impurity’ in the lattice and make one of the beads, the one at the origin of coordinates for concreteness, abnormally heavy with mass $M + \Delta M$. This will introduce a perturbation or interaction term to the Hamiltonian so we can write

$$H = H_0 + H_{\text{int}} \quad (17.49)$$

with H_0 the free-field Hamiltonian and with

$$H_{\text{int}}(t) = \frac{1}{2} \Delta M \dot{\phi}(0, t)^2 \quad (17.50)$$

the interaction term.

Classically, we expect that if we were to ‘illuminate’ such an impurity with a monochromatic, beamed sound wave then this would give rise to a spherical outgoing scattered wave of the same frequency, quite analogous to Thomson scattering of an electromagnetic wave by an electron.

To treat this quantum mechanically, we first write $\dot{\phi}(0, t)$ in $H_{\text{int}}(t)$ as a sum over creation and destruction operators:

$$\dot{\phi}(0, t) = \sum_{\mathbf{k}} \sqrt{\frac{\hbar \omega_{\mathbf{k}}}{2M_{\text{tot}}}} (a_{\mathbf{k}}^{\dagger}(t) - a_{\mathbf{k}}(t)) \quad (17.51)$$

and so the interaction term can be written as a double sum

$$H_{\text{int}}(t) = \frac{\Delta M \hbar}{4M_{\text{tot}}} \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \sqrt{\omega_{\mathbf{k}} \omega_{\mathbf{k}'}} (a_{\mathbf{k}}^{\dagger}(t) - a_{\mathbf{k}}(t)) (a_{\mathbf{k}'}^{\dagger}(t) - a_{\mathbf{k}'}(t)) \quad (17.52)$$

which contains all possible pairs of $a_{\mathbf{k}}$ ’s and $a_{\mathbf{k}}^{\dagger}$ ’s.

Applying (17.25) to some initial state $|i\rangle$ gives the time evolution of the state

$$\begin{aligned} |\psi(t)\rangle &= U(t, t_1) |i\rangle = |i\rangle + \frac{1}{i\hbar} \int_{t_1}^t dt H_{\text{int}}(t) |i\rangle = |i\rangle + \frac{\Delta M}{4iM_{\text{tot}}} \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \sqrt{\omega_{\mathbf{k}} \omega_{\mathbf{k}'}} \\ &\quad \times \int_{t_1}^t dt (a_{\mathbf{k}}^{\dagger} e^{i\omega_{\mathbf{k}} t} - a_{\mathbf{k}} e^{-i\omega_{\mathbf{k}} t}) (a_{\mathbf{k}'}^{\dagger} e^{i\omega_{\mathbf{k}'} t} - a_{\mathbf{k}'} e^{-i\omega_{\mathbf{k}'} t}) |i\rangle. \end{aligned} \quad (17.53)$$

Let’s suppose the system starts out in some multi-particle eigenstate of the unperturbed Hamiltonian $|i\rangle = |\dots, n_{\mathbf{k}}, \dots\rangle$. After applying the perturbation for some time T the state will no longer be in this pure eigenstate, but will contain a superposition of the original state and states obtained from the initial state by applying a pair of creation and/or destruction operators. We can ask for the amplitude that the system be in some specific eigenstate $|f\rangle = |\dots, n'_{\mathbf{k}}, \dots\rangle$ with a different set of occupation numbers. For example, let’s ask for the amplitude $\langle f | U | i \rangle$ that the occupation numbers differ only for two modes $\mathbf{k}_1, \mathbf{k}_2$ and where $n'_1 = n_1 - 1$ and $n'_2 = n_2 + 1$ i.e. an excitation has been annihilated from mode \mathbf{k}_1 and one has been created in mode \mathbf{k}_2 . With this choice of initial and final states, the only terms in the double sum over operators which produce a non-zero contribution to the amplitude are those containing the destruction operator for the mode \mathbf{k}_1 and a creation operator for mode \mathbf{k}_2 , for which we have

$$\langle f | a_{\mathbf{k}_2}^{\dagger} a_{\mathbf{k}_1} | i \rangle = \sqrt{n_1(1 + n_2)}, \quad (17.54)$$

and we get an identical result if we change the order of these operators (since this is a scattering we can assume $\mathbf{k}_1 \neq \mathbf{k}_2$, so the operators $a_{\mathbf{k}_2}^{\dagger}$ and $a_{\mathbf{k}_1}$ commute). Any other combination of operators will produce a mixing into a state which is orthogonal to $|f\rangle$ and therefore gives zero when sandwiched

between $\langle f|$ and $|i\rangle$. Taking account of both possibilities (i.e. $\mathbf{k}, \mathbf{k}' = \mathbf{k}_1, \mathbf{k}_2$ and $\mathbf{k}, \mathbf{k}' = \mathbf{k}_2, \mathbf{k}_1$) give the transition amplitude

$$\langle f|U|i\rangle = \frac{\Delta M}{2iM_{\text{tot}}} \sqrt{\omega_1 \omega_2} \sqrt{n_1(n_2 + 1)} \int dt e^{i(\omega_2 - \omega_1)t}. \quad (17.55)$$

For large T , the integral becomes $2\pi\delta(\omega_2 - \omega_1)$. Had we asked, instead, for the amplitude to make the transition to the state $|n_1 + 1, n_2 + 1\rangle$, we would have had $2\pi\delta(\omega_2 + \omega_1)$, which vanishes because the energies are positive.

The *probability* to make this transition in time T is the squared modulus of the amplitude:

$$p(\mathbf{k}_1 \rightarrow \mathbf{k}_2) = \left(\frac{\Delta M}{2M_{\text{tot}}} \right)^2 \omega_1 \omega_2 n_1(n_2 + 1) \int dt \int dt' e^{i(\omega_1 - \omega_2)(t - t')}. \quad (17.56)$$

For large T , the double integral becomes $\int dt \int d\tau e^{i(\omega_1 - \omega_2)\tau} = 2\pi T \delta(\omega_1 - \omega_2)$ so the *transition rate* (i.e. the transition probability per unit time) is

$$R(\mathbf{k}_1 \rightarrow \mathbf{k}_2) = \lim_{T \rightarrow \infty} \frac{p(\mathbf{k}_1 \rightarrow \mathbf{k}_2)}{T} = \left(\frac{\pi \Delta M}{M_{\text{tot}}} \right)^2 \omega_1 \omega_2 n_1(n_2 + 1) \delta(\omega_1 - \omega_2). \quad (17.57)$$

The δ -function here says that the phonon frequency, and therefore the energy, is unchanged in the scattering; this is elastic scattering. More generally, if the perturbation only acts for finite time T , the energy conserving δ -function is replaced by a function with width $\delta\omega \sim 1/T$. Equation (17.57) gives the rate for a transition to a specific final mode \mathbf{k}_2 , and if we integrate over all modes at the scattering frequency we can obtain the net scattering rate out of state \mathbf{k}_1 . This is quite analogous to Rayleigh scattering. Note that wave-momentum is not conserved here. This is to be expected since the presence of the impurity violated the invariance of the system under translations.

A key feature here is the dependence of the scattering rate on the initial occupation numbers:

$$R(\mathbf{k}_1 \rightarrow \mathbf{k}_2) \propto n_1(n_2 + 1). \quad (17.58)$$

This is a result of profound significance. That the rate should scale with the occupation number n_1 in the initial mode is reasonable, but we see that the rate also depends non-trivially on the initial occupation number n_2 of the *final* mode. This is the phenomenon of *stimulated emission* as originally deduced by Einstein from considering a two-state system in thermal equilibrium with black-body radiation. Here we see it arising as a fundamental property of field theory — it will apply to any bosonic field and is quite general — and it says that if a certain state has a high occupation number then the probability that particles will scatter into that state is increased.

Note that the wave-momentum is *not* conserved in scattering off an impurity, since the Lagrangian density is not independent of position.

17.4.2 Self Interactions

As discussed in the previous chapter, another interesting model for the interaction of a field is a self-interaction. In our ‘scalar-elasticity’ model this can be introduced by letting the springs K have a slightly ‘non-harmonic’ or non-quadratic behavior, so the contribution to the potential energy (and therefore to the Hamiltonian) is

$$V(\phi) = \frac{\mu^2}{2} \phi^2 \rightarrow \frac{\mu^2}{2} \phi^2 + \lambda \phi^4 \quad (17.59)$$

for example.

This type of modification will again break the non-interacting nature of the normal modes. At the classical level, the interaction term

$$H_{\text{int}} = \int d^3x \mathcal{H}_{\text{int}} = \lambda \int d^3x \phi^4 \quad (17.60)$$

will introduce a coupling between the otherwise independent normal modes. Expanding the field as a sum over creation and destruction operators $\phi \rightarrow \sum_{\mathbf{k}} \omega_{\mathbf{k}}^{-1/2} (a_{\mathbf{k}}^\dagger e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} + a_{\mathbf{k}} e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})})$ the interaction Hamiltonian will now contain a four-fold sum over wave-vectors with terms consisting of all possible combinations of four a 's or a^\dagger 's:

$$H_{\text{int}} \propto \lambda \int d^3x \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \sum_{\mathbf{k}''} \sum_{\mathbf{k}'''} \frac{(a_{\mathbf{k}}^\dagger e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} + a_{\mathbf{k}} e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})}) \dots}{\sqrt{\omega_{\mathbf{k}} \omega_{\mathbf{k}'} \omega_{\mathbf{k}''} \omega_{\mathbf{k}'''}}} \quad (17.61)$$

where \dots stands for three more copies of the term in parentheses, but with \mathbf{k} replaced by \mathbf{k}' , \mathbf{k}'' and \mathbf{k}''' . We can use this with the S -matrix expansion to compute the transition amplitude between initial and final states with definite occupation numbers. As with scattering off an impurity, once we specify the initial and final states, only a very limited subset of all of the possible combinations of operators are relevant.

For example, consider initial and final states $|i\rangle = |n_1, n_2, n_3, n_4\rangle$ and $|f\rangle = |n_1 - 1, n_2 - 1, n_3 + 1, n_4 + 1\rangle$. We are only labelling the modes for which the occupation numbers actually change. This is a scattering reaction $\mathbf{k}_1 \mathbf{k}_2 \rightarrow \mathbf{k}_3 \mathbf{k}_4$. Obviously, the only effective combinations of operators contain destruction operators for modes \mathbf{k}_1 , \mathbf{k}_2 and creation operators for \mathbf{k}_3 , \mathbf{k}_4 for which

$$\langle f | a_{\mathbf{k}_4}^\dagger a_{\mathbf{k}_3}^\dagger a_{\mathbf{k}_2} a_{\mathbf{k}_1} | i \rangle = \sqrt{n_1 n_2 (1 + n_3) (1 + n_4)}. \quad (17.62)$$

The operators can appear in any order — since we are assuming that all of the modes are distinct — and this yields $4!$ identical contributions for the $4!$ ways to assign \mathbf{k} , \mathbf{k}' , \mathbf{k}'' , \mathbf{k}''' to \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 and \mathbf{k}_4 .

With the operator algebra taken care of we can proceed to performing the spatial and time integrals:

$$\begin{aligned} \langle f | U | i \rangle &= \langle f | \frac{1}{i\hbar} \int dt \int d^3x \mathcal{H}_{\text{int}} | i \rangle \\ &\propto \lambda \sqrt{\frac{n_1 n_2 (n_3 + 1) (n_4 + 1)}{\omega_1 \omega_2 \omega_3 \omega_4}} \int_{t_1}^{t_2} dt e^{-(\omega_1 + \omega_2 + \omega_3 + \omega_4)t} \int d^3x e^{i(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3 - \mathbf{k}_4) \cdot \mathbf{x}}. \end{aligned} \quad (17.63)$$

The spatial integral is $(2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3 - \mathbf{k}_4)$. This means that the sum of wave vectors of the phonons is conserved in the scattering. We will assume that the interaction is on permanently, in which case $t_1 \rightarrow -\infty$ and $t_2 \rightarrow \infty$ and the time integral becomes $2\pi \delta(\omega_1 + \omega_2 - \omega_3 - \omega_4)$, so the sum of the temporal frequencies is also conserved in the interaction. Squaring the amplitude gives a transition probability per unit time, or *scattering rate*, of

$$\frac{p(\mathbf{k}_1, \mathbf{k}_2 \rightarrow \mathbf{k}_3, \mathbf{k}_4)}{T} \propto \lambda^2 \delta^{(4)}(\vec{k}_1 + \vec{k}_2 - \vec{k}_3 - \vec{k}_4) \frac{n_1 n_2 (1 + n_3) (1 + n_4)}{\omega_1 \omega_2 \omega_3 \omega_4}. \quad (17.64)$$

where we have defined the ‘frequency-wave number 4-vector’ $\vec{k} \equiv (\omega, \mathbf{k})$ — though this should not be construed as a Lorentz 4-vector — and $\delta^{(4)}(\mathbf{k})$ is shorthand for $\delta^{(3)}(\mathbf{k})\delta(\omega)$.

This is a scattering of phonons off each other induced by non-linearity in the springs. We use the symbol shown in figure 17.1 to denote the contribution to the complex amplitude (17.63). This is a single-vertex *Feynman diagram* with four external legs. A Feynman diagram tells us at a glance the initial and final states, and the nature of the interaction term (here the vertex with four emerging lines tells us that we are considering an interaction Hamiltonian consisting of the product of four fields). It also tells us at what order in the general S -matrix expansion we are working; here there is one vertex which tells us this is a first order contribution to the amplitude. It also happens to look like a space-time diagram of a collisional interaction between a pair of particles.

The self-interaction scattering conserves total energy and wave-momentum, as it should since the classical Lagrangian density on which the theory is based has no explicit dependence on position.

Conservation of the sum of the frequencies of the quanta involved is readily interpreted as conservation of energy, since the energy of each mode is $(n + 1/2)\hbar\omega$. Conservation of the vector sum of the wave-numbers similarly implies conservation of the total wave *momentum*. In what follows we will sometimes refer to ‘energy ω ’ or ‘momentum \mathbf{k} ’. In such blatantly dimensionally incorrect statements we are implicitly setting $\hbar = 1$.

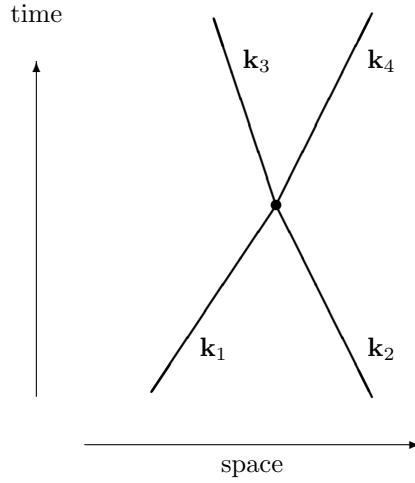


Figure 17.1: Feynman diagram for first order (i.e. single-vertex) scattering of phonons induced by an an-harmonic spring potential energy $V(\phi) = K\phi^2/2 + \lambda\phi^4$.

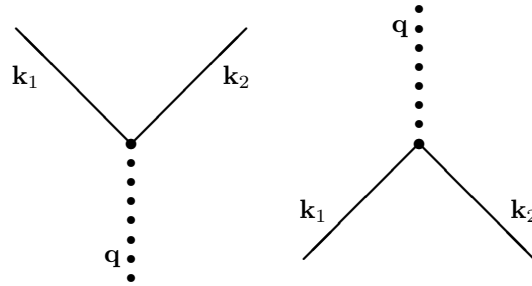


Figure 17.2: Feynman diagrams for the decay of a chion (dotted line) into a pair of phions (solid lines), and the inverse process. The interaction Hamiltonian is $H_{\text{int}} = \alpha\phi^2\chi$.

17.4.3 Second Order Scattering

The examples above were processes that can be described at first order in perturbation theory. Many important scattering processes involve the exchange of a virtual particle, and the rates for these processes require second order perturbation theory. With a slight modification to our ‘scalar-elasticity’ model we can explore such processes.

To this end, consider a medium that can support two types of phonon fields of the type we have been discussing : ϕ and χ , each with their own free-field Lagrangian density, though with different parameters $\mu = \mu_\chi, \mu_\phi$. Let the fields be coupled by a term

$$\mathcal{H}_{\text{int}} = \alpha\phi^2\chi \quad (17.65)$$

where α is a coupling constant. It should be clear that this has the potential to describe first order processes such as a ‘chion’ decaying to a pair of ‘phions’, or a pair of phions annihilating to form a chion as shown in figure 17.2. If we denote the frequency of the χ field by $\Omega(\mathbf{q})$, with \mathbf{q} the spatial frequency, then the rate for such a process would involve an energy conserving δ -function $\delta(\omega(\mathbf{k}_1) + \omega(\mathbf{k}_2) - \Omega(\mathbf{q}))$ and the momentum conserving factor $\delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{q})$.

Now what if the minimum energy (i.e. \hbar times the minimum frequency) of our chions is greater than the sum of the energy of the available phions? This is easy to arrange; recall that the minimum energy is set by the strength of the K -springs (or by the $-c\chi^2/2$ term in the continuous field Lagrangian density). In that case the first order transition $\phi + \phi \rightarrow \chi$ cannot take place. There

is, however, still the possibility of interesting second-order effects such as phion-phion scattering mediated by the exchange of a virtual chion. Calculating the evolution of the initial state in the interaction picture at second order using (17.24) gives

$$|\phi(t_f)\rangle = U|i\rangle = |i\rangle + \left(\frac{1}{i\hbar}\right)^2 \int_{t_i}^{t_f} dt H_{\text{int}}(t) \int_{t_i}^t dt' H_{\text{int}}(t') |i\rangle \quad (17.66)$$

where we have discarded the first order term.

In terms of the interaction Hamiltonian density these time integrals become integrals over space-time

$$U|i\rangle = |i\rangle + \left(\frac{1}{i\hbar}\right)^2 \int dt \int d^3x \mathcal{H}(\mathbf{x}, t) \int dt' \int d^3x' \mathcal{H}(\mathbf{x}', t') |i\rangle \quad (17.67)$$

or

$$|\phi(t_f)\rangle = |i\rangle + \left(\frac{\alpha}{i\hbar}\right)^2 \int dt \int d^3x (\phi\phi\chi)_{\mathbf{x},t} \int dt' \int d^3x' (\phi\phi\chi)_{\mathbf{x}',t'} |i\rangle \quad (17.68)$$

As usual, we now expand each of the field operators here as a sum of creation and destruction operators:

$$\phi(\mathbf{x}, t) = \sum_{\mathbf{k}} \omega(\mathbf{k})^{-1/2} (a_{\mathbf{k}}^\dagger e^{i(\omega_{\mathbf{k}}t - \mathbf{k}\cdot\mathbf{x})} + a_{\mathbf{k}} e^{-i(\omega_{\mathbf{k}}t - \mathbf{k}\cdot\mathbf{x})}) \quad (17.69)$$

and

$$\chi(\mathbf{x}, t) = \sum_{\mathbf{q}} \Omega(\mathbf{q})^{-1/2} (a_{\mathbf{q}}^\dagger e^{i(\Omega_{\mathbf{q}}t - \mathbf{q}\cdot\mathbf{x})} + a_{\mathbf{q}} e^{-i(\Omega_{\mathbf{q}}t - \mathbf{q}\cdot\mathbf{x})}) \quad (17.70)$$

where, for clarity, we have dropped the constant factor $\sqrt{2\hbar/M_{\text{tot}}}$. Since there are six fields involved this yields a six-fold sum over wave-vectors, each element of which consists of various combinations of a s and a^\dagger s.

Now let us specify the initial state as $|i\rangle = |\mathbf{k}_1, \mathbf{k}_2; 0\rangle$, by which we mean one phion with momentum \mathbf{k}_1 , and one with \mathbf{k}_2 and no chions, and the final state as $|f\rangle = |\mathbf{k}_3, \mathbf{k}_4; 0\rangle$. That is, we are considering the scattering process $\mathbf{k}_1\mathbf{k}_2 \rightarrow \mathbf{k}_3\mathbf{k}_4$. The amplitude to make the transition $\langle f|U|i\rangle$ then picks out a very limited subset of all the possible second order terms in our 6-fold sum. First, any relevant term must contain a pair of destruction operators $a_{\mathbf{k}_1}$, $a_{\mathbf{k}_2}$ to annihilate the initial particles, and a pair of creation operators $a_{\mathbf{k}_3}^\dagger$, $a_{\mathbf{k}_4}^\dagger$ to create the outgoing particles. Second, it must also contain a creation operator $a_{\mathbf{q}}^\dagger$ to create a chion, with some as yet unspecified momentum \mathbf{q} , and must then also contain a destruction operator $a_{\mathbf{q}}$ for the same momentum. This eliminates five of the sums over momenta and we are left with a single sum over the momentum \mathbf{q} of the virtual chion. One such combination of operators which gives a non-zero contribution to the amplitude $\langle f|U|i\rangle$ is

$$(a_{\mathbf{k}_2} a_{\mathbf{k}_4}^\dagger a_{\mathbf{q}}) (a_{\mathbf{k}_1} a_{\mathbf{k}_3}^\dagger a_{\mathbf{q}}^\dagger). \quad (17.71)$$

where the first triplet is associated with the space-time point (\mathbf{x}, t) and the second triplet with (\mathbf{x}', t') . This would describe the scattering of phonons via the exchange of a chion, the Feynman diagram for which is shown in figure 17.3.

The contribution to the transition amplitude from this term is then

$$\langle f|U|i\rangle \sim \alpha^2 \sum_{\mathbf{q}} \frac{\langle \mathbf{k}_3, \mathbf{k}_4; 0 | a_{\mathbf{k}_2} a_{\mathbf{k}_4}^\dagger a_{\mathbf{q}} a_{\mathbf{k}_1} a_{\mathbf{k}_3}^\dagger a_{\mathbf{q}}^\dagger | \mathbf{k}_1, \mathbf{k}_2; 0 \rangle}{\Omega_{\mathbf{q}} \sqrt{\omega_1 \omega_2 \omega_3 \omega_4}} \int dt \int d^3x \int_{t_i}^t dt' \int d^3x' \quad (17.72)$$

$$e^{i(\mathbf{k}_2 - \mathbf{k}_4 + \mathbf{q}) \cdot \mathbf{x}} e^{i(\mathbf{k}_1 - \mathbf{k}_3 - \mathbf{q}) \cdot \mathbf{x}'} e^{i(\omega_4 - \omega_2 - \Omega_{\mathbf{q}})t} e^{i(\omega_3 - \omega_1 + \Omega_{\mathbf{q}})t'}.$$

Performing the spatial integrations results in a pair of δ -functions, and the expectation value is $\langle \mathbf{k}_3, \mathbf{k}_4; 0 | \dots | \mathbf{k}_1, \mathbf{k}_2; 0 \rangle = 1$ which takes care of all the operator arithmetic. Changing the integration variable from t' to $\tau = t' - t$, the amplitude becomes

$$\langle f|U|i\rangle \sim \alpha^2 (\omega_1 \omega_2 \omega_3 \omega_4)^{-1/2} \sum_{\mathbf{q}} \Omega_{\mathbf{q}}^{-1} \delta^{(3)}(\mathbf{k}_2 - \mathbf{k}_4 + \mathbf{q}) \delta^{(3)}(\mathbf{k}_1 - \mathbf{k}_3 - \mathbf{q}) \quad (17.73)$$

$$\times \int dt e^{-i(\omega_1 + \omega_2 - \omega_3 - \omega_4)t} \int_{-\infty}^0 d\tau e^{i(\omega_3 - \omega_1 + \Omega_{\mathbf{q}})\tau}$$

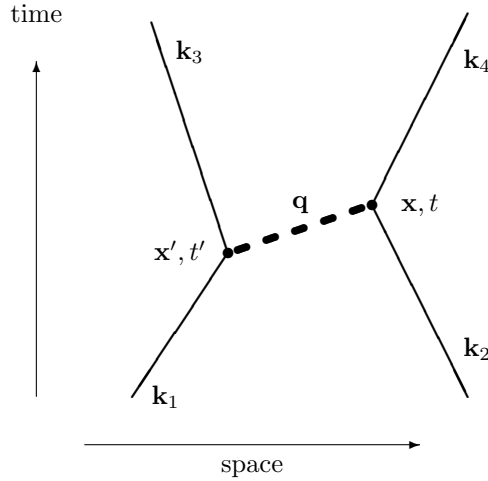


Figure 17.3: Feynman diagram for second order scattering of a phonon. The total Lagrangian density here is assumed to be that for two free fields ϕ (solid lines) and χ (dashed) with an interaction term $\mathcal{L}_{\text{int}} = \alpha\phi^2\chi$.

The two three dimensional δ -functions express conservation of 3-momentum at each of the vertices. We can use one of these to perform the summation over \mathbf{q} , since $\sum_{\mathbf{q}} \delta^{(3)}(\mathbf{k}_1 - \mathbf{k}_3 - \mathbf{q})f(\mathbf{q}) = f(\mathbf{k}_1 - \mathbf{k}_3)$, i.e. we erase this δ -function, and the summation, and replace \mathbf{q} by $\mathbf{k}_1 - \mathbf{k}_3$ throughout. The remaining δ -function becomes $\delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3 - \mathbf{k}_4)$ which conserves total wave-momentum. The dt integral similarly enforces conservation of total energy. All that remains is the $d\tau$ integral. Introducing an infinitesimal positive constant δ to ensure convergence, and defining $\epsilon = \omega_1 - \omega_3$, this is

$$\int_{-\infty}^0 d\tau e^{-i(\epsilon - \Omega_{\mathbf{q}})\tau} = \lim_{\delta \rightarrow 0} \int_{-\infty}^0 d\tau e^{-i(\epsilon - \Omega_{\mathbf{q}} + i\delta)\tau} = \frac{-1}{\epsilon - \Omega_{\mathbf{q}}}. \quad (17.74)$$

Combining the δ -functions the amplitude is then

$$\langle f|U|i \rangle \sim \alpha^2 \frac{\delta^{(4)}(\vec{k}_1 + \vec{k}_2 - \vec{k}_3 - \vec{k}_4)}{\sqrt{\omega_1 \omega_2 \omega_3 \omega_4}} \frac{1}{\Omega_{\mathbf{q}}} \frac{1}{\epsilon - \Omega_{\mathbf{q}}}. \quad (17.75)$$

The final amplitude is then a rather simple function of the external phonon momenta and energies (since $\epsilon = \omega_1 - \omega_3$ and $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_3$). However, what we have computed here is only part of the story; we have only considered the one possible combination of operators in (17.71). We need now to enumerate all of the allowed terms, and sum their contributions to the amplitude.

How many combinations like (17.71) are there that give a non-zero contribution? Let's call the vertex at (t', \mathbf{x}') vertex A and that at (t, \mathbf{x}) vertex B. The S -matrix expansion stipulates that t' precede t . Vertex A must then contain the chion creation operator and vertex B must contain the corresponding chion destruction operator. The operators for the external phions can, however, be assigned in any way we like, subject only to the condition that we have two phion operators per vertex; this being dictated by the form of the interaction (17.65). The number of ways of connecting the 4 external particles in pairs to two vertices is $4!/(2!)^2 = 6$. These are shown in figure 17.4. Evidently there are three pairs of diagrams which differ only in that in the lower member of each pair the right vertex has been 'pulled back' to precede the left vertex, and the vertex labels have been swapped.

Consider the center pair of diagrams. We have already calculated the contribution to the amplitude for the upper diagram. The lower diagram gives a very similar contribution. If we swap primed by unprimed coordinates, all of the complex exponential factors associated with the external

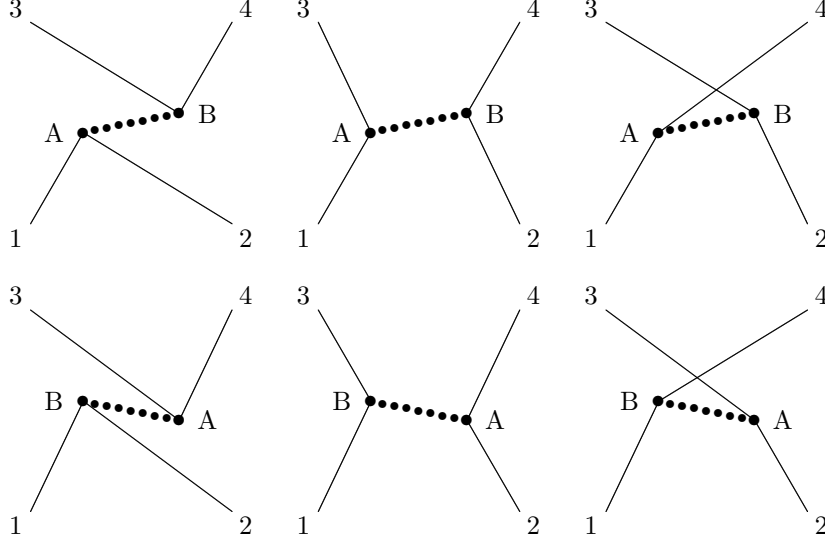


Figure 17.4: The six possible contributions to phonon-phonon scattering *via* a $\alpha\phi\phi\chi$ interaction potential. These are ‘time-ordered’ processes; in each case the vertex A precedes B .

particles are identical to what we had before. The only changes are that we must now integrate over $t < t'$ (or equivalently $t' > t$) and that we must replace the chion creation operator associated with the vertex at \mathbf{x}', t' with a destruction operator and *vice versa*. This has the effect of changing the sign in the exponent for factors like $e^{i\Omega t'}$ and $e^{i\mathbf{q}\cdot\mathbf{x}}$.

The contribution to the transition amplitude from both of these diagrams is

$$\frac{\alpha^2}{\omega_1\omega_2\omega_3\omega_4} \sum_{\mathbf{q}} \int dt \int d^3x \int d^3x' e^{i(\mathbf{k}_2 - \mathbf{k}_4 + \mathbf{q}) \cdot \mathbf{x}} e^{i(\mathbf{k}_1 - \mathbf{k}_3 - \mathbf{q}) \cdot \mathbf{x}'} e^{i(\omega_1 + \omega_2 - \omega_3 - \omega_4)t} \times \frac{1}{\Omega_{\mathbf{q}}} \left\{ \int_{-\infty}^0 d\tau e^{-i(\epsilon - \Omega_{\mathbf{q}})\tau} + \int_0^{\infty} d\tau e^{-i(\epsilon + \Omega_{\mathbf{q}})\tau} \right\} \quad (17.76)$$

Where, as before, $\epsilon = \omega_1 - \omega_3$; i.e. it is the amount of energy being transferred from the left-hand phion to the one on the right. It might appear that we should have included the $e^{i\mathbf{q}\cdot\mathbf{x}}$ etc. factors within the term in braces and with appropriate sign factors for the two cases $t' < t$ and $t' > t$. However, this is not necessary since we are integrating over all possible \mathbf{q} , so we can replace, for instance $\sum_{\mathbf{q}} e^{i\mathbf{q}\cdot\mathbf{x}}$ by $\sum_{\mathbf{q}} e^{-i\mathbf{q}\cdot\mathbf{x}}$.

The $d\tau$ integrals can be performed as above, to obtain

$$\frac{1}{\Omega_{\mathbf{q}}} \left\{ \int_{-\infty}^0 d\tau e^{-i(\epsilon - \Omega_{\mathbf{q}})\tau} + \int_0^{\infty} d\tau e^{-i(\epsilon + \Omega_{\mathbf{q}})\tau} \right\} = \frac{-1}{\epsilon - \Omega_{\mathbf{q}}} + \frac{-1}{\epsilon + \Omega_{\mathbf{q}}} = \frac{2}{\Omega_{\mathbf{q}}^2 - \epsilon^2}. \quad (17.77)$$

As before, the dt , d^3x and d^3x' integrals and the sum over \mathbf{q} become a 4-dimensional δ -function conserving energy and total 3-momentum for the diagram as a whole.

The final amplitude for this pair of diagrams taken together is then

$$\langle f|U|i \rangle \sim \alpha^2 \frac{\delta^{(4)}(\vec{k}_1 + \vec{k}_2 - \vec{k}_3 - \vec{k}_4)}{\sqrt{\omega_1\omega_2\omega_3\omega_4}} \frac{1}{\Omega_{\mathbf{q}}^2 - \epsilon^2}. \quad (17.78)$$

We symbolize this contribution by the center diagram in 17.5. The two vertices are drawn here at the same time to emphasize that this diagram accounts for both of the two cases where the left vertex precedes the right and *vice versa*. There are three such diagrams, as depicted in figure 17.5.

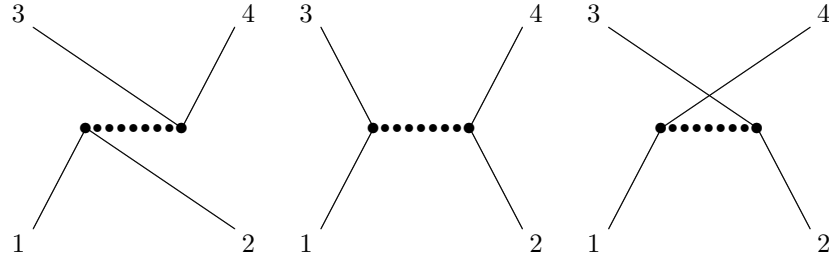


Figure 17.5: Diagrams for phonon-phonon scattering, each of which combines the contribution to the amplitude from the pair of time-ordered graphs in figure 17.4. The diagram on the left, for example, symbolizes the contribution from two separate time-ordered processes. In one, the pair of phions 1, 2 annihilate to create a chion which later decays into the phions 3, 4. In the other, the phions 3, 4 and the chion (which must have negative energy) are first created, and the chion later merges with the phions 1 and 2.

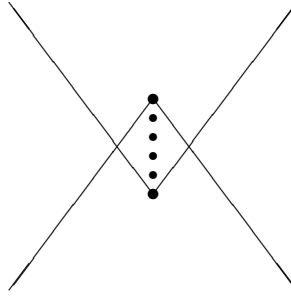


Figure 17.6: An alternative view of the process shown in the left panel of figure 17.5.

Each contains a factor $1/(\Omega_{\mathbf{p}}^2 - \epsilon^2)$ where \mathbf{p} is the momentum transferred in the exchange and ϵ is the energy transferred. For the left hand diagram, for example, $\epsilon = \omega_1 + \omega_2$, and $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_2$ while for the right diagram $\epsilon = \omega_1 - \omega_4$, and $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_4$.

Each diagram is *topologically* distinct from the others. Imagine the external lines as rubber bands connected to the corners of a frame, and the internal line as a rigid rod. We can hold the rod in any orientation, and look at it from any angle. This generates diagrams which look superficially different, but we do not add any extra contribution to the amplitude. For example, we could have drawn the diagram on the left in figure 17.5 as in 17.6, but this is topologically equivalent.

That's it — we're done. Squaring the amplitude, summed over all topologically distinct diagrams, gives the transition probability, and thereby the scattering rate for this process.

The final amplitude is very similar to that for 1st order scattering *via* a $\lambda\phi^4$ interaction (17.63). In both cases we have conservation of total momentum and energy, and in both cases there are factors $\omega^{-1/2}$ associated with each incoming and outgoing particle. Had we considered here initial states with occupation number $n_i > 1$ or final states with initial occupation number $n_f > 0$ then we would have also obtained an extra factor $\sqrt{n_1 n_2 (1 + n_3)(1 + n_4)}$ exactly as in (17.63).

An important new ingredient is the factor

$$\frac{1}{\Omega_{\mathbf{q}}^2 - \epsilon^2} \quad (17.79)$$

which we will call the *chion propagator*. It is a function of the incoming and outgoing particle energies and momenta. Now the function $\Omega(\mathbf{q})$ is just the chion dispersion relation; it gives the angular frequency for a real chions of momentum $\mathbf{p} = \hbar\mathbf{q}$. The relation between energy and momentum for real phions is then

$$E(\mathbf{p}) = \hbar\Omega(\mathbf{p}/\hbar). \quad (17.80)$$

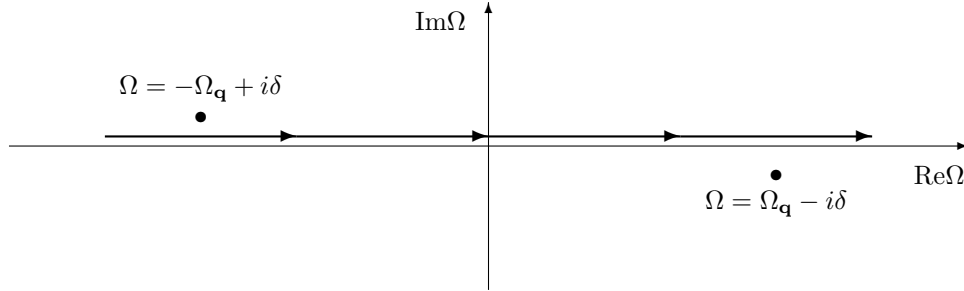


Figure 17.7: Contour integral for evaluating the chion propagator.

For the free-field Lagrangian we have been considering the dispersion relation is $a\Omega^2(\mathbf{q}) = b|\mathbf{q}|^2 + c$, so the real particle states lie on a 3-dimensional hyper-surface in energy-momentum 4-space; the ‘energy shell’. This is exactly analogous to relativistic particles which live on the ‘mass-shell’ $E^2 = p^2c^2 + m^2c^4$.

Now consider the left hand phion leg of figure 17.3. This particle comes in with energy $\hbar\omega_1$ and leaves with energy $\hbar\omega_3$, so it transfers to the chion energy $\Delta E = \hbar(\omega_1 - \omega_3)$. Similarly, it transfers an amount of momentum $\Delta\mathbf{p} = \hbar(\mathbf{k}_1 - \mathbf{k}_3)$. We have seen that 3-momentum is exactly conserved at each vertex. We can also say that energy is also conserved, since whatever energy is lost (or gained) in the interaction by the left-hand phion is gained (lost) by the right-hand phion. However, for the situation we are considering here, there is insufficient energy to create a *real* chion with momentum $\Delta\mathbf{p}$; i.e. $\Delta E < \hbar\Omega(\Delta\mathbf{p})$; the energy ΔE transferred by the chion exchange is not the same as that obtained from the dispersion relation (17.80). This is what we mean when we say that the exchanged chion is a *virtual particle*. The energy-momentum 4-vector lies off the energy shell. (In the relativistic context, we say that the exchanged particle is ‘off mass-shell’). Note that the strength of the interaction becomes larger the closer the exchanged energy is to that for a real chion with momentum $\Delta\mathbf{p}$. If we increase the energy and momentum of the incoming particles we will eventually reach energies where, if the outgoing momentum is sufficiently small, one can create a real chion, and the interaction strength then formally diverges. This is not unphysical, however, since in that case we should not have ignored the first order scattering amplitude.

Note also that for phion energies much less than $\hbar\Omega_{\min}$ the phonon propagator is effectively independent of the external particle energies. The reaction rate is then the same that one would obtain from 1st order scattering with a $\lambda\phi^4$ interaction term with a suitable choice of coupling constant λ . However, with increasing reactant energy one would see an increase in the reaction rate which would signify that one is really dealing with collisions which are being mediated by the exchanged of virtual particles for which the minimum (real particle) energy is large.

17.4.4 Contour Integral Formalism

An alternative and elegant way to evaluate the transition amplitude is *via* contour integration. The second line of (17.76) can be written as

$$\int d\tau \left[\int d\Omega \frac{e^{i(\Omega - \epsilon)\tau}}{\Omega^2 - (\Omega_{\mathbf{q}} - i\delta)^2} \right] \quad (17.81)$$

Here Ω is a complex variable and the integral is to be taken along the real axis, as indicated in figure 17.7. The integrand has poles at $\Omega = \pm(\Omega_{\mathbf{q}} - i\delta)$. For $\tau > 0$ we can complete the contour in the positive half of the complex Ω plane and we pick up the residue at $\Omega = -(\Omega_{\mathbf{q}} - i\delta)$ and for $\tau < 0$ we can complete the integral in the lower half of the plane and enclose the other pole.

The amplitude can then be written as

$$\begin{aligned} \langle f|U|i\rangle &= \frac{1}{\sqrt{\omega_1\omega_2\omega_3\omega_4}} \sum_{\mathbf{q}} \int d\Omega \frac{1}{\Omega^2 - (\Omega_{\mathbf{q}} - i\delta)^2} \\ &\times \int dt \int d^3x \int dt' \int d^3x' e^{i(\mathbf{k}_2 - \mathbf{k}_4 - \mathbf{q}) \cdot \mathbf{x}} e^{i(\mathbf{k}_1 - \mathbf{k}_3 + \mathbf{q}) \cdot \mathbf{x}'} e^{i(\omega_4 - \omega_2 - \Omega)t} e^{i(\omega_3 - \omega_1 + \Omega)t'} \\ &= \frac{1}{\sqrt{\omega_1\omega_2\omega_3\omega_4}} \sum_{\mathbf{q}} \int d\Omega \frac{1}{\Omega^2 - (\Omega_{\mathbf{q}} - i\delta)^2} \delta^{(4)}(\vec{k}_2 - \vec{k}_4 - \vec{q}) \delta^{(4)}(\vec{k}_1 - \vec{k}_3 + \vec{q}) \end{aligned} \quad (17.82)$$

This is an integral over energy-momentum space and the other integrals over the vertex coordinates are now taken without let or hindrance over the entire space-time.

This form for the amplitude is very convenient, particularly for computing more complicated graphs. If there are multiple internal lines then we have an integration over all energy and momentum values for each internal line.

17.4.5 Feynman Rules

Feynman realised that there was a rather mechanical procedure for generating the amplitude for a diagram. For our simple $\alpha\phi^2\chi$ model these are:

1. Write a factor $1/\sqrt{\omega}$ for each external leg.
2. Label all lines with momenta, and pencil in the assumed direction of momentum flow.
3. Write a factor $1/(\Omega^2 - \Omega_{\mathbf{q}}^2)$ for each internal line.
4. Write a total energy-momentum conserving δ -function.
5. Integrate over all energy-momenta that are not determined by the external particle momenta.

17.4.6 Kinematics of Scattering

Conservation of total energy and momentum imposed important *kinematic constraints* on the allowed interactions. For example, one immediate consequence is that the amplitude for processes where one particle decays into several lower energy particles of the same kind is identically zero. To see this, one can invoke the invariance under Lorentz-like transformations to transform into the frame where the momentum of the initial particle vanishes. Since the outgoing particles is necessarily greater than their rest-energies, such a process cannot conserve energy.

17.4.7 Discussion

The treatment here has been incomplete and somewhat superficial; we have ignored fermions entirely and we have considered only a rather simple bosonic field theory, a rather idealized ‘scalar-elasticity’ model. The point here is not to learn about real solids though, but to illustrate the way in which quantum field theories are constructed and how transition rates are computed. What we have sketched here is so called ‘second-quantization’, which is rather different in flavor from elementary quantum mechanics where one solves the Schroedinger equation for a wave-function. The program, in outline, is to take some classical system (defined by its wave equation, or equivalently by its Lagrangian density) as an idealized non-interacting system, find the normal modes and construct the creation and destruction operators which act on the multi-particle states (15.1). One then adds interactions, usually in the form of assumed weak couplings between different fields, express the interaction Hamiltonian in terms of the a ’s and a^\dagger ’s and compute rates for transitions using the S -matrix expansion. This proves to be very powerful.

17.5 Problems

17.5.1 Ladder Operators

a) Compute the mean square displacement $\langle n|x^2|n\rangle$ for a simple harmonic oscillator $H = p^2/2m + \omega^2 x^2/2$ using the creation and destruction operators a^\dagger , a . Note that the energy eigenstates $|n\rangle$ are orthonormal: $\langle m|n\rangle = \delta_{mn}$.

b) Show that the Heisenberg equations of motion for the creation and destruction operators a^\dagger , a are

$$\begin{aligned} da(t)/dt &= -i\omega a(t) \\ da^\dagger(t)/dt &= i\omega a^\dagger(t) \end{aligned} \tag{17.83}$$

Chapter 18

Relativistic Field Theory

18.1 The Klein-Gordon Field

The system we have been discussing is a mechanical one, with waves propagating on an underlying medium consisting of real rods, springs, beads etc. However, abstracting away the underlying medium and choosing suitable coefficients in the Lagrangian density it becomes immediately transformed into something much more interesting and exotic; a 3-dimensional relativistically covariant scalar field.

- The scalar field is denoted by $\phi(x) = \phi(\mathbf{x}, t)$, and is a Lorentz invariant quantity. Different observers assign different values to the coordinates (\mathbf{x}, t) of space-time events but they agree on the value of the field ϕ .
- The Lagrangian density for the scalar field is

$$\mathcal{L} = \frac{1}{2}\dot{\phi}^2 - \frac{1}{2}(\nabla\phi)^2 - \frac{1}{2}m^2\phi^2. \quad (18.1)$$

- This is in natural units such that $c = \hbar = 1$. In physical units

$$\mathcal{L} = \frac{1}{2}\frac{1}{c^2}\dot{\phi}^2 - \frac{1}{2}(\nabla\phi)^2 - \frac{1}{2}\frac{m^2c^2}{\hbar^2}\phi^2. \quad (18.2)$$

- The Lagrangian density has *precisely* the same form as the scalar elasticity model. However, in that model, the field ϕ had dimensions of length. Here, since the Lagrangian density has units of energy density $\mathcal{L} = ML^{-1}T^{-2}$ the field has units $\phi = M^{1/2}L^{1/2}T^{-1}$.
- The Lagrangian density (18.1) can be written

$$\mathcal{L} = -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2, \quad (18.3)$$

which is clearly covariant. The Lagrangian density is a Lorentz scalar.

- The action integral is

$$S = \int d^4x \mathcal{L}(\phi, \phi_{,\mu}). \quad (18.4)$$

This is also Lorentz invariant.

- This yields the field equation

$$\frac{\partial(\partial\mathcal{L}/\partial\phi_{,\mu})}{\partial x_\mu} - \frac{\partial\mathcal{L}}{\partial\phi} = 0. \quad (18.5)$$

or

$$\ddot{\phi} - \nabla^2\phi + m^2\phi = 0 \quad (18.6)$$

or again, equivalently,

$$\square\phi + m^2\phi = 0 \quad (18.7)$$

which is the *Klein-Gordon equation* for a massive scalar field.

- We are working here in ‘natural’ units such that both \hbar and c are unity. If we put these back in, the field equation is $\ddot{\phi}/c^2 - \nabla^2\phi + (m^2c^2/\hbar^2)\phi = 0$.
- This is a purely classical field equation analogous to the electromagnetic field equations for the field \vec{A} . Planck’s constant appears only as a parameter in the mass term.
- The traveling wave solutions of this equation have dispersion relation $\omega^2 = k^2 + m^2$ (or $\hbar^2\omega^2 = \hbar^2c^2k^2 + m^2c^4$ in physical units) which we recognize as the relativistically correct relation between energy and momentum for a particle of mass m . The scale k_* in the phonon model now becomes the (inverse of) the Compton wavelength for the particle.
- Quantization of the normal modes of this classical equation yields massive, neutral spin-less non-interacting ‘pions’.
- The propagator (17.79) becomes $1/(\vec{k} \cdot \vec{k} - m^2)$ which is again covariant.
- As discussed earlier, a complex field can represent charged particles, though we shall not explore that here.

This is interesting. What started out as a simple physical model with beads on springs has become a relativistic massive scalar field. But aside from rescaling of the parameters in the Lagrangian it is the same physical system, so we can draw useful analogies between the concrete beads and springs system and the more abstract relativistic field. With further slight modifications, our beads and springs system can also illustrate some other interesting features of field theory. Let us explore a few of these.

What if the K -spring were made an-harmonic? Specifically, what would happen if we were to replace the potential energy term $K\phi^2/2$ (which becomes the mass term $m^2\phi^2/2$ in the scalar field Lagrangian) by some more general potential

$$V(\phi) = \frac{1}{2}m^2\phi^2 + \lambda\phi^4 \quad (18.8)$$

say? As discussed above, this leads to scattering of the quanta which would be described by a single-vertex Feynman diagram with four legs. An interesting feature of the transition matrix element for the relativistically covariant field is that the 1-dimensional energy conserving δ -function and the 3-dimensional wave-number conserving δ -function in (17.64) combine to form a 4-dimensional δ -function which ensures conservation of total 4-momentum in the scattering process.

What if the potential is of the form $V(\phi) = \text{constant} - a\phi^2 + b\phi^4$ with a, b positive constants? This kind of w -shaped potential (see figure 18.1) leads to *spontaneous symmetry breaking* where at low temperatures the field will want to settle into one of the two minima, and leads to the formation of ‘domains’ within which the field is spatially constant, separated by ‘domain-walls’ where there is concentration of field gradients (and therefore potential energy). Higher dimensional fields are possible, and these lead to other possible *topological defects*. These have been considered as possible mechanism for the formation of cosmological structures.

What if we have multiple scalar fields? Clearly if we add a new independent field ψ to give a total Lagrangian density $\mathcal{L} = \mathcal{L}(\phi, \mu, \phi) + \mathcal{L}(\psi, \mu, \psi)$ this will give decoupled non-interacting quanta. However, what if we add an interaction term $\mathcal{L}_{\text{int}} = -\alpha\phi^2\psi^2$? Such an interaction would allow scattering of pions off psions *via* a single vertex diagram (i.e. a first order process). Similarly, a term $\alpha\phi^2\psi$ would allow pion-pion scattering *via* the exchange of a virtual psion.

What if the Lagrangian for the second field ψ has potential $V(\psi) = \text{constant} - a\psi^2 + b\psi^4$ and this field has settled into one of the asymmetric minima? This means that within any domain ψ will be constant, and from the point of view of the ϕ -field there appears to be a potential $\alpha\psi^2\phi^2 = M_{\text{eff}}\phi^2/2$ so this gives rise to an effective mass term for the pions even if the ϕ field started out massless. This

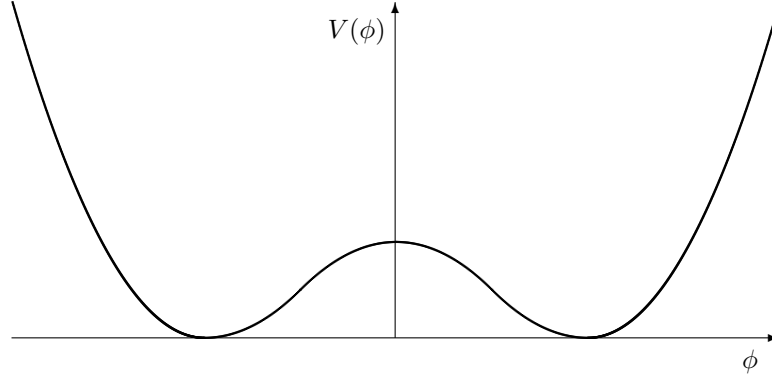


Figure 18.1: Illustration of highly an-harmonic w -shaped potential that arises in theories with spontaneous symmetry breaking.

may sound baroque, but this is what underlies the Weinberg-Salaam model for weak interactions where the ψ field is the Higgs field and the ϕ field represents the W or Z vector bosons, which acquire a mass in just this manner.

18.2 Quantum Electrodynamics

Other theories can be constructed by introducing different fields, which need not be scalars, and specifying the couplings between them. Once such example is quantum electrodynamics, where the coupling is

$$\mathcal{H}_{\text{int}} \propto \psi A \psi \quad (18.9)$$

where ψ represents the field for an electron, which is a 4-component ‘spinor’ it turns out, and A the electromagnetic 4-potential.

This is different from the scalar field analysis since the electron is a fermion. Fermions obey the exclusion principle. This means that the occupation numbers can take only the values 0, and 1, and that the creation operator applied to an occupied state must vanish. This is accomplished by replacing the commutation relation for bosonic ladder operators by an anti-commutation relation with $[a, a^\dagger] \rightarrow \{a, a^\dagger\}$. This has the interesting consequence that the rates for scattering will involve factors $1 - n_f$ where n_f represents the initial occupation number of the final state, rather than $1 + n_f$; the transition rate is proportional to the ‘unoccupation number’ of the final state. In many other respects, interactions involving fermionic fields are treated much as above.

The 3-field interaction (18.9) does not admit any real single vertex reactions because 4-momentum cannot be conserved. Real reactions appear in second order perturbation theory through terms like

$$\psi^\dagger A^\dagger \psi \psi^\dagger A \psi \quad (18.10)$$

which (reading right to left as usual) destroys an initial electron and photon, creates a virtual electron (which may be off mass-shell) which then gets destroyed, followed by creation of an outgoing electron and an outgoing photon. This type of term describes Compton scattering (figure 18.2). Other products of factors describe scattering of electrons by exchange of a virtual photon etc. with diagrams similar to those in figure 18.2. Because electrons are charged, not all of the processes depicted in figure 17.5 are allowed. For electron-electron scattering, there is no analog of the left-hand diagram since two electrons cannot annihilate to make a photon. Such a diagram does, however, contribute to electron-positron scattering, but then only one of the other diagrams contributes since an electron cannot transmute into a positron.

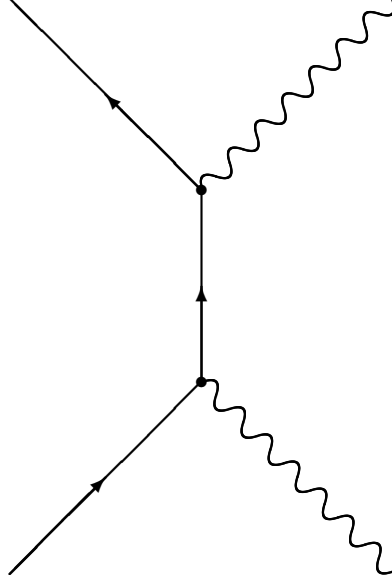


Figure 18.2: Feynman diagram for the lowest order contribution to the amplitude for Compton scattering. Solid lines denote electrons and wiggly lines photons.

18.3 Connection to Kinetic Theory

The reaction rates obtained as above provide a key input to *kinetic theory*, which describes the evolution of the 6-dimensional phase space density for particles $f(\mathbf{r}, \mathbf{p})$.

The phase space density is very closely related to the occupation number $n_{\mathbf{k}}$. These were derived for a unit volume box, for which the separation of states is $\delta k = 2\pi/L$. If the density of particles is uniform in space, so f is a function only of momentum, the number of particles in 3-dimensional volume element of momentum space is $N = f(\mathbf{p})L^3 d^3 p$, whereas, with $\mathbf{p} = \hbar \mathbf{k}$, the sum of the occupation numbers for the modes in this volume is $N = \bar{n} d^3 p / (2\pi\hbar/L)^3$ where \bar{n} is the mean occupation number, and therefore

$$\bar{n} = fh^3. \quad (18.11)$$

Now the mean number of reactions taking place in volume V and time T in which there are initial particles in momentum space elements $d^3 p_1$ and $d^3 p_2$ and final particles in momentum space elements $d^3 p_3$ and $d^3 p_4$ is then, for the $\lambda\phi^4$ self-interaction model,

$$N(\vec{p}_1, \vec{p}_2 \rightarrow \vec{p}_3, \vec{p}_4) \propto VT \lambda^2 \delta^{(4)}(\vec{p}_1 + \vec{p}_2 - \vec{p}_3 - \vec{p}_4) \times n_1 n_2 (1 + n_3)(1 + n_4) \frac{d^3 p_1}{E_1} \frac{d^3 p_2}{E_2} \frac{d^3 p_3}{E_3} \frac{d^3 p_4}{E_4} \quad (18.12)$$

where the energy factors arise from the frequency factors in the denominator in (17.64). This is already manifestly covariant; the factor VT is Lorentz invariant, the occupation number is proportional to the phase space density which is also Lorentz invariant, and each of the momentum-space volumes is paired with the corresponding energy in the combination $d^3 p/E$ which is also Lorentz invariant.

One can also express the number of reactions N as the minus the rate of change of n_1 with time times the volume V , time interval T , and momentum space volume $d^3 p_1$ as

$$N(\vec{p}_1, \vec{p}_2 \rightarrow \vec{p}_3, \vec{p}_4) = -VT \dot{n}_1 d^3 p_1. \quad (18.13)$$

However, we also need to allow for the possibility of inverse reactions, where particles are removed from modes \mathbf{k}_3 and \mathbf{k}_4 and particles are created in modes \mathbf{k}_1 and \mathbf{k}_2 . This introduces an extra term in the rates which has opposite sign, but has the input and output states exchanged. Both forward

and inverse reactions can be treated together if we replace the *quantum mechanical statistical-factor* $n_1 n_2 (1 + n_3)(1 + n_4)$ by

$$n_1 n_2 (1 + n_3)(1 + n_4) - n_3 n_4 (1 + n_1)(1 + n_2). \quad (18.14)$$

Fermionic reactants can be included here by changing plus signs to minus signs.

The rate of change of the mean occupation number is then given by

$$E_1 \frac{dn_1}{dt} = -\lambda^2 \int \frac{d^3 p_2}{E_2} \int \frac{d^3 p_3}{E_3} \int \frac{d^3 p_4}{E_4} \delta^{(4)}(\vec{p}_1 + \vec{p}_2 - \vec{p}_3 - \vec{p}_4) \times (n_1 n_2 (1 + n_3)(1 + n_4) - n_3 n_4 (1 + n_1)(1 + n_2)). \quad (18.15)$$

This is a remarkable and powerful equation. It is fully relativistic — since all of the factors are explicitly Lorentz invariant — and it is also fully quantum mechanical, since it includes e.g. the stimulated emission factors for bosons and Fermi-blocking for fermions. There are three similar equations giving the rate of change of n_2 , n_3 and n_4 . This was obtained for a specific interaction, but other reactions can be included by replacing the coupling constant λ with the transition amplitude, which, as we have seen, is also an invariant function of the particle 4-momenta.

- This is a deterministic system which can be integrated forward in time to compute the reactions in a gas of relativistic particles. For example, one can split momentum space up into a finite grid of cells and assign initial values to the occupation numbers. Then, for each cell \mathbf{k}_1 one could loop over the subset of all triplets $\mathbf{k}_2, \mathbf{k}_3, \mathbf{k}_4$ such that $\vec{p}_1 + \vec{p}_2 - \vec{p}_3 - \vec{p}_4$ vanishes, and increment or decrement n_1 appropriately for some small time interval Δt . Repeating this laborious, but conceptually straightforward, process gives the evolution of the occupation numbers $n(\mathbf{p})$ with time.
- The right hand side is really a 5-dimensional integral (three triple integrals containing four δ -functions). This makes sense. To compute the rate of loss of particles from mode \mathbf{k}_1 we need to integrate over the momenta \mathbf{k}_2 of the other input particle. The output particles, being real, have six degrees of freedom, four of which are fixed by energy-momentum conservation, leaving two free variables to integrate out. These could be taken to be the direction of one of the outgoing particles, for example.
- The output of this program is the number density of particles $n = \int d^3 p f \rightarrow \int dE E p f$, the energy density $\rho = \int dE E^2 p f$, and other quantities such as the pressure, entropy etc can also be extracted from the final occupation numbers.
- An important and direct consequence of this expression is that in equilibrium, where the mean rates must all vanish, we require that the statistical factor (18.14) must vanish and this implies that the mean occupation numbers as a function of energy are given by the Bose-Einstein and Fermi-Dirac expressions for the bosons and fermions respectively.

18.4 The Scalar Field in an Expanding Universe

Scalar fields are cherished by cosmologists. They have been invoked as the ‘inflaton’ which drives inflation; scalar fields have been considered as candidates for the dark matter (the classical example here being the ‘axion’), and the recent evidence for an accelerating universe has rekindled interest in the possibility that the universal expansion has recently become controlled by another inflaton-like field.

To help understand how such fields behave in a cosmological context let us reformulate the equations of motion in expanding coordinates. Specifically let us transform from physical spatial coordinates \mathbf{x} to ‘comoving’ coordinates \mathbf{r} defined such that $\mathbf{x} = a(t)\mathbf{r}$ with $a(t)$ the cosmological scale factor.

The action $S = \int d^4x \mathcal{L}(\dot{\phi}, \nabla_x \phi, \phi)$ becomes

$$S = \int dt \int d^3r \mathcal{L}(\dot{\phi}, \nabla_r \phi, \phi, t) \quad (18.16)$$

where now the Lagrangian, in terms of these new variables, is

$$\mathcal{L}(\dot{\phi}, \nabla_r \phi, \phi, t) = \frac{a^3}{2} \dot{\phi}^2 - \frac{a}{2} (\nabla_r \phi)^2 - \frac{a^3 m^2}{2} \phi^2. \quad (18.17)$$

Taking the various partial derivatives and combining them in the usual way yields the *Klein-Gordon equation in expanding coordinates*

$$\ddot{\phi} + 3H\dot{\phi} - \frac{1}{a^2} \nabla_r^2 \phi + m^2 \phi = 0 \quad (18.18)$$

where $H \equiv \dot{a}/a$ is the expansion rate. This differs most significantly from the equation in non-expanding coordinates by the inclusion of the ‘damping term’ $3H\dot{\phi}$ which, as its name suggests, causes the field amplitude to decay as the Universe expands.

To compute how the field amplitude decays we proceed as follows:

- We consider a single plane wave with (comoving) spatial frequency \mathbf{k} and let $\phi(\mathbf{r}, t) = \phi_0(t) e^{i\mathbf{k} \cdot \mathbf{r}}$.
- We re-scale the field, and define an auxiliary variable $\psi(t) = a(t)^{3/2} \phi_0(t)$.
- This re-scaling gets rid of the ‘damping term’ and the equation of motion for ψ is

$$\ddot{\psi} + \Omega^2(t) \psi = 0. \quad (18.19)$$

- This is an oscillator equation with time varying frequency

$$\Omega^2(t) = k^2/a^2(t) + m^2 \quad (18.20)$$

where we have dropped some terms which become negligible if $k^2/a^2 + m^2 \gg H^2$, or equivalently if the period of oscillation of the ϕ field is small compared to the age of the Universe $t \sim 1/H$.

- We then apply adiabatic invariance, which tells us that the re-scaled field amplitude varies as $\psi \propto \Omega(t)^{-1/2}$.
- We divide ψ by $a^{3/2}$ to obtain the true field amplitude ϕ_0 .

There are two limiting cases of interest: The first is where $m \gg k/a$, or equivalently that the physical wavelength $\lambda \sim a/k \gg 1/m$ or again equivalently $\lambda \gg \lambda_c$ the Compton wavelength. We expect this to correspond to non-relativistic quanta. The frequency Ω is then nearly constant in time so the amplitude of the re-scaled field is nearly constant and the amplitude of the real field ϕ_0 decays as $\phi_0 \propto a^{-3/2}$. The energy density is proportional to the kinetic energy term in the Lagrangian density, and so the energy density of the particles decays as $\dot{\phi}^2 \sim \Omega^2 \phi^2 \propto 1/a^3$. This is just what one would expect for non-relativistic quanta — the energy of each particle is just the rest mass, which is constant, and the number density falls as $1/a^3$ as the Universe expands.

The other extreme is $m \ll k/a$, corresponding to $\lambda \ll \lambda_c$. The frequency Ω then varies with the expansion as $\Omega \simeq k/a \propto 1/a$ and the re-scaled field amplitude is $\psi \propto \Omega^{-1/2} \propto a^{1/2}$. The amplitude of the true field then decays as $\phi \propto 1/a$ and the energy density is $\dot{\phi}^2 \sim \Omega^2 \phi^2 \propto 1/a^4$. However, this again is exactly what one would expect for relativistic quanta; the number of quanta dilutes as $1/a^3$ as before, but each quantum is losing energy proportional to $1/a$ so the density falls as $1/a^4$.

The foregoing discussion applies to spatially incoherent field configurations representing quanta of radiation or randomly moving thermal particles etc. The scalar field is bosonic, like the electromagnetic field, and so can also appear in macroscopic configurations in a manner analogous to macroscopic magnetic fields. In the long wavelength limit, the field equation is $\ddot{\phi} + 3H\dot{\phi} + m^2 \phi = 0$.

This has the decaying oscillatory solutions $\phi \propto a^{-3/2}e^{imt}$ already discussed, provided $m \gg H$, but at sufficiently early times or for sufficiently light fields this condition will be broken and the field equation becomes effectively $\ddot{\phi} + 3H\dot{\phi} = 0$ which admits solutions with ϕ constant in time. These nearly constant field configurations have the rather interesting property that the Lagrangian density, and therefore the energy density are constant. This means that such a field configuration is under enormous tension, since the universal expansion must be doing work at a great rate to maintain the constant mass-energy density. This is somewhat analogous to the tension along a static magnetic field, but here we have isotropic tension. Another consequence, as we shall see, is that the universe must expand exponentially if it is dominated by such a field, and this is the basis for inflation.

18.5 Non-Relativistic Scalar Fields

It is interesting to look at the evolution of the scalar field in the non-relativistic limit. This corresponds to long-wavelength variations such that $\nabla^2\phi \ll m^2\phi$ (or equivalently that the wavelength greatly exceeds the Compton wavelength).

Since it is the $\nabla^2\phi$ term which couples the different oscillators, to a first approximation this means that if we start with $\phi(\mathbf{r}, t=0) = \phi_0(\mathbf{r})$ then the field will just sit there and oscillate at the Compton frequency without spreading: $\phi(\mathbf{r}, t) \simeq \phi_0(\mathbf{r})e^{imt}$. This becomes exact in the limit that $\lambda \rightarrow \infty$. For a large but finite wavelength disturbance we expect that the disturbance will tend to diffuse away from the initial location.

To describe this mathematically let us factor out the common rapid oscillation factor e^{imt} and set

$$\phi(\mathbf{r}, t) = \psi(\mathbf{r}, t)e^{imt} + \psi^*(\mathbf{r}, t)e^{-imt} \quad (18.21)$$

where $\psi(\mathbf{r}, t)$ is a slowly varying field. By construction ϕ will be real. Taking the time derivative of the field gives

$$\dot{\phi} = im\psi e^{imt} - im\psi^* e^{-imt} + \dot{\psi}e^{imt} + \dot{\psi}^*e^{-imt} \quad (18.22)$$

where the first two terms here are much larger than the last two. Taking a further time derivative yields

$$\ddot{\phi} = -m^2(\psi e^{imt} + \psi^* e^{-imt}) + 2im(\dot{\psi}e^{imt} - \dot{\psi}^*e^{-imt}) + \mathcal{O}(\ddot{\psi}e^{imt}). \quad (18.23)$$

The Laplacian of the field is

$$\nabla^2\phi = \nabla^2\psi e^{imt} + \nabla^2\psi^* e^{-imt} \quad (18.24)$$

and combining these in the field equation $\ddot{\phi} - \nabla^2\phi + m^2\phi = 0$ gives

$$2im(\dot{\psi}e^{imt} - \dot{\psi}^*e^{-imt}) - \nabla^2\psi e^{imt} + \nabla^2\psi^* e^{-imt} = 0. \quad (18.25)$$

Since ψ is supposed to be relatively slowly varying compared to e^{imt} this requires that both the coefficient of e^{imt} and of e^{-imt} must vanish, which means that

$$i\dot{\psi} - \frac{\nabla^2\psi}{2m} = 0 \quad (18.26)$$

which is just the Schroedinger equation.

This is a familiar equation appearing in a perhaps unfamiliar context. More usually, this equation is used to describe the quantum mechanical wave function of a fermionic particle like the electron, whereas here it appears as the non-relativistic limit of a classical wave equation for a bosonic field!

Returning to our beads and springs model, we have argued that if we set this oscillating with some long-wavelength amplitude modulation pattern then to first order the energy density (squared amplitude of the field) remains localized, but over long periods of time the energy will diffuse away from its initial location, and this is described by the Schroedinger equation. One consequence of this is that a non-relativistic scalar field, while being a fundamentally wave mechanical system, can mimic the behavior of classical non-relativistic particles by virtue of the correspondence principle.

One can also introduce the effect of Newtonian gravity by letting the mass parameter $m \rightarrow m(1 + \Phi(\mathbf{r}))$ where $\Phi(\mathbf{r})$ is the Newtonian gravitational potential. One then finds that the Schroedinger

equation contains an additional $m\Phi(\mathbf{r})\psi$ energy term. This could provide an alternative to N-body integration for instance. Rather than treat a set of classical particles with some given initial phase-space distribution function, one can set up some corresponding initial ϕ field with the same macroscopic properties. Specifically this would mean that if one were to take a region of space which is small, but still much larger than the wavelength, and Fourier transform the field, the resulting energy density as a function of \mathbf{k} should be proportional to $f(\mathbf{p})$. The gravitational field could then be calculated using Poisson's equation $\nabla^2\Phi = 4\pi G\rho$ with $\rho \propto \psi^2$. This provides a pair of coupled equations with which we can evolve the system.

18.6 Problems

18.6.1 Stress-Energy Tensor

Consider the Lagrangian density $\mathcal{L}(\partial_\mu\phi, \phi)$ for a field $\phi(\vec{x})$ (as usual $\partial_\mu\phi$ denotes partial derivative with respect to space-time coordinates).

- Write down the Euler-Lagrange equation for this system.
- Use this to show that the partial derivative of the Lagrangian density with respect to space time coordinate x^σ can be written as

$$\frac{\partial\mathcal{L}}{\partial x^\sigma} = \frac{\partial}{\partial x^\mu} \left(\frac{\partial\mathcal{L}}{\partial\phi_{,\mu}} \frac{\partial\phi}{\partial x^\sigma} \right). \quad (18.27)$$

- Show, thereby, that

$$T^{\mu\nu}{}_{,\nu} = 0 \quad (18.28)$$

where the *stress-energy tensor* $T^{\mu\nu}$ is

$$T^{\mu\nu} = \eta^{\mu\nu}\mathcal{L} - \eta^{\nu\sigma} \frac{\partial\mathcal{L}}{\partial\phi_{,\mu}} \frac{\partial\phi}{\partial x^\sigma}. \quad (18.29)$$

This analysis is very similar to the analysis of energy conservation in the dynamics section of the notes (though technically considerably more challenging).

Just as the equation $j^\mu{}_{,\nu}$ expresses the conservation of the total electric charge $Q = \int d^3x j^0$, the equation $T^{\mu\nu}{}_{,\nu} = 0$ expresses the conservation of a four-vector $P^\mu = \int d^3x T^{0\mu}$; the total four-momentum of the system. See Landau and Lifshitz “Classical Theory of Fields” §32 for a detailed discussion.

18.6.2 Klein-Gordon Field

- For the Klein-Gordon Lagrangian density

$$\mathcal{L} = -\frac{1}{2}\eta^{\mu\nu}\partial_\nu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2 \quad (18.30)$$

show that the stress-energy tensor is

$$T^{\mu\nu} = \partial^\mu\phi\partial^\nu\phi + \eta^{\mu\nu}\mathcal{L}; \quad (18.31)$$

that the time-time component of $T^{\mu\nu}$, which is the energy density, is

$$\rho = T^{00} = \frac{1}{2}[\dot{\phi}^2 + (\nabla\phi)^2 + m^2\phi^2], \quad (18.32)$$

and that the average of the spatial diagonal components, which is the pressure, is

$$P = \frac{1}{3}T^{ii} = \frac{1}{2}[\dot{\phi}^2 - (\nabla\phi)^2/3 - m^2\phi^2]. \quad (18.33)$$

- What is the relation between the pressure and density — the ‘equation of state’ — for the following field configurations.

1. A sea of incoherent random waves with $k \gg m$.
2. A sea of incoherent random waves with $k \ll m$.
3. A spatially uniform and static field $\phi = \text{constant}$.

18.6.3 Scalar Field Pressure

Starting from the Klein-Gordon Lagrangian density (in natural units)

$$\mathcal{L} = \frac{1}{2}(\dot{\phi}^2 - (\nabla_x \phi)^2 - m^2 \phi^2) \quad (18.34)$$

obtain the Euler-Lagrange equation in an expanding Universe where comoving and physical coordinates are related by $\mathbf{r} = \mathbf{x}/a(t)$. Here ∇_x denotes derivative with respect to physical spatial coordinate.

The energy density for the field is

$$\rho = \frac{1}{2} \left[\dot{\phi}^2 + \frac{1}{a^2} (\nabla \phi)^2 + m^2 \phi^2 \right] \quad (18.35)$$

where ∇ denotes derivative with respect to comoving spatial coordinate. In the ‘beads and springs’ analog what do the three terms represent?

Combine these to show that the rate of change of the density, averaged over some comoving volume, is

$$\frac{d\langle \rho \rangle}{dt} = -\frac{\dot{a}}{a} \left[3\langle \dot{\phi}^2 \rangle + \frac{1}{a^2} \langle (\nabla \phi)^2 \rangle \right] + \text{surface terms.} \quad (18.36)$$

where the surface terms become negligible if the averaging volume becomes large, and vanish if we impose periodic boundary conditions.

Now the first law of thermodynamics (and $E = Mc^2$) imply that for a homogeneous expanding Universe, the rate of change of the density is

$$\frac{d\rho}{dt} = -3\frac{\dot{a}}{a}(\rho + P). \quad (18.37)$$

Show thereby that the mean pressure is

$$\langle P \rangle = \frac{1}{2} \left[\langle \dot{\phi}^2 \rangle - \langle (\nabla_x \phi)^2 \rangle / 3 - m^2 \langle \phi^2 \rangle \right]. \quad (18.38)$$

What is the equation of state (relation between P and ρ) for highly relativistic waves.

18.6.4 Domain Walls and Strings

a) Consider a scalar field $\phi(\mathbf{x}, t)$ with ‘spontaneous symmetry breaking’ potential

$$V(\phi) = \text{constant} - a\phi^2 + b\phi^4 \quad (18.39)$$

with a, b positive constants. Let the minima of the potential be $V = 0$ at $\phi = \pm\phi_0$ and let $V(0) = V_0$.

- Sketch the minimum energy field configuration, if we require that $\phi = \pm\phi_0$ as $x_1 \rightarrow \pm\infty$.
- What is the width of the domain wall?
- What is the surface density of the wall?

b) Now consider a two component field with the analogous ‘wine bottle bottom’ potential:

$$V(\phi_1, \phi_2) = \text{constant} - a(\phi_1^2 + \phi_2^2) + b(\phi_1^2 + \phi_2^2)^2 \quad (18.40)$$

with potential minimum $V = 0$ at $\phi_1^2 + \phi_2^2 = \phi_0^2$ and again $V(0, 0) = V_0$.

- Describe the minimum energy field configuration, if we require that $(\phi_1, \phi_2) \rightarrow (x_1, x_2)/\sqrt{x_1^2 + x_2^2}$ as $x_1^2 + x_2^2 \rightarrow \infty$?
- What is the thickness of the string?
- What is the line density of the string?

Order of magnitude estimates are sufficient.

Part IV

Matter

Chapter 19

Kinetic Theory

Kinetic Theory provides the microscopic basis for gas dynamics. In the dilute gas approximation, such that the de Broglie wavelength is much less than the mean particle separation (which implies that the occupation number be small so stimulated emission and exclusion principle are unimportant) we can consider the gas to be a set of classical particles moving ballistically with occasional collisions.

The collisions are described by the *cross-section* σ which gives the rate of collisions and the distribution of deflection angles for the particles. This may be computed quantum mechanically or determined experimentally and is assumed here to be a given function of deflection angle and relative velocity. We shall also focus on non-relativistic gases such that $kT \ll mc^2$.

We shall derive the Boltzmann equation which describes the evolution of the phase-space density $f(\mathbf{r}, \mathbf{v})$ for atoms in the gas, first for a collisionless gas and then study the effect of collisions to obtain the *Boltzmann transport equation*. We discuss equilibrium solutions of this equation, and Boltzmann's H-theorem. We obtain the ideal gas laws as a limiting idealization in the case of short mean free path, and discuss qualitatively the effects of viscosity and heat conduction.

19.1 The Collisionless Boltzmann Equation

Consider a 6-dimensional volume $d^6w = d^3r d^3v$ of phase space (we shall assume equal mass particles, so velocity and momentum are effectively equivalent). Imagine the volumes to be small compared to the scale over which macroscopic conditions vary but still large enough to contain a huge number of particles.

The number of particles in d^6w is $f(w)d^6w$. The rate of change of this number is the integral of the flux of particles across the boundary of d^6w . For a finite volume w we have

$$\frac{d}{dt} \int_w d^6w f = - \int_S dS \mathbf{n} \cdot (f \dot{\mathbf{w}}) \quad (19.1)$$

with S the surface of w and \mathbf{n} its normal. The 6-dimensional divergence theorem relates such a surface integral to the volume integral of the 6-dimensional divergence $\int dS \mathbf{n} \cdot (f \dot{\mathbf{w}}) = \int d^6w \nabla \cdot (f \dot{\mathbf{w}})$ and therefore

$$\int_w d^6w \left[\frac{\partial f}{\partial t} + \nabla \cdot (f \dot{\mathbf{w}}) \right] = 0 \quad (19.2)$$

from which follows the 6-dimensional continuity equation

$$\frac{\partial f}{\partial t} + \nabla \cdot (f \dot{\mathbf{w}}) = 0. \quad (19.3)$$

This result can also be obtained by considering the flux of particles through the sides of a small 6-cube.

Writing this out in more detail

$$\frac{\partial f}{\partial t} + \sum_{i=1}^3 \frac{\partial}{\partial x_i} (f \dot{x}_i) + \sum_{i=1}^3 \frac{\partial}{\partial v_i} (f \dot{v}_i) = 0 \quad (19.4)$$

or

$$\frac{\partial f}{\partial t} + \dot{\mathbf{x}} \cdot \frac{\partial f}{\partial \mathbf{x}} + \dot{\mathbf{v}} \cdot \frac{\partial f}{\partial \mathbf{v}} = -f \sum_i \left(\frac{\partial \dot{x}_i}{\partial x_i} + \frac{\partial \dot{v}_i}{\partial v_i} \right) \quad (19.5)$$

but using Hamilton's equations $\dot{x}_i = \partial H / \partial p_i$ and $\dot{p}_i = -\partial H / \partial x_i$ with $\mathbf{p} = m\mathbf{v}$ shows that the right hand side vanishes identically. The left hand side is the *convective derivative* of the phase-space density; it gives the rate of change of f as seen by an atom moving with 6-velocity $\dot{\mathbf{w}} = (\dot{\mathbf{x}}, \dot{\mathbf{v}})$. We denote this by $df/dt = Lf$ where the differential *Liouville operator* is $L \equiv \partial/\partial t + \dot{\mathbf{x}} \cdot \partial/\partial \mathbf{x} + \dot{\mathbf{v}} \cdot \partial/\partial \mathbf{v}$ and we obtain the collisionless Boltzmann equation, or Liouville's equation

$$\frac{df}{dt} = Lf = \left(\frac{\partial}{\partial t} + \dot{\mathbf{x}} \cdot \nabla_{\mathbf{x}} + \dot{\mathbf{v}} \cdot \nabla_{\mathbf{v}} \right) f = 0. \quad (19.6)$$

This equation says that the phase-space density behaves like an incompressible fluid; the phase-space density remains constant along the path of a particle. The space-density of particles may vary, but it must be accompanied by a change in the momentum-space density so that the product $n(\mathbf{x})n(\mathbf{v})$ remains constant. The phase-space density in the vicinity of different particles will, in general, differ.

This should not be confused with the invariance of $f(\mathbf{x}, \mathbf{v})$ under Lorentz boosts noted earlier. It is however intimately related to the adiabatic invariance of $\oint dp dq$. Consider a collection of particles oscillating in a pig-trough. These particles will fill some region of the 2-dimensional phase space. If we change the energy or the profile of the trough, this will change the boundary of the region of phase-space they occupy, but the density of each volume of this 'fluid' remains constant, so the total area of this region remains fixed.

19.2 The Boltzmann Transport Equation

To describe the effect of collisions we augment the collisionless Boltzmann equation with a collision term

$$Lf = \left(\frac{\partial}{\partial t} + \dot{\mathbf{x}} \cdot \nabla_{\mathbf{x}} + \dot{\mathbf{v}} \cdot \nabla_{\mathbf{v}} \right) f = \left(\frac{\partial f}{\partial t} \right)_{\text{coll}}. \quad (19.7)$$

The instantaneous rate of change of $f(\mathbf{x}, \mathbf{v})$ due to collisions involves collisions which scatter particles out of this region of phase space and collisions which scatter particles in. We define the loss rate R such that $Rd^3x d^3v$ is the number of collisions per unit time where one of the *initial* particles is in $d^3x d^3v$ and the gain rate \bar{R} so that $\bar{R}d^3x d^3v$ is the number of collisions per unit time where one of the *final* particles is in $d^3x d^3v$ so the collision term can be written as

$$\left(\frac{\partial f}{\partial t} \right)_{\text{coll}} d^3x d^3v = N_{\text{gain}} - N_{\text{loss}} = (\bar{R} - R) d^3x d^3v \quad (19.8)$$

Let us consider exclusively 2-body collisions of equal mass atoms which scatter particles labeled 1, 2 from velocities $\mathbf{v}_1, \mathbf{v}_2$ to $\mathbf{v}'_1, \mathbf{v}'_2$ as illustrated in figure 19.1. To obtain the loss rate from particles with some specific velocity \mathbf{v}_1 we need to integrate over all possible velocities \mathbf{v}_2 for the other colliding particle. Once \mathbf{v}_1 and \mathbf{v}_2 are specified, this leaves the six variables $\mathbf{v}'_1, \mathbf{v}'_2$ to be determined. Momentum and energy conservation impose four constraints, leaving two variables to fully determine the collision. We can take these to be the direction $\Omega = \hat{\mathbf{v}}'_1$ of particle 1 after the collision. We define the *differential cross-section* $\sigma(\Omega)$ such that the number of collisions per unit time per unit spatial volume between particles in streams with space densities $n_1 = f(\mathbf{v}_1)d^3v_1$ and $n_2 = f(\mathbf{v}_2)d^3v_2$ and in which particle 1 is deflected into direction $d\Omega$ is

$$\frac{dN}{dt d^3x} = n_1 n_2 |\mathbf{v}_1 - \mathbf{v}_2| \sigma(\Omega) d\Omega = d^3v_1 d^3v_2 f(\mathbf{v}_1) f(\mathbf{v}_2) |\mathbf{v}_1 - \mathbf{v}_2| \sigma(\Omega) d\Omega. \quad (19.9)$$

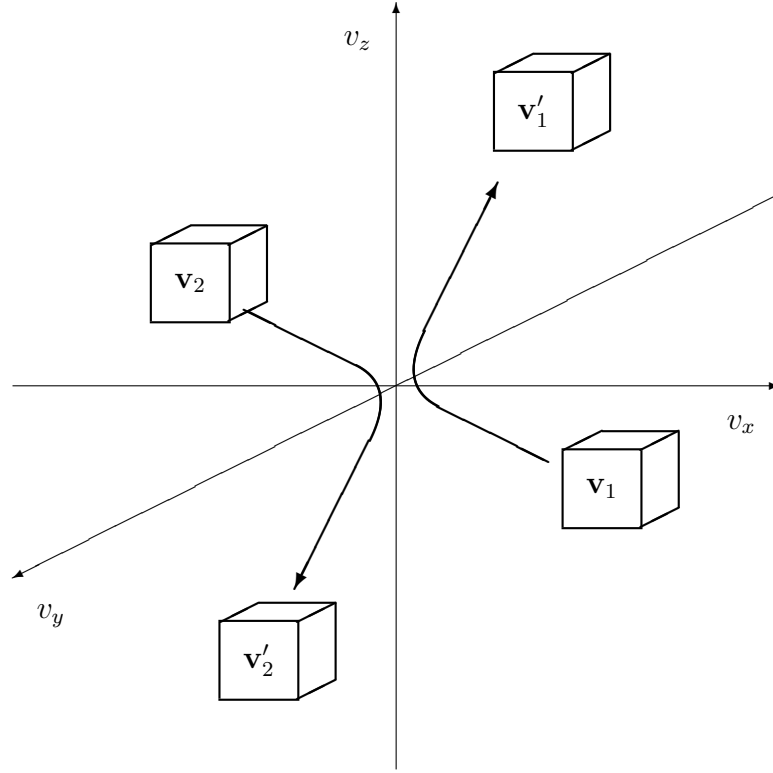


Figure 19.1: Illustration of a collision between two particles where initial particles are in momentum space cells labeled $\mathbf{v}_1, \mathbf{v}_2$ and end up in cells labeled $\mathbf{v}'_1, \mathbf{v}'_2$. In calculating the rate of change of occupation number for cell \mathbf{v}_1 say, we have a negative term corresponding to the ‘forward’ reactions as shown, but we also have a positive term arising from ‘inverse’ reactions. For interaction processes of most interest, the ‘microscopic rates’ for the forward and inverse reactions are the same, and the actual rates are then just proportional to $f_1 f_2$ and $f'_1 f'_2$ respectively.

Integrating this over all \mathbf{v}_2 and Ω gives the loss term

$$R d^3 v_1 = d^3 v_1 f(\mathbf{v}_1) \int d^3 v_2 \int d\Omega \sigma(\Omega) |\mathbf{v}_1 - \mathbf{v}_2| f(\mathbf{v}_2). \quad (19.10)$$

The gain term is trickier, as we need to consider the inverse collisions $\mathbf{v}'_1, \mathbf{v}'_2 \rightarrow \mathbf{v}_1, \mathbf{v}_2$ where one of the particles ends up in $d^3 v_1$ and integrate over all possible values of $\mathbf{v}'_1, \mathbf{v}'_2$ and \mathbf{v}_2 consistent with energy and momentum conservation. However, we know from field theory that the net rate of reactions scattering particles out of state \mathbf{v}_1 (ie including backward reactions as a negative rate) is

$$R(\mathbf{v}_1, \mathbf{v}_2 \rightarrow \mathbf{v}'_1, \mathbf{v}'_2) \propto d^3 v_1 d^3 v_2 d^3 v'_1 d^3 v'_2 \delta^{(4)}(\vec{p}_1 + \vec{p}_2 - \vec{p}'_1 - \vec{p}'_2) (n'_1 n'_2 - n_1 n_2) |T(\vec{p}_i)|^2. \quad (19.11)$$

The transition amplitude T is an invariant function of its arguments, and for electromagnetic interactions is symmetric under exchange of initial and final states. Thus for real reactions we expect the rates to be symmetric with respect to forward and backward collisions, and so we have for the total collision term

$$\left(\frac{\partial f_1}{\partial t} \right)_{\text{coll}} = \int d^3 v_2 \int d\Omega \sigma(\Omega) |\mathbf{v}_1 - \mathbf{v}_2| (f'_1 f'_2 - f_1 f_2) \quad (19.12)$$

where $f'_1 = f(\mathbf{v}'_1)$ etc and where in the integral on the right \mathbf{v}_1 is assumed to be fixed, and $\mathbf{v}'_1, \mathbf{v}'_2$ are considered functions of $\mathbf{v}_1, \mathbf{v}_2$ and Ω . Also, it should be noted that $\sigma(\Omega)$ is really also a function of the relative speed of the particles.

19.3 Applications of the Transport Equation

19.3.1 Equilibrium Solutions

In equilibrium, the collision term (19.12) must vanish, which implies

$$f'_1 f'_2 - f_1 f_2 = 0 \quad (19.13)$$

This can be thought of as a conservation law

$$\log f'_1 + \log f'_2 = \log f_1 + \log f_2 \quad (19.14)$$

Now we also have conservation of energy, or

$$(v'_1)^2 + (v'_2)^2 = v_1^2 + v_2^2 \quad (19.15)$$

The equilibrium phase space distribution function is some function only of the *speed* $f = f(|\mathbf{v}|)$, which is compatible with the above conservation laws if

$$\log f = a - bv^2. \quad (19.16)$$

The coefficient b can be related to the temperature using the Boltzmann formula. Since the kinetic energy is $E = mv^2/2$ and f measures the probability that a cell of phase-space is occupied we must have

$$\frac{f(\mathbf{v}_1)}{f(\mathbf{v}_2)} = \exp(-b(v_1^2 - v_2^2)) = \exp(-\beta(E_1 - E_2)) \quad (19.17)$$

so $b = \beta m/2 = m/(2kT)$.

The coefficient a is a normalization factor and is fixed once we specify the number density of particles

$$n = \int d^3v f \quad (19.18)$$

and performing the integration gives

$$f(\mathbf{v}) = \frac{n}{(2\pi kT/M)^{3/2}} e^{-\frac{1}{2}mv^2/kT}. \quad (19.19)$$

One can thereby show that the mean kinetic energy per particle

$$\epsilon = \frac{\int d^3v f(v) mv^2/2}{\int d^3v f} = \frac{3}{2}kT. \quad (19.20)$$

If we include the ‘quantum mechanical’ factors for the final states then the condition for equilibrium becomes

$$f'_1 f'_2 (1 \pm h^3 f_1)(1 \pm h^3 f_2) - f_1 f_2 (1 \pm h^3 f'_1)(1 \pm h^3 f'_2) = 0 \quad (19.21)$$

and one then obtains the Fermi-Dirac and Bose-Einstein distributions. One can similarly derive other thermodynamic properties such as the pressure and the entropy density s , the latter being given, aside from a constant, by

$$s = -\frac{1}{n} \int d^3v f(v) \log h^3 f(v) = -\langle \log h^3 f \rangle \quad (19.22)$$

since one can easily show that with the equilibrium distribution function this integral is

$$\int d^3v f(v) \log h^3 f(v) = \frac{3n}{2} \log(kT/n^{2/3}) + \text{constant} \quad (19.23)$$

19.3.2 Boltzmann's H -Theorem

Consider a gas of particles with uniform space density (so $\nabla_{\mathbf{x}}f = 0$) and with no external force acting (so $\dot{\mathbf{v}} \cdot \nabla_{\mathbf{v}}f = 0$ also), so the Liouville operator is $L = \partial/\partial t$ and Liouville's equation becomes

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t} \right)_{\text{coll}}. \quad (19.24)$$

Let us define Boltzmann's H -function

$$H(t) = \int d^3v f(\mathbf{v}, t) \log f(\mathbf{v}, t). \quad (19.25)$$

Where, for this to make sense, f must be understood to be the *numerical* value of f for some particular choice of units, rather than a dimensional quantity. Also, f must really be considered to be the average of the occupation number n over a large number of fundamental cells of size h^3 in phase-space, since the actual occupation numbers are mostly zero in the dilute gas approximation.

The time derivative of H is

$$\frac{dH}{dt} = \int d^3v_1 \frac{\partial f_1}{\partial t} (1 + \log f_1) \quad (19.26)$$

or, using (19.12),

$$\frac{dH}{dt} = \int d^3v_1 \int d^3v_2 \int d\Omega \sigma(\Omega) |\mathbf{v}_1 - \mathbf{v}_2| (f'_1 f'_2 - f_1 f_2) (1 + \log f_1). \quad (19.27)$$

Now this 8-dimensional integral is really just a way of enumerating all the possible reaction paths, and could have been written as

$$\frac{dH}{dt} \propto \int d^3v_1 \int d^3v_2 \int d^3v'_1 \int d^3v'_2 \delta^{(4)}(\vec{p}_1 + \vec{p}_2 - \vec{p}'_1 - \vec{p}'_2) (f'_1 f'_2 - f_1 f_2) |T(\vec{p}_i)|^2 (1 + \log f_1) \quad (19.28)$$

which, aside from the last factor, is symmetric under exchanging particle labels, except for a factor -1 if we switch final and initial state labels. This gives us four equivalent statements for dH/dt which we can average to obtain

$$\frac{dH}{dt} = \frac{1}{4} \int d^3v_1 d^3v_2 d\Omega \sigma'(\Omega) |\mathbf{v}'_1 - \mathbf{v}'_2| (f'_1 f'_2 - f_1 f_2) (\log f_1 f_2 - \log f'_1 f'_2). \quad (19.29)$$

Now since \log is a monotonically increasing function of its argument the last two factors here have opposite sign and so we obtain the H -theorem

$$\frac{dH}{dt} \leq 0. \quad (19.30)$$

Since, as we have seen, the quantity $H(t)$ is proportional to (minus) the entropy, we see that the statistical mechanical entropy can never decrease. Also, the condition that the entropy should be constant is the same as the condition for equilibrium. Therefore, if $f(\mathbf{v})$ differs from the Maxwellian form, the entropy must necessarily increase until the phase-space density becomes Maxwellian.

We can perhaps see this more clearly if we consider a single reaction path $\mathbf{v}_1, \mathbf{v}_2 \leftrightarrow \mathbf{v}'_1, \mathbf{v}'_2$ as in figure 19.1. If the phase-space density for cell v_1 changes by an amount Δf_1 then the contribution to the H -function from that cell changes by an amount $\Delta H_1 = d^3v_1 \Delta f_1 \log f_1$. Now $d^3v f$ is the space number density of those particles falling in this cell, so if we take the total volume of space to be unity, for a single forward interaction we have $d^3v_1 f_1 = d^3v_2 f_2 = -1$ and $d^3v_1 f'_1 = d^3v_2 f'_2 = +1$ so the total change in H for a single forward reaction is

$$\Delta H = -\log f_1 - \log f_2 + \log f'_1 + \log f'_2 = \log f'_1 f'_2 - \log f_1 f_2. \quad (19.31)$$

However, the mean net rate of forward reactions (i.e. the mean number of forward reactions minus the mean number of backwards reactions) is proportional to $f_1 f_2 - f'_1 f'_2$ so we reach the conclusion that, on average, collisions will cause H to decrease (assuming the gas is not in equilibrium). This is true regardless of which reaction pathway we consider so dH/dt must be non positive.

19.4 Conserved Quantities

If some quantity $\chi(\mathbf{v})$ is conserved in collisions, ie

$$\chi(\mathbf{v}_1) + \chi(\mathbf{v}_2) = \chi(\mathbf{v}'_1) + \chi(\mathbf{v}'_2), \quad (19.32)$$

then

$$\int d^3v \chi(\mathbf{v}) \left(\frac{\partial f}{\partial t} \right)_{\text{coll}} = 0. \quad (19.33)$$

The proof is similar to that of the H -theorem. Using (19.12) we can write this as

$$\int d^3v \chi(\mathbf{v}) \left(\frac{\partial f}{\partial t} \right)_{\text{coll}} = \int d^3v_1 d^3v_2 d\Omega \sigma(\Omega) |\mathbf{v}_1 - \mathbf{v}_2| \chi(\mathbf{v}_1) (f'_1 f'_2 - f_1 f_2) \quad (19.34)$$

and interchanging $1 \leftrightarrow 1'$ etc as before we can write this as

$$\int d^3v \chi(\mathbf{v}) \left(\frac{\partial f}{\partial t} \right)_{\text{coll}} = \frac{1}{4} \int d^3v_1 d^3v_2 d\Omega \sigma(\Omega) |\mathbf{v}_1 - \mathbf{v}_2| (\chi_1 + \chi_2 - \chi'_1 - \chi'_2) (f'_1 f'_2 - f_1 f_2) \quad (19.35)$$

which vanishes by (19.32).

Using $Lf = (\partial f / \partial t)_{\text{coll}}$ this can be written as

$$\int d^3v \chi(\mathbf{v}) \left(\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} - \frac{\partial \Phi}{\partial \mathbf{x}} \cdot \frac{\partial f}{\partial \mathbf{v}} \right) = 0 \quad (19.36)$$

where we have set $\dot{\mathbf{v}} = -\partial \Phi / \partial \mathbf{x}$, with Φ the gravitational potential, or as

$$\frac{\partial}{\partial t} \int d^3v \chi f + \frac{\partial}{\partial \mathbf{x}} \int d^3v \chi \mathbf{v} f - \frac{\partial \Phi}{\partial \mathbf{x}} \cdot \int d^3v \left(\frac{\partial(f\chi)}{\partial \mathbf{v}} - f \frac{d\chi}{d\mathbf{v}} \right) = 0 \quad (19.37)$$

where we have used $\chi \partial f / \partial \mathbf{v} = \partial(f\chi) / \partial \mathbf{v} - f d\chi / d\mathbf{v}$. Now the first term in the last integral in (19.37) vanishes since $f(v) \rightarrow 0$ as $v \rightarrow \infty$. Defining the velocity average of a function Q as

$$\langle Q \rangle = \frac{\int d^3v f Q}{\int d^3v f} = \frac{1}{n} \int d^3v f Q \quad (19.38)$$

(19.37) becomes

$$\frac{\partial n \langle \chi \rangle}{\partial t} + \frac{\partial n \langle \chi \mathbf{v} \rangle}{\partial \mathbf{x}} + n \frac{\partial \Phi}{\partial \mathbf{x}} \cdot \left\langle \frac{d\chi}{d\mathbf{v}} \right\rangle = 0. \quad (19.39)$$

There are 5 dynamical quantities which are conserved in collisions: The mass of the particles m , the three components of the momentum $m\mathbf{v}$, and the energy $mv^2/2$. Substituting these in turn for χ in (19.39) yields a set of 5 useful conservation laws.

19.4.1 Mass Conservation

If we set $\chi = m$ in (19.39) and define the mass density $\rho \equiv nm$ then the last term vanishes (since m is independent of \mathbf{v}) and we obtain the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad (19.40)$$

where we have defined the mean velocity (or ‘streaming velocity’) as

$$\mathbf{u} = \langle \mathbf{v} \rangle = \frac{\int d^3v f \mathbf{v}}{\int d^3v f}. \quad (19.41)$$

Since $\nabla \cdot (\rho \mathbf{u}) = \rho \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla \rho$ we can write (19.40) as

$$\frac{D\rho}{Dt} = -\rho \nabla \cdot \mathbf{u} \quad (19.42)$$

where we have defined the *Lagrangian derivative* $D/Dt \equiv \partial/\partial t + \mathbf{u} \cdot \nabla$ so DQ/Dt is the rate of change of Q as seen by an observer moving at the streaming velocity.

19.4.2 Momentum Conservation

Setting $\chi(\mathbf{v}) = m\mathbf{v}$ in (19.39) yields

$$\frac{\partial \rho u_i}{\partial t} + \frac{\partial \rho \langle v_i v_k \rangle}{\partial x_k} + \rho \frac{\partial \Phi}{\partial x_i} = 0. \quad (19.43)$$

If we let $\mathbf{v} = \mathbf{u} + \mathbf{w}$, so \mathbf{w} is the ‘random’ motion of the particle relative to the local mean velocity \mathbf{u} , and define $\rho \langle w_i w_k \rangle = P\delta_{ik} - \pi_{ik}$ where

$$P \equiv \rho \langle |\mathbf{w}|^2 \rangle / 3 \quad (19.44)$$

is the ‘gas pressure’ and

$$\pi_{ik} \equiv \rho \langle |\mathbf{w}|^2 \delta_{ik} / 3 - w_i w_k \rangle \quad (19.45)$$

is the ‘viscous stress-tensor’. What we are doing here is decomposing the *velocity dispersion tensor* $\langle w_i w_k \rangle$ into a diagonal part describing isotropic kinetic pressure and the traceless part π_{ik} . Note that π_{ik} is symmetric and therefore has 5 degrees of freedom.

The momentum conservation law (19.43) then becomes

$$\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_k}(\rho u_i u_k + P\delta_{ik} - \pi_{ik}) = -\rho \frac{\partial \Phi}{\partial x_i} \quad (19.46)$$

and if we combine this with (19.40) one obtains the ‘force equation’

$$\rho \frac{D\mathbf{u}}{Dt} = -\rho \nabla \Phi - \nabla P + \nabla \cdot \boldsymbol{\pi} \quad (19.47)$$

where $(\nabla \cdot \boldsymbol{\pi})_i \equiv \partial \pi_{ik} / \partial x_k$.

19.4.3 Energy Conservation

Setting

$$\chi(\mathbf{v}) = \frac{1}{2}mv^2 = \frac{1}{2}mu^2 + m\mathbf{w} \cdot \mathbf{u} + \frac{1}{2}mw^2 \quad (19.48)$$

in (19.39) yields the energy conservation law

$$\frac{\partial}{\partial t} \left[\frac{\rho}{2} (u^2 + \langle w^2 \rangle) \right] + \frac{\partial}{\partial x_k} \left[\frac{\rho}{2} \langle (u_k + w_k) |\mathbf{u} + \mathbf{w}|^2 \rangle \right] + \rho \frac{\partial \Phi}{\partial x_k} u_k = 0. \quad (19.49)$$

Now

$$\langle (u_k + w_k)(u_i + w_i)(u_i + w_i) \rangle = u^2 u_k + 2u_i \langle w_i w_k \rangle + u_k \langle w^2 \rangle + \langle w_k w^2 \rangle \quad (19.50)$$

and defining the *specific internal energy* ϵ by

$$\rho \epsilon = \rho \langle w^2 / 2 \rangle = 3P/2 \quad (19.51)$$

and the *conduction heat flux* \mathbf{F}

$$F_k = \frac{1}{2} \rho \langle w_k w^2 \rangle \quad (19.52)$$

(19.49) gives the *total energy equation*

$$\frac{\partial}{\partial t} \left(\frac{1}{2} \rho u^2 + \rho \epsilon \right) + \frac{\partial}{\partial x_k} \left(\frac{1}{2} \rho u^2 u_k + u_i (P\delta_{ik} - \pi_{ik}) + \rho \epsilon u_k + F_k \right) = -\rho u_k \frac{\partial \Phi}{\partial x_k} \quad (19.53)$$

which is a continuity equation where the first term is the time derivative of the total energy density, the second term is the divergence of the energy flux and the term on the right hand side is the rate at which external forces are doing work.

Using (19.47) this can also be recast as the *internal energy equation*

$$\frac{\partial(\rho\epsilon)}{\partial t} + \frac{\partial(\rho\epsilon u_k)}{\partial x_k} = -P \frac{\partial u_k}{\partial x_k} - \frac{\partial F_k}{\partial x_k} + \Psi \quad (19.54)$$

where we have defined the *rate of viscous dissipation* as

$$\Psi = \pi_{ik} \frac{\partial u_i}{\partial x_k} = \frac{1}{2} \pi_{ik} \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} \right). \quad (19.55)$$

Finally, using the equation of continuity this becomes an expression of the first law of thermodynamics

$$\rho \frac{D\epsilon}{Dt} = P \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{F} + \Psi \quad (19.56)$$

where now the external potential Φ and the bulk motion \mathbf{u} no longer appear. This says that the rate of change of internal energy of a parcel of fluid comes from a combination of PdV work, heat conduction and dissipation of anisotropic shear.

19.5 Fluid Equations

To recapitulate, taking moments of the Boltzmann transport equation have yielded

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \right) \rho &= -\rho \nabla \cdot \mathbf{u} \\ \rho \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \right) \mathbf{u} &= -\rho \nabla \Phi - \nabla P + \nabla \cdot \boldsymbol{\pi} \\ \rho \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \right) \epsilon &= -P \nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{F} + \Psi \end{aligned} \quad (19.57)$$

with constituent relations

$$\begin{aligned} \rho &= m \int d^3v f(\mathbf{x}, \mathbf{v}, t) \\ \mathbf{u} &= \langle \mathbf{v} \rangle \\ \epsilon &= \frac{3}{2} P / \rho = \frac{1}{2} \langle w^2 \rangle \\ \pi_{ik} &\equiv \rho \langle |\mathbf{w}|^2 \delta_{ik} / 3 - w_i w_k \rangle \\ F_k &= \frac{1}{2} \rho \langle w_k w^2 \rangle \\ \Psi &= \pi_{ik} \frac{\partial u_i}{\partial x_k} \end{aligned} \quad (19.58)$$

this gives 5 equations, but unfortunately there are 13 unknowns these are ρ , ϵ , and the three components of \mathbf{u} plus the three components of the heat flux \mathbf{F} and the 5 components of the viscous stress tensor π_{ij} (which is a trace-free symmetric tensor).

However, in the limit that the mean free path λ is small compared to the scales L over which conditions are varying progress can be made by making an expansion in powers of λ/L . In the limit $\lambda \rightarrow 0$ collisions will be very effective and will force the velocity distribution to become locally Maxwellian, on a short time-scale on the order of the collision time. Inhomogeneities on large scales though will take much longer to damp out. Now for a Maxwellian, the random velocity distribution is isotropic, so the viscous stress and the heat flux vanish. If we simply set $\mathbf{F} = 0$ and $\pi_{ik} = 0$ then one ends up with a set of 5 equations for the 5 unknowns ρ , ϵ and \mathbf{u} . These are known as the ideal fluid equations, and the momentum equation is known as the *Euler equation*. Taking the ideal fluid as a zeroth order solution, one can then compute the small anisotropy of the random velocity distribution function to first order in the mean free path. For instance, if there is velocity shear, then this will drive an anisotropy of the random velocity dispersion $\langle \omega_i \omega_j \rangle$ resulting in a viscous stress $\pi \sim \lambda \nabla \mathbf{u} / \langle w^2 \rangle^{1/2}$. Similarly, if there is a temperature gradient in the zeroth order solution then this will give rise to a heat flux \mathbf{F} etc. Including the effects of viscosity and heat conduction result in the *Navier-Stokes equations*.

19.6 Problems

19.6.1 Boltzmann distribution

According to elementary field theory, and in the ‘dilute gas’ approximation, the net rate for reactions $p_{i1} \dots p_{iN} \leftrightarrow p_{f1} \dots p_{fM}$ with N incoming particles and M outgoing particles (or *vice versa* for the

inverse reaction) contains a factor

$$(f_{i1} \dots f_{iN}(1 \pm f_{f1}) \dots (1 \pm f_{fM}) - f_{f1} \dots f_{fM}(1 \pm f_{i1}) \dots (1 \pm f_{iN})) \quad (19.59)$$

where \pm signs apply for outgoing bosons and fermions respectively. This expression says that the rate of reactions depends on the product of the densities of incoming particles, as expected classically, and the $1 \pm f$ factors associated with the outgoing particles express quantum mechanical corrections for stimulated emission (for bosons) and the exclusion principle for fermions.

- a. For a gas of particles which have isotropically distributed momenta in the lab frame the occupation number f is a function only of the energy of the particle. Show that the equilibrium distribution function for the species i is $f_i(E_i) = (e^{(\beta E_i + \alpha_i)} \mp 1)^{-1}$ where β is the inverse temperature and α_i is chemical potential for species i and the sum of the chemical potentials for the incoming particles is equal to the sum for the outgoing particles.
- b. Show that for a reaction $L_1, L_2 \leftrightarrow L'_1, L'_2, B$ (i.e. two L -particles scattering off each other, creating in the process a bosonic B particle) that $\alpha_B = 0$ and that the bosons must therefore have a Planckian distribution.

19.6.2 Kinetic theory and entropy

Write down expressions for the energy U and pressure P of an ideal monatomic gas of N particles at temperature T in a volume V . Use these relations and the first law of thermodynamics $dU = TdS - PdV$ to show that the entropy is given by $S = Nk(3/2 \ln T + \ln V) + \text{constant}$.

Now consider the statistical mechanical definition of the entropy density $s = -k \int d^3v f \ln f$ where $f(\mathbf{v})$ is the phase space distribution function. For a Maxwellian distribution $f = A \exp -mv^2/2kT$, compute the normalisation factor A in terms of the temperature T and the space density of particles n , and compute the total statistical mechanical entropy $S = \int d^3x s$ and compare with the thermodynamic result.

The collision term in the Boltzmann transport equation is

$$\left(\frac{\partial f(\mathbf{v}_1)}{\partial t} \right)_c = \int d^3v_2 \int d\Omega \frac{d\sigma}{d\Omega} |\mathbf{v}_1 - \mathbf{v}_2| (f(\mathbf{v}'_1)f(\mathbf{v}'_2) - f(\mathbf{v}_1)f(\mathbf{v}_2)) \quad (19.60)$$

Show that for a Maxwellian, $(\partial f / \partial t)_c = 0$.

19.6.3 Kinetic theory

Compare the de Broglie wavelength and the mean separation of air molecules at atmospheric pressure; discuss the validity of a description of such a gas as a collection of effectively distinguishable particles following classical trajectories interrupted by brief collision events.

Estimate the thermal speed for air molecules at room temperature. Estimate the mean free path and collision time assuming a collision cross section $\sigma \sim 10^{-16} \text{cm}^2$.

Estimate the thermal velocities of electrons and ions in ionised gas at $T \sim 10^4 K$. Estimate the mean free time between hard electron-electron collisions (i.e. collisions which change the direction of the electron's motion substantially) and the corresponding mean free path assuming a density of $n_e \sim 1 \text{cm}^{-3}$.

19.6.4 Massive neutrinos

According to the standard big bang model the Universe is suffused with a thermal gas of neutrinos which, at high redshift has a number density and temperature essentially the same as the photons which now comprise the 3K microwave background. These neutrinos, if they had a mass on the order of ten electron volts, could provide the 'missing mass' inferred from dynamics of clusters of galaxies etc.

- a) Estimate the redshift at which these neutrinos became non-relativistic (i.e. the epoch when a MBR photon had a typical energy ~ 10 eV) and estimate their phase space density at that time.
- b) What does the collisionless Boltzmann equation tell us about the subsequent evolution of their phase space density?
- c) Use this to derive, to order of magnitude, the ‘Tremaine-Gunn’ bound in the size - velocity dispersion plane for structures which are gravitationally bound by these particles.
- d) Compare this limit with cores of galaxy clusters ($\sigma \sim 10^3$ km/s and $R \sim 0.25$ Mpc).

Chapter 20

Ideal Fluids

Setting $\mathbf{F} = 0$ and $\pi_{ik} = 0$ in (19.57) yields the *ideal fluid equations*

$$\begin{aligned} \frac{D\rho}{Dt} &= -\rho \nabla \cdot \mathbf{u} \\ \rho \frac{D\mathbf{u}}{Dt} &= -\rho \nabla \Phi - \nabla P \\ \frac{D\epsilon}{Dt} &= -\frac{2}{3} \epsilon \nabla \cdot \mathbf{u} \end{aligned} \quad (20.1)$$

These are known as the *continuity equation*, *Euler's equation*, and the *energy equation*.

20.1 Adiabatic Flows

If one multiplies the energy equation in (20.1) by $3\rho/2\epsilon$ and subtracts the continuity equation one obtains

$$\frac{3}{2} \rho \frac{D\epsilon}{Dt} - \frac{D\rho}{Dt} = 0 \quad (20.2)$$

or equivalently

$$\frac{D}{Dt}(\rho \epsilon^{-3/2}) = 0 \quad (20.3)$$

which tells us that along a *streamline* the specific energy and the mass density are related by

$$\epsilon \propto \rho^{2/3} \quad (20.4)$$

This makes physical sense. Each particle is bouncing off other particles. If there is a net contraction, so $\nabla \cdot \mathbf{u} < 0$ in some region, then a particle will tend to gain energy as it bounces off particles moving inwards.

This is much like what happens to a particle in a box with reflecting walls which are changing with time so that $L = L(t)$. In each reflection off a wall perpendicular to the x -axis the x -component of the velocity changes by $\Delta v_x = -\dot{L}$ and such reflections occur once per time $\Delta t = L/v_x$, and so the rate of change of the velocity is

$$\frac{dv}{dt} = \frac{\Delta v}{\Delta t} = \frac{v \dot{L}}{L} \quad \rightarrow \quad \frac{dv}{v} = -\frac{dL}{L} \quad (20.5)$$

with solution $v(t) \propto 1/L$. Since the density of particles scales as $\rho \propto 1/L^3$ the thermal energy scales as $\epsilon \propto v^2 \propto 1/L^2 \propto \rho^{2/3}$ in accord with (20.4).

Another line of argument that leads to the same result is to consider standing de Broglie waves in a cavity.

Equation (20.4) is equivalent to constancy of the entropy. The first law of thermodynamics is $dU = TdS - PdV$, so for constant S , $dU = -PdV$, but $U = M\epsilon$, $V = M/\rho$ where M is the mass of gas, so $Md\epsilon = MPd\rho^{-1}$ which, with $P = 2/3\epsilon\rho$, gives $d\epsilon/\epsilon = (2/3)d\rho/\rho$ which again implies $\epsilon \propto \rho^{2/3}$.

20.2 Hydrostatic Equilibrium

Setting $D\mathbf{u}/Dt = 0$ in Euler's equation gives the *equation of hydrostatic equilibrium*

$$\nabla P = -\rho \nabla \Phi \quad (20.6)$$

with Φ the gravitational potential.

For a stratified medium this becomes

$$\frac{\partial P}{\partial z} = \rho g \quad (20.7)$$

with $\mathbf{g} = -\nabla \Phi$ the gravity.

More generally the density and potential are related by *Poisson's equation*

$$\nabla^2 \Phi = 4\pi G \rho \quad (20.8)$$

and combining this with (20.6) gives

$$\nabla \cdot \left(\frac{1}{\rho} \nabla P \right) = -\nabla^2 \Phi = -4\pi G \rho \quad (20.9)$$

or, for a spherically symmetric system

$$\frac{1}{r^2} \frac{d}{dr} \left(\frac{r^2}{\rho} \frac{dP}{dr} \right) = -4\pi G \rho. \quad (20.10)$$

With an assumed *equation of state* $P = P(\rho)$ for instance, (20.10) can be integrated to give $P(r)$, $\rho(r)$ etc.

20.3 Convective Stability

Solutions of (20.10) are *mechanically stable* but may be *convectively unstable*. To determine whether a solution $\rho(z)$, $P(z)$ is convectively stable, consider a blob of gas, and imagine displacing it upwards by an amount Δz . After rising, the pressure falls by an amount $\Delta P = -\Delta z \nabla P$ and it will expand (adiabatically) with $P \propto \rho^{5/3}$ and the fractional change in density will be $\Delta \rho / \rho = (3/5) \Delta P / P$, whereas the fractional change in the ambient density is $\Delta \rho / \rho = \Delta z \nabla \rho / \rho$. If, after rising, it finds itself more buoyant than its surroundings (ie lower density) then it will continue to rise and is clearly unstable. It is not difficult to see that the condition for stability is that the entropy of the gas should increase with height, or equivalently $d \log P / d \log \rho < 5/3$.

20.4 Bernoulli's Equation

Bernoulli's Equation applies to *steady flows* for which $\partial \mathbf{u} / \partial t = 0$ (this should not be confused with hydrostatic equilibrium, for which $D\mathbf{u}/Dt = 0$). Euler's equation then becomes

$$(\mathbf{u} \cdot \nabla) \mathbf{u} = -\nabla \Phi - \nabla P / \rho \quad (20.11)$$

but $(\mathbf{u} \cdot \nabla) \mathbf{u} = \frac{1}{2} \nabla u^2 - \mathbf{u} \times (\nabla \times \mathbf{u})$ so this is

$$\frac{1}{2} \nabla u^2 + (\nabla \times \mathbf{u}) \times \mathbf{u} + \nabla \Phi + \nabla P / \rho = 0. \quad (20.12)$$

Now under various conditions, the last term here is a total gradient. One such situation is if the flow is adiabatic (ie isentropic) in which case $\nabla P / \rho$ is the gradient of the specific enthalpy. The *total enthalpy* H is defined to be $H = U + PV$, and its derivative is $dH = dU + PdV + VdP = TdS + VdP$, and dividing by the mass gives $dh = Tds + dP/\rho$. Thus, for adiabatic gas, $\nabla P / \rho = \nabla h$, where

$$h = \epsilon + P/\rho \quad (20.13)$$

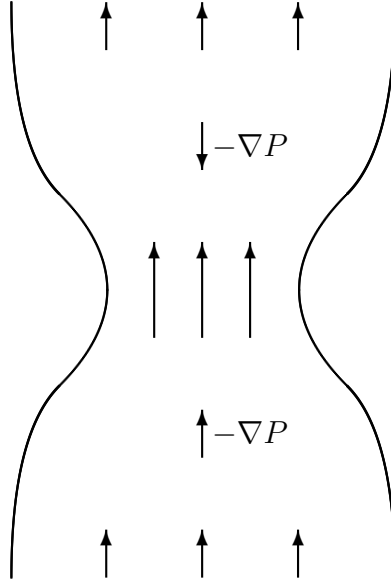


Figure 20.1: Illustration of the Bernoulli effect. An ideal fluid moves through a venturi tube. Constancy of flux means the velocity must be highest at the waist of the tube. In a steady state this requires that the fluid be accelerated, which requires a pressure gradient so that the pressure be lowest where the velocity is highest.

is the *specific enthalpy*. The quantity $\nabla P/\rho$ is also a total gradient for any *barytropic* equation of state $P = P(\rho)$, and we shall assume this is the case.

If we dot equation (20.12) above with the unit vector $\hat{\mathbf{u}}$ the term involving $\nabla \times \mathbf{u}$ drops out, and noting that $\hat{\mathbf{u}} \cdot \nabla$ is the derivative along the direction of motion, which we will write as d/dl , we have

$$\frac{d}{dl} \left(\frac{1}{2} u^2 + h + \Phi \right) = 0. \quad (20.14)$$

Now the specific enthalpy is just proportional to the pressure, which means that, if we neglect the effect of gravity, an increase in fluid or gas velocity must be accompanied by a drop of pressure. This is called the *Bernoulli effect* and is used in the carburetor where the air is channelled through a ‘venturi’ or constricting nozzle. Since the flux of gas is fixed, this requires that the gas velocity scale inversely as the cross-sectional area, resulting in a drop in pressure in the nozzle. This pressure drop causes fuel to be sucked through the jet into the air-stream. The phenomenon is also said to occur when ships travel parallel to one another. Here the water is forced to flow faster between the ships, causing a drop in pressure which causes the ships to be attracted to one another. This also explains why if one opens a car window a crack, air will be sucked out. Finally, it can be used effectively to separate pages of a book or newspaper simply by blowing past the leaves to cause a drop in pressure.

20.5 Kelvin’s Circulation Theorem

We can write the Euler equation as

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \left(\frac{1}{2} u^2 \right) + (\nabla \times \mathbf{u}) \times \mathbf{u} = -\nabla \Phi - \nabla P/\rho. \quad (20.15)$$

If we take the curl of this, and assume adiabaticity, so $\nabla P/\rho = \nabla h$, or more generally assume a *barytropic equation of state* $P = P(\rho)$, then all but the first and third terms on the left hand side drop out. Defining the *vorticity* $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ we obtain

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \nabla \times (\boldsymbol{\omega} \times \mathbf{u}) = 0. \quad (20.16)$$

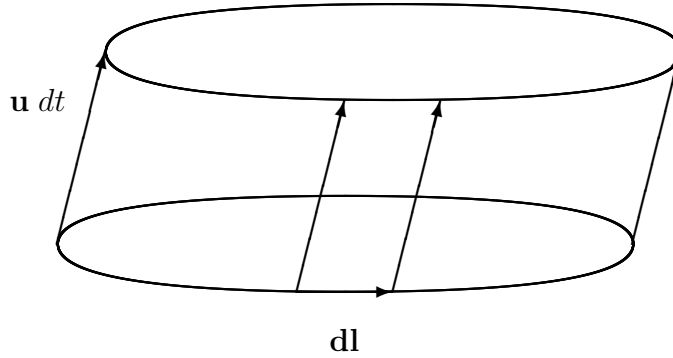


Figure 20.2: Kelvin's circulation theorem concerns the rate of change of Γ , the integral of the vorticity over a surface, as the boundary of the surface moves with the fluid flow. This figure illustrates that the change in the surface as it moves a distance $\mathbf{u} dt$ can be considered to be the ribbon-like strip connecting the old loop to the new loop.

Now consider the *circulation*

$$\Gamma \equiv \oint \mathbf{u} \cdot d\mathbf{l} \quad (20.17)$$

which is defined for a closed loop. Using Stokes' theorem this can be expressed as

$$\Gamma = \int dA (\nabla \times \mathbf{u}) \cdot \mathbf{n} = \int dA \boldsymbol{\omega} \cdot \mathbf{n} = \int d\mathbf{A} \cdot \boldsymbol{\omega} \quad (20.18)$$

How does Γ change with time for a loop which moves with the fluid? The value of Γ for a given loop does not depend on the surface that one chooses. The simplest way to compute the change in Γ as we move the loop a small distance $\mathbf{u}dt$ along the flow is to take the new surface to be the old surface plus a ribbon like wall which connects the old loop to the new loop as illustrated in figure 20.2. The change in Γ will then consist of two terms; the change in Γ for the old surface if the velocity field has changed with time plus the contribution from the ribbon:

$$\delta\Gamma = \delta t \int d\mathbf{A} \cdot \frac{\partial \boldsymbol{\omega}}{\partial t} + \int \delta \mathbf{A} \cdot \boldsymbol{\omega}. \quad (20.19)$$

The element of surface of the ribbon shown in figure 20.2 is

$$\delta \mathbf{A} = d\mathbf{l} \times \mathbf{u} \delta t \quad (20.20)$$

so in the second integral $\delta \mathbf{A} \cdot \boldsymbol{\omega} = \delta t (d\mathbf{l} \times \mathbf{u}) \cdot \boldsymbol{\omega} = \delta t d\mathbf{l} \cdot (\boldsymbol{\omega} \times \mathbf{u})$ (since $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ is the volume of the parallelepiped with edges \mathbf{a} , \mathbf{b} , \mathbf{c} , which is independent of the order of the vectors) and so dividing through by δt gives

$$\frac{d\Gamma}{dt} = \int d\mathbf{A} \cdot \frac{\partial \boldsymbol{\omega}}{\partial t} + \oint d\mathbf{l} \cdot (\boldsymbol{\omega} \times \mathbf{u}). \quad (20.21)$$

Using Stokes' theorem to convert the loop integral to a surface integral gives

$$\frac{d\Gamma}{dt} = \int d\mathbf{A} \cdot \left(\frac{\partial \boldsymbol{\omega}}{\partial t} + \nabla \times (\boldsymbol{\omega} \times \mathbf{u}) \right) \quad (20.22)$$

but by (20.16) the integrand vanishes and we have *Kelvin's circulation theorem*

$$\frac{d\Gamma}{dt} = 0 \quad (20.23)$$

which tells us that the circulation is conserved.

20.6 Potential Flows

Consider a stream of ideal fluid and assume that the vorticity at some point \mathbf{r} upstream vanishes. Now according to (20.16), the convective derivative of the vorticity is

$$\frac{D\boldsymbol{\omega}}{Dt} = \frac{\partial\boldsymbol{\omega}}{\partial t} + (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} = \nabla \times (\boldsymbol{\omega} \times \mathbf{u}) + (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} \quad (20.24)$$

which implies that if $\boldsymbol{\omega} = 0$ initially (ie at \mathbf{r}) then it will stay that way and must vanish at all points on the streamline passing through \mathbf{r} .

One can reach the same conclusion using Kelvin's circulation theorem for loops of various orientation in the vicinity of \mathbf{r} (which have $\Gamma = 0$) and following them downstream.

If the vorticity vanishes at *all* points upstream (as is the case if one has laminar flow, for instance) then the vorticity must vanish everywhere. If $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ this means that the flow velocity must be the gradient of some scalar function Ψ , known as the *velocity potential*:

$$\mathbf{u}(\mathbf{r}, t) = \nabla \Psi(\mathbf{r}, t). \quad (20.25)$$

A general flow has three degrees of freedom at each point in space, whereas a potential flow has only one since it is derived from a single scalar function Ψ . One way to see this is to write the flow as a Fourier synthesis

$$\mathbf{u}(\mathbf{r}, t) = \sum_{\mathbf{k}} \tilde{\mathbf{u}}_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{r}}. \quad (20.26)$$

For a general flow we need to specify three values $\tilde{\mathbf{u}}_{\mathbf{k}}$ for each \mathbf{k} whereas for a potential flow $\tilde{\mathbf{u}}_{\mathbf{k}} = i\mathbf{k}\tilde{\Psi}_{\mathbf{k}}$.

Now, neglecting gravity and assuming adiabaticity (so $\nabla P/\rho = \nabla h$) Euler's equation is

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{1}{2} \nabla u^2 + \mathbf{u} \times \boldsymbol{\omega} = -\nabla h \quad (20.27)$$

but for a potential flow, $\boldsymbol{\omega} = 0$, and $\partial \mathbf{u} / \partial t = \partial(\nabla \Psi) / \partial t$, and so we have

$$\nabla \left(\frac{\partial \Psi}{\partial t} + \frac{1}{2} u^2 + h \right) = 0 \quad (20.28)$$

implying

$$\frac{\partial \Psi}{\partial t} + \frac{1}{2} u^2 + h = f(t). \quad (20.29)$$

In fact, we can always set $f = 0$ without loss of generality since one can always make the transformation $\Psi \rightarrow \Psi' = \Psi + F(t)$ for arbitrary $F(t)$ without changing the flow $\mathbf{u} = \nabla \Psi$ and one can therefore choose $F(t)$ such that $\partial F / \partial t = f$.

For a steady flow $\partial \Psi / \partial t = 0$ and $f(t) = \text{constant}$ and we have

$$\frac{1}{2} u^2 + h = \text{constant}. \quad (20.30)$$

It is interesting to compare this with Bernoulli's equation (20.14) which looks very similar. However, the latter states that $u^2/2 + h$ is constant *along streamlines*, and allows the possibility that the constant is different for different streamlines. Equation (20.30), in contrast, is more restrictive and states that for a steady *potential* flow $u^2/2 + h$ is the same for all streamlines.

20.7 Incompressible Potential Flows

Potential flows have $\nabla \times \mathbf{u} = 0$, implying $\mathbf{u} = \nabla \Psi$. If in addition we require that the flow be *incompressible*, or approximately so, then the continuity equation $D\rho/Dt = -\rho \nabla \cdot \mathbf{u}$ implies that $\nabla \cdot \mathbf{u} = 0$ also, and therefore

$$\nabla^2 \Psi = 0. \quad (20.31)$$

Thus the velocity potential is the solution of Poisson's equation with vanishing source term.

An example is $\Psi = -1/|\mathbf{r}|$ for $\mathbf{r} \neq 0$, which gives a velocity $\mathbf{u} = \hat{\mathbf{r}}/r^2$ which is the $1/r^2$ mass conserving flow one would expect to find for a constant source of fluid injected at $\mathbf{r} = 0$.

In this context the Euler equation (now including the gravitational potential Φ) is

$$\nabla \frac{\partial \Psi}{\partial t} + \frac{1}{2} \nabla u^2 + \nabla \Phi + \nabla(P/\rho) = 0 \quad (20.32)$$

or

$$\frac{\partial \Psi}{\partial t} + \frac{u^2}{2} + \Phi + P/\rho = 0 \quad (20.33)$$

where we have assumed an appropriate $F(t)$ in Φ to set the right hand side to zero.

If the external pressure on some boundary is specified then (20.33) (with $u^2 = (\nabla \Psi)^2$) supplies the boundary condition for (20.31).

Note that Bernoulli's equation for incompressible potential flows is

$$\frac{1}{2} u^2 + \Phi + P/\rho = 0 \quad (20.34)$$

or, neglecting gravity,

$$\frac{1}{2} u^2 + P/\rho = 0 \quad (20.35)$$

which provides a direct relationship between flow velocity and pressure. If we consider an obstruction in a potential flow this implies that the pressure will be the greatest at the *stagnation points* on the surface of the obstruction where $\mathbf{u} = 0$. This also allows one to see how an aerofoil can provide lift if it is shaped such that the fluid must flow faster over the upper surface than over the lower surface.

20.8 Gravity Waves

An interesting application of potential flow theory is provided by *gravity waves* such as occur in the ocean or in atmospheres of stars and planets.

For concreteness, let us consider waves in a bath of incompressible fluid with negligible external pressure.

One can easily derive the main features of waves in such a system from order of magnitude arguments. These reveal an important distinction between the cases when the bath is deep or shallow as compared to the wavelength of the wave. In the former case, if we displace some fluid so that the surface in a region of size L is raised by an amount h , then we expect the fluid to respond by sinking back causing a flow extending to depth $\sim L$ below the surface. The mass of fluid involved in the flow is therefore $M \sim \rho L^3$ and the gravitational force is $F \sim L^2 g h$. Setting $F = Ma = -M\ddot{h}$ gives

$$\ddot{h} \sim -\frac{L^2 g h}{\rho L^3} \sim -\frac{g}{L} h \quad (20.36)$$

which is a simple harmonic oscillator with frequency

$$\omega^2 \simeq g/L. \quad (20.37)$$

One can read from this that gravity waves in a deep body of fluid are dispersive, since $\omega \sim \sqrt{k}$ (with $k \sim 2\pi/\lambda$ the wave-number) and therefore phase and group velocities $v_{\text{phase}} = \omega/k$ and $v_{\text{group}} = \partial\omega/\partial k$ have non-trivial dependence on wavelength. Also, one can see that the period of gravity waves of length-scale L is on the order of the period of a pendulum of that length.

One can perform a similar argument for a shallow bath of depth $D \ll L$, in which case the mass of fluid involved is $M \sim \rho L^2 D$ whereas the restoring force is the same as before $F \sim L^2 g h$ and one obtains an equation of motion

$$\ddot{h} \sim \frac{g}{D} h \quad (20.38)$$

which is a non-dispersive wave equation.

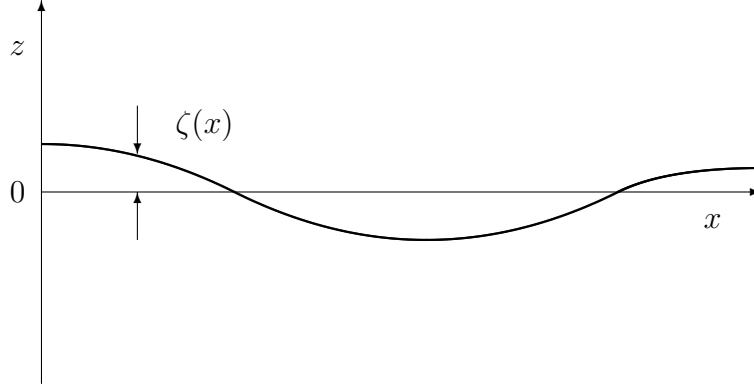


Figure 20.3: Geometry for the calculation of gravity wave dispersion relation.

Now let's see how this can be made more precise. We start with the Euler equation as above, replace $\Phi \rightarrow gz$, and linearize in the amplitude of the waves. This means we drop the u^2 term since it is of second order. Letting ζ denote the displacement of the surface (see figure 20.3), and with zero (or negligible) external pressure, Euler's equation provides the boundary condition

$$\left(\frac{\partial \Psi}{\partial t} \right)_{z=\zeta} + g\zeta = 0 \quad (20.39)$$

or, more usefully, taking the time derivative of this and using $\dot{\zeta} = u_z = \partial \Psi / \partial z$, this is

$$\left(g \frac{\partial \Psi}{\partial z} + \frac{\partial^2 \Psi}{\partial t^2} \right)_{z=\zeta} = 0. \quad (20.40)$$

This condition must be satisfied on the surface $z = \zeta$. However, since both Ψ and ζ are first order quantities, we can effectively apply the above condition at $z = 0$ and make only an error of second order. The pair of equations we need to solve are

$$\begin{aligned} \nabla^2 \Psi &= 0 \\ \left(g \frac{\partial \Psi}{\partial z} + \frac{\partial^2 \Psi}{\partial t^2} \right)_{z=0} &= 0 \end{aligned} \quad (20.41)$$

As usual, we want to look for traveling wave solutions $\Psi = f(z)e^{i(\omega t - kx)}$ and let's guess that $f(z)$ is some kind of exponential, so substituting $\Phi = ce^{az+i(\omega t - kx)}$ in Poisson's equation gives $\nabla^2 \Psi = (a^2 - k^2)\Psi = 0$ and therefore $a = \pm k$. Since z is negative below the surface, for the infinitely deep bath (which is typically the most interesting case) we need to take the solution $a = -k$, so the velocity field disturbance falls off exponentially with depth with e-folding scale $k^{-1} = \lambda/2\pi$. This justifies the assumption in the order of magnitude analysis that the velocity field extends a distance $\sim \lambda$ below the surface. The boundary condition equation now becomes $gk\Psi = -\omega^2\Psi$ which gives the dispersion relation

$$\omega(k) = \sqrt{gk} \quad (20.42)$$

again in accord with the hand-waving analysis.

For a bath of finite depth one needs to provide also a boundary condition that $u_z = \partial \Psi / \partial z$ vanish on the floor of the bath, and one finds then that the solution is a combination of the growing and decaying exponential solutions.

The dispersive nature of these waves is quite interesting. The group velocity is $\partial \omega / \partial k \propto k^{-1/2}$ so the group velocity scales as the square root of the wavelength. Thus a localized disturbance (an impulse) will evolve into a 'chirp' with the low-frequencies arriving first.

Note that $v_{\text{phase}} = \sqrt{g/k}$ whereas $v_{\text{group}} = \frac{1}{2}\sqrt{g/k}$ which means that wave crests travel twice as fast as the group velocity. For a packet of waves this means that wave crests will appear as if from nowhere at the tail of the packet, march forward through the packet, and disappear to the front. All of this is familiar to anyone who has idly tossed pebbles into a pond.

20.9 Sound Waves

To obtain the wave equation for small amplitude acoustic oscillations we start with the equations of continuity and the Euler equation

$$\begin{aligned} \left(\frac{\partial \rho}{\partial t} + \mathbf{u} \cdot \nabla \rho \right) &= -\rho \nabla \cdot \mathbf{u} \\ \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} \right) &= -\nabla P / \rho \end{aligned} \quad (20.43)$$

We then set

$$\begin{aligned} \rho &= \rho_0 + \rho_1 \\ P &= P_0 + P_1 \\ \mathbf{u} &= 0 + \mathbf{u}_1 \end{aligned} \quad (20.44)$$

where a subscript 0 denotes the equilibrium solution in the absence of waves (so $\mathbf{u}_0 = 0$ of course) and a subscript 1 denotes an assumed small perturbation about the equilibrium.

Substituting (20.44) in (20.43) and keeping only terms of first order gives

$$\begin{aligned} \frac{\partial \rho_1}{\partial t} + \rho_0 \nabla \cdot \mathbf{u}_1 &= 0 \\ \frac{\partial \mathbf{u}_1}{\partial t} + \frac{1}{\rho_0} \nabla P_1 &= 0 \end{aligned} \quad (20.45)$$

now if the oscillations are *adiabatic* then

$$P_1 = \left(\frac{\partial P}{\partial \rho} \right)_s \rho_1 \quad (20.46)$$

and taking the time derivative of the first of (20.45) and subtracting ρ_0 times the gradient of the second gives

$$\frac{d^2 \rho_1}{dt^2} - \left(\frac{\partial P}{\partial \rho} \right)_s \nabla^2 \rho_1 = 0 \quad (20.47)$$

which is a wave equation in the scalar quantity ρ_1 with sound speed c_s given by

$$c_s^2 = \left(\frac{\partial P}{\partial \rho} \right)_s. \quad (20.48)$$

For the special case of a planar disturbance $\partial/\partial y = \partial/\partial z = 0$ we have

$$\frac{d^2 \rho_1}{dt^2} - c_s^2 \frac{d^2 \rho_1}{dx^2} = 0 \quad (20.49)$$

with general solution

$$\rho_1(x, t) = f_1(x - ct) + f_2(x + ct) \quad (20.50)$$

where f_1 and f_2 are two arbitrary functions which must be determined from the boundary conditions.

In general the solutions of the 3-dimensional wave equation are superpositions of traveling waves

$$\rho_1(\mathbf{r}, t) = \sum_{\mathbf{k}} A_{\mathbf{k}} e^{i(\omega t - \mathbf{k} \cdot \mathbf{r})}. \quad (20.51)$$

The dispersion relation for these waves is

$$\omega^2 - k^2/c_s^2 = 0 \quad \rightarrow \quad \omega = k/c_s \quad (20.52)$$

so these waves are non-dispersive.

We can see under what conditions we are allowed to neglect terms of 2nd order (and higher) in the wave amplitude. For example, in passing from (20.43) to (20.47) we have, for example, dropped $(\mathbf{u} \cdot \nabla) \mathbf{u}$ as compared to $\partial \mathbf{u} / \partial t$ but for a plane wave $(\mathbf{u} \cdot \nabla) \mathbf{u} \sim k u^2$ while $\partial \mathbf{u} / \partial t \sim \omega u$ so this is justified provided $k u \ll \omega$ or, since $\omega = k c_s$ if $u \ll c_s$ so the linearized wave equation is valid provided the velocity associated with the waves is small compared to the sound speed. It is easy

to see from the first of (20.45) that $\rho_1/\rho_0 \sim ku/\omega \sim u/c_s$ so the linearized wave equation is valid provided the fractional amplitude of the density fluctuation is small: $\rho_1/\rho_0 \ll 1$ also.

The sound speed is given by (20.48). For adiabatic compression or rarefaction we found $\epsilon \propto \rho^{2/3}$ and $P \propto \rho \epsilon \propto \rho^{5/3}$ so $(dP/d\rho)_s = (5/3)(P_0/\rho_0)$. Now $P_0 = \rho_0 \langle w^2 \rangle / 3$ and therefore we have

$$c_s = \sqrt{\frac{5}{9} \langle w^2 \rangle} \quad (20.53)$$

ie the sound speed is on the order of the rms random thermal velocity. Note that the sound speed is independent of the density and is proportional to the square root of the temperature.

20.10 Problems

20.10.1 Ideal fluids

Write down the continuity and energy equations for an ideal gas in the form

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \right) \rho &= \dots \\ \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \right) \varepsilon &= \dots \end{aligned} \quad (20.54)$$

where ε is the specific thermal energy density.

Combine these equations to show that $\rho \varepsilon^{-3/2}$ is constant along fluid trajectories, so $\varepsilon \propto \rho^{2/3}$, and that the specific entropy s is therefore constant.

Estimate the rate (in degrees per km) at which dry air cools adiabatically if raised in height from sea level.

20.10.2 Potential flows

Potential flows have vanishing vorticity: $\nabla \times \mathbf{u} = 0$. A sufficient condition for this is that the velocity be the gradient of some scalar potential field $\phi(\mathbf{r})$. By considering the fourier transform of a general 3-dimensional vector field (or otherwise) show that this is also a necessary condition.

Show that it is possible to reconstruct the full 3-dimensional velocity field for a potential flow from measurements only of the line-of-sight component of the velocity. This technique is used in studies of cosmological ‘bulk-flows’ caused by nearby superclusters.

20.10.3 Hydrostatic equilibrium

a) Write down the equations of hydrostatic equilibrium for gas of density ρ and pressure P . Show that if $P = a\rho^2$ these can be combined to yield

$$\nabla^2 \rho + \frac{2\pi G}{a} \rho = 0 \quad (20.55)$$

b) Show that this equation admits spherically symmetric solutions of the form $\rho(r) = A \sin kr/r$ if $r < \pi/k$, $\rho = 0$ otherwise, where $k = \sqrt{2\pi G/a}$ and A is an arbitrary constant (which is fixed once the total mass is specified).

c) Show that there are also solutions of the form $\rho(r) = B \cos kr/r$, but that these require an addition point mass at the origin.

d) Interestingly, the above equation also admits non-spherical solutions:

$$\rho(r) = A \cos(k_x x) \cos(k_y y) \cos(k_z z) \quad (20.56)$$

for $|x| < \pi/2k_x$ etc. Find the condition that must be satisfied by the coefficients k_x, k_y, k_z and sketch some isodensity contours in the plane $z = 0$ and for $k_x = k_y = 2/\pi$. Does this solution seem physically reasonable to you? Discuss what may be wrong with this type of solution.

e) Returning to the spherical case, consider the run of entropy with radius for the $\rho(r) = A \sin kr/r$ solution. Does it increase or decrease with r ? Is the solution convectively stable?

Chapter 21

Viscous Fluids

21.1 Transport Coefficients

The ideal fluid equations (20.1) were obtained from the general fluid equations (19.57) by dropping the terms involving the viscous shear tensor π_{ij} and the heat flux \mathbf{F} . These can be calculated from kinetic theory, in the limit that the mean free path is small compared to the scales over which macroscopic conditions vary. To lowest order the solution of the Boltzmann transport equation $Lf = (\partial f / \partial t)_{\text{coll}}$ is a *shifted Maxwellian*

$$f(\mathbf{r}, \mathbf{v}) = \frac{n(\mathbf{r})}{(2\pi kT(\mathbf{r})/m)^{3/2}} e^{-\frac{1}{2}m|\mathbf{v}-\mathbf{u}(\mathbf{r})|^2/kT(\mathbf{r})}. \quad (21.1)$$

However, this cannot be an exact solution of the BTE since the collision term vanishes identically, but the Liouville operator contains, for example, the spatial gradient term $\mathbf{v} \cdot \nabla_{\mathbf{x}} f$ which is non zero, so one needs to augment the locally Maxwellian zeroth order approximation with correction terms. The detailed calculation of the *transport coefficients* is known as the *Chapman-Enskog procedure* which is described in Huang's book for example. The key results are that the viscous shear tensor is given by

$$\pi_{ij} = \mu D_{ij} \quad (21.2)$$

where

$$D_{ij} = \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}(\nabla \cdot \mathbf{u})\delta_{ij} \quad (21.3)$$

where D_{ij} is the *deformation rate tensor* which is simply a symmetrized and traceless version of the shear tensor $\partial u_i / \partial x_j$. This says that the viscous shear tensor (which is essentially the traceless part of the velocity dispersion tensor) is driven by the velocity shear. The *coefficient of shear viscosity*, or more briefly the *viscosity*, μ in (21.2) is given by

$$\mu = \frac{5}{8} \frac{(\pi m k T)^{1/2}}{\sigma} \quad (21.4)$$

where σ is the velocity averaged cross-section. To order of magnitude $\mu \sim m v_T / \sigma$ where $v_T \sim \sqrt{kT/m}$ is the typical thermal random velocity.

The *heat flux* is given by

$$\mathbf{F} = -\kappa \nabla T \quad (21.5)$$

where $\kappa/\mu = (5/2)C_V$ is known as *Eucken's constant*.

These results are well confirmed predictions of kinetic theory. The derivation is mathematically involved, but the general form of the results can be well understood from simple arguments. Let us model a shearing fluid as a set of slabs of thickness on the order of the mean free path λ , as illustrated in figure 21.1. Each mean free time (or collision time) $\tau = \lambda/v_T$ most of the particles in a slab will be replaced by particles from the neighboring slabs. Consider the situation where there

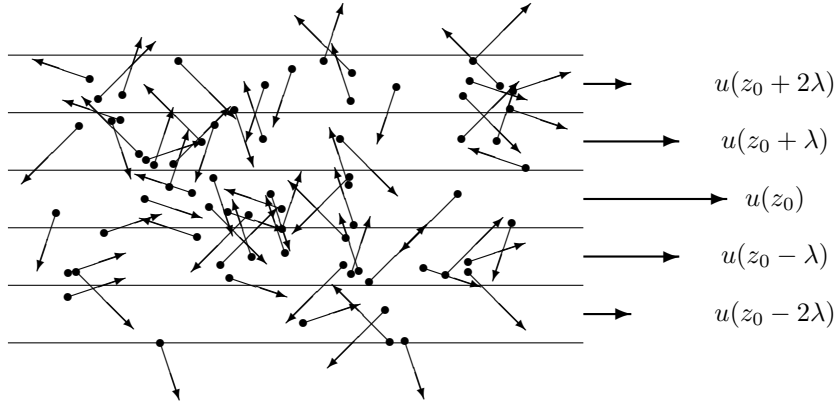


Figure 21.1: The effect of shear viscosity on irregularities in the velocity field can be understood crudely as a smoothing effect due to random motion of atoms. Here we have a stack of slabs each of which is roughly one mean free path thick. In one collision time $\tau \sim \lambda/v_T$, a good fraction of the particles in the center slab will have departed, to be replaced by the particles which were in the neighboring slabs. The new velocity of the center slab is approximatedly the average velocity of the two neighbors. This means that the acceleration of the center slab is $(\partial u/\partial t)_{\text{visc}} \sim \lambda^2 \partial^2 u / \partial z^2 / (\lambda/v_T) \sim \lambda v_T \partial^2 u / \partial z^2$. Another way to look at this is that there is a flux of x -component of momentum in the z -direction $\Delta P / \Delta A \Delta t \sim (m v_T / \sigma) \partial u / \partial z$. For the slabs adjacent to the center slab, for instance, there is a gradient of velocity, but no acceleration since the same amount of momentum flows through each face. For the center slab, however, there is a loss of x -momentum through both faces, so the slab decelerates.

is a continuous variation of the x -component of the flow velocity $u(z)$. If a particular slab at z_0 is moving faster than the average of its two neighbors then, after one collision time, this slab will have velocity $u' = u(z_0, t + \tau) \simeq (u(z_0 + \lambda, t) + u(z_0 - \lambda, t))/2$ and the acceleration of this slab is therefore

$$\frac{\partial u}{\partial t} \sim \frac{\Delta u}{\Delta t} = \frac{u(z_0, t + \tau) - u(z_0, t)}{\tau} \simeq \frac{u(z_0 + \lambda, t) - 2u(z_0, t) + u(z_0 - \lambda, t)}{2\tau} \quad (21.6)$$

but if the velocity field $u(z)$ varies smoothly the numerator in the final expression is approximately $\lambda^2 \partial^2 u / \partial z^2$, and since $\tau \sim \lambda/v_T$, the viscous acceleration is

$$\left(\frac{\partial u}{\partial t} \right)_{\text{visc}} \sim v_T \lambda \nabla^2 u. \quad (21.7)$$

On the other hand, in the fluid equations (19.57) this term is

$$\left(\frac{\partial u}{\partial t} \right)_{\text{visc}} = \rho^{-1} \nabla \cdot \pi \sim \frac{\mu}{\rho} \nabla^2 u. \quad (21.8)$$

since $\pi \sim \mu \nabla \mathbf{u}$. These are compatible if $\mu \sim \rho v_T \lambda \sim (mn)(kT/m)^{1/2} (n\sigma)^{-1} \sim \sqrt{mkT}/\sigma$ in agreement with (21.4).

The exchange of particles between slabs gives a *momentum flux* (in this case a flux of x -momentum in the direction in which u_x is decreasing). The exchange momentum between two slabs of area A is $\Delta P \sim \rho A \lambda \Delta u \sim mn A \lambda^2 \partial u / \partial z$ which is exchanged in time $\Delta t = \lambda/v_T$ and the momentum transport per unit time per unit area (or the force per unit area) is

$$\frac{\Delta P}{A \Delta t} \sim \frac{m v_T}{\sigma} \frac{\partial u}{\partial z} \quad (21.9)$$

so the momentum flux is proportional to the shear. If there is just a pure gradient of velocity then the momentum flux is constant, and the momentum of any slab does not change with time. If there

is a second derivative of the flow velocity then the force per unit area will be different on the top and bottom of the slab and there will be an acceleration of the slab (see figure 21.1).

It is interesting that the density of particles does not appear in the viscosity or in the momentum flux. While the number of particles per second crossing a surface increases with n , the distance they travel decreases since $\lambda \sim 1/(n\sigma)$ and these effects cancel.

The effect of viscosity in the Euler equation is to gradually smooth out the velocity flow. If $u(z)$ is initially fluctuating then the faster slabs will decelerate and *vice versa* until the flow becomes uniform with $u \rightarrow \bar{u}$ the mean of the initial velocity. Since the bulk kinetic energy density is ρu^2 , this smoothing of the velocity field results in a net loss (or dissipation) of the kinetic energy. This energy ends up in heat, and appears in the internal energy equation (19.54) as the ‘viscous dissipation’ term $\Psi \sim \pi \nabla u$.

The form of the thermal conductivity can be understood from the same kind of argument. If there is a gradient of temperature then the difference in energy per particle over a distance λ is $\Delta E_1 \sim k\lambda \nabla T$ and since the number of particles in a slab of volume $A\lambda$ is $N \sim nA\lambda$ and these exchange in a time $\tau = \lambda/v_T$, the flux of energy per unit area is

$$F \sim \frac{N\Delta E_1}{A\tau} \sim \frac{kv_T}{\sigma} \nabla T \quad (21.10)$$

again in accord with (21.5).

21.2 Damping of Sound Waves

Keeping the viscous term in the Euler equation and the dissipation term in the internal energy equation and linearizing as before gives

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} &= -\frac{1}{\rho_0} \nabla P_1 + \alpha \lambda v_T \nabla^2 \mathbf{u} \\ \frac{\partial \epsilon_1}{\partial t} &= -\frac{P_0}{\rho_0} \nabla \cdot \mathbf{u} + \beta \lambda v_T \nabla^2 \epsilon_1 \end{aligned} \quad (21.11)$$

with α, β dimensionless factors of order unity. Strictly speaking, the dissipation term here will cause the gas to heat up, and the background solution ϵ_0 will not be exactly constant, as we have assumed. However, the energy in the waves is second order in the wave amplitude and we can ignore this and take the gas to be effectively isentropic, so $\nabla P_1 = (5/2)(P_0/\epsilon_0) \nabla \epsilon_1 = (5/3)\rho_0 \nabla \epsilon_1$.

If we try solutions

$$\begin{aligned} \mathbf{u} &= \hat{\mathbf{k}} U e^{i(\omega t - \mathbf{k} \cdot \mathbf{x})} \\ \epsilon_1 &= E e^{i(\omega t - \mathbf{k} \cdot \mathbf{x})} \end{aligned} \quad (21.12)$$

and the differential equations (21.11) become the algebraic equations

$$\begin{aligned} i\omega U - \frac{5}{3}ikE + \nu k^2 U &= 0 \\ i\omega E - \frac{3}{5}ikc_s^2 U + \frac{\beta}{\alpha} \nu k^2 E &= 0 \end{aligned} \quad (21.13)$$

where we have defined the *kinematic viscosity* $\nu = \alpha \lambda v_T$.

The first of these equations gives $U = \frac{5}{3}ikE/(\omega + \nu k^2)$ and using this in the second equation and dividing through by E gives the dispersion relation

$$(i\omega + \nu k^2)(i\omega + \frac{\alpha}{\beta} \nu k^2) = -c_s^2 k^2 \quad (21.14)$$

and if we assume that the damping is weak (so the fractional change in the amplitude of the wave per cycle is small) then the frequency is

$$\omega = c_s k + i\Gamma \quad (21.15)$$

where the damping rate is $\Gamma \sim \nu k^2$.

- This damping rate is proportional to k^2 so short waves damp out fastest.
- The damping time is $t_{\text{damp}} \sim L^2/\lambda v_T \sim (L/\lambda)^2 t_{\text{coll}}$ (with λ the mean free path and L the wavelength), but this is just the time it takes for a particle to random walk a distance L .

21.3 Reynold's Number

The Euler equation for a potential flow and neglecting gravity is

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{1}{2} \nabla u^2 = -\frac{1}{\rho} \nabla P + \nu \nabla^2 \mathbf{u}. \quad (21.16)$$

Ignoring the last term on each side gave non-dissipating sound waves. Including viscosity, as we have seen, causes the sound waves to decay. The second term on the left hand side, which we shall call the ‘inertial term’, in contrast, leads to instability in the sense that irregularities will tend to grow. One way to see this is consider a single planar sound wave with spatial frequency \mathbf{k}_0 . The term $\nabla u^2/2$ therefore contains a component at spatial frequency $2\mathbf{k}_0$. This term therefore appears as a driving term for the oscillation mode $\mathbf{k} = 2\mathbf{k}_0$. The inertial term therefore couples modes of different spatial frequencies. In the absence of viscosity this leads to the ‘turbulent cascade’ with energy propagating from large scales to smaller scales.

This natural tendency to instability for non-viscous fluids can be stabilized if the viscosity is sufficiently large. If we have a velocity disturbance of amplitude u on scale L then the relative importance of the viscous and inertial terms is given by the dimensionless ratio

$$\text{Re} \equiv \frac{|\nabla u|^2}{\nu \nabla^2 u} \simeq \frac{u^2/L}{\nu u/L^2} \simeq \frac{uL}{\nu} \quad (21.17)$$

which is called the *Reynold's number*.

- If the Reynold's number for a system is large compared to unity the viscous effects are relatively unimportant and *vice versa*.
- Since $\nu \sim v_T \lambda$ the Reynold's number can be written as $\text{Re} \sim (u/c_s)(L/\lambda)$. Now for a fluid description to be at all valid we need $L \gg \lambda$, so the Reynold's number will typically be large (unless the velocity associated with the wave is tiny compared to the sound speed).
- If two systems have the same Reynold's number the solutions are *mathematically similar*, which means that the solutions for one can be obtained from those for another by applying a suitable scaling of lengths and times.

21.4 Problems

21.4.1 Viscous hydrodynamics

Water has a kinematic viscosity $\nu \sim 10^{-2} \text{cm}^2/\text{sec}$. Infer the mean free path for water molecules, and estimate the damping time for gravity waves on the ocean as a function of their wavelength. Review the derivation of the dispersion relation for deep ocean ‘gravity waves’: $\omega(k) = \sqrt{kg}$. Estimate the shortest wavelength waves which can reach Hawaii effectively unattenuated from a distant storm in the North Pacific.

21.4.2 Damping of Sound Waves

Use the ideal fluid continuity and energy equations to derive the linearised wave equation for sound waves. Show that the most general solution for planar sound wave disturbances is $f_1(x-ct) + f_2(x+ct)$ where f_1 and f_2 are arbitrary functions. Air has a kinematic viscosity $\nu \sim 0.15 \text{cm}^2/\text{sec}$. Using random walk arguments, or otherwise, estimate the damping time for a sound wave at ‘middle-C’ (256Hz).

21.4.3 Sound waves

The continuity, Euler and energy equations for an ideal gas can be written as

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla\right) \rho &= -\rho \nabla \cdot \mathbf{u} \\ \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla\right) \mathbf{u} &= -\nabla P / \rho - \nabla \phi \\ \left(\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla\right) \varepsilon &= -\frac{P}{\rho} \nabla \cdot \mathbf{u} \end{aligned} \quad (21.18)$$

a) What do the symbols $\mathbf{u}, \rho, P, \phi, \varepsilon$ represent?

b) Assuming a simple monatomic gas, combine the continuity and energy equations to show that $\rho \varepsilon^{-3/2}$ is constant along fluid trajectories, so $\varepsilon \propto \rho^{2/3}$, and that the specific entropy s is therefore constant.

c) Let $\rho(\mathbf{r}, t) = \rho_0 + \rho_1(\mathbf{r}, t)$, $P(\mathbf{r}, t) = P_0 + P_1(\mathbf{r}, t)$ where ρ_1, P_1 are understood to be small perturbations about a static uniform solution. Neglecting, for the moment, the effect of gravity, linearise the continuity and Euler equations and combine these to obtain the wave equation for adiabatic sound waves

$$\frac{\partial^2 \rho_1}{\partial t^2} - \left(\frac{\partial P}{\partial \rho}\right)_s \nabla^2 \rho_1 = 0 \quad (21.19)$$

d) Find the relation between the sound speed and the rms thermal velocity of atoms for a simple monatomic gas. How does the sound speed depend on density? How does the sound speed depend on temperature?

e) For finite mean free path, diffusion effects will modify the fluid equations. Indicate the general form of the extra terms in the fluid equations at the next level of approximation (Navier-Stokes) and describe their effect on sound waves.

f) Discuss qualitatively the effect of including the self-gravity of the sound wave. You should explain what is meant by the ‘Jeans length’, and indicate how it is related to the sound speed and the dynamical time $t_{\text{dyn}} \sim (G\rho)^{-1/2}$.

Chapter 22

Fluid Instabilities

Fluids are susceptible to various instabilities. We have already discussed convective instability. We first briefly mention the Rayleigh-Taylor and Kelvin-Helmholtz instabilities, and then discuss in a little more depth gravitational instability, thermal instability and turbulence.

22.1 Rayleigh-Taylor and Kelvin-Helmholtz

Rayleigh-Taylor instability arises when a heavy fluid lies over a lighter fluid; any small irregularity in the height of the interface will grow exponentially and rapidly become large.

The term is also used to describe what happens in an explosion with dense ejected material incident on lighter surrounding material.

In both situations, ripples in the interface grow into ‘fingers’ of dense matter penetrating the lighter material.

Kelvin-Helmholtz instability arises, for instance, when wind blows across the surface of a fluid, or more generally when there is wind shear in a stratified medium. In the former case the instability can be stabilized by surface tension of the interface. For a shearing stratified fluid the instability can be stabilized by a sufficiently large entropy gradient, much as we found for convective instability. We will discuss this later when we consider turbulence, the onset of which is closely related to Kelvin-Helmholtz instability.

22.2 Gravitational Instability

Gravitational instability appears if we add the gravitational term to the linearized sound wave equations. Starting with the continuity and Euler equations

$$\begin{aligned}\dot{\rho}_1 &= -\rho_0 \nabla \cdot \mathbf{u} \\ \dot{\mathbf{u}} &= -c_s^2 \frac{\nabla \rho_1}{\rho_0} - \nabla \Phi\end{aligned}\tag{22.1}$$

with $\nabla^2 \Phi = 4\pi G \rho_1$ the *peculiar gravity* caused by the density fluctuation. Taking the time derivative of the first and using the second to eliminate the velocity gives

$$\ddot{\rho}_1 = c_s^2 \nabla^2 \rho_1 + 4\pi G \rho_0 \rho_1\tag{22.2}$$

and using trial solution $\rho_1 \propto e^{i(\omega t - \mathbf{k} \cdot \mathbf{x})}$ yields the dispersion relation

$$\omega^2 = k^2 c_s^2 - 4\pi G \rho_0.\tag{22.3}$$

- The last term is the inverse square of the dynamical-time or collapse-time for a body of density ρ_0 .

- The squared frequency becomes negative (signalling exponentially growing instability) for waves with wavelength exceeding a critical value

$$\lambda_c = 2\pi/k_c = 2\pi c_s / \sqrt{4\pi G \rho_0} \quad (22.4)$$

- Crudely speaking, the criterion for stability of an over-dense region is that a sound wave should be able to cross the perturbation in less than the dynamical time.
- We have performed the ‘Jeans swindle’ in side-stepping the issue of the fact that the assumed stable background about which we are making a linearized perturbation is in fact unstable. We will see how this can be justified when we consider growth of density fluctuations in cosmology.

22.3 Thermal Instability

Consider a box with transparent walls full of atomic or molecular gas with uniform density ρ and temperature T . The atoms will have a Maxwellian velocity distribution and the atoms will have a thermal distribution of internal excitation levels. In the absence of any external radiation field the gas can cool through collisionally induced excitation followed by spontaneous emission, and there will be an emissivity which is a function of the temperature and the density of the gas: $j = j(T, \rho)$. If we switch on an external radiation source, then one should be able to find combinations of temperature and pressure such that the radiation loss is just canceled by absorption, and the system will then be in equilibrium. However, for most temperatures, this equilibrium will be unstable. This is because the energy loss due to collisions scales as the density squared, whereas the energy gain rate is proportional to the density times the absorption cross-section.

Imagine one has gas with thermodynamic parameters (ρ_0, T_0) which is just in balance between heating and cooling. If one were to perturb a parcel of this gas by injecting a little heat then it will expand and move to a new point with a lower density. It must, however, have a higher temperature than it had initially, since it must remain in pressure equilibrium. Similarly, if we remove a little heat it must contract slightly, and become denser (and therefore cooler if it is to remain in pressure equilibrium). This is illustrated in figure 22.1. If the cooling rate is insensitive to the temperature, then an over-dense, and therefore cooler, perturbation will radiate more effectively and will become still denser and cooler. An under-dense, and therefore hotter, region will radiate less efficiently, so the radiative energy input will exceed the rate at which energy is radiated and the region will become even more under-dense and hotter. The initial density inhomogeneity seed, no matter how small, will grow exponentially and the system is said to be ‘thermally unstable’.

The only way to avoid this instability is if the cooling rate is strongly dependent on the gas temperature, with the cooling being more effective at higher temperature. If the temperature dependence is sufficiently strong, then the cooling rate under isobaric conditions may *decrease* with increasing density, and the gas can then be stable.

This strong, positive dependence of cooling rate on temperature can occur for certain, rather specific, temperatures. For example, for atomic gas at temperatures below about 10^4K the probability that an atom is excited is exponentially small; thus a small increase in temperature results in an exponential increase in the fraction of excited atoms, and consequently in the radiative cooling rate. A similar transition occurs around 10^2K for molecules where molecular vibration states get excited.

A body of gas at an intermediate temperature, say around 10^3K , will be unstable, and will precipitate over-dense clouds which will end up at around 100K surrounded by a hotter gas at around 10^4K . This leads to the ‘two-phase medium’ picture (or more generally multi-phase medium) of the interstellar medium.

The condition for stability is $(\partial\mathcal{L}/\partial T)_P > 0$, where \mathcal{L} is the cooling rate.

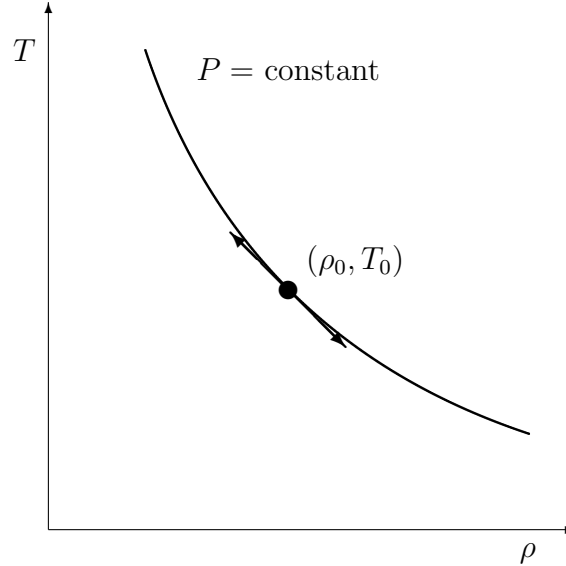


Figure 22.1: The curve shows an isobar passing through the state (ρ_0, T_0) , for which the collisionally induced cooling just balances the energy input radiation. If we perturb an element of the gas by adding or removing some heat then it will move along this isobar. Since cooling ordinarily tends to be more efficient for higher densities, this is unstable (see text). To avoid the instability it is necessary to have the cooling rate increase strongly with temperature.

22.4 Turbulence

Empirically, flows with small Reynold's number are stable, whereas high Re flows are unstable and lead to chaotic behavior. Examples are stirring cups of honey and coffee respectively. The latter type of flow tends to form *eddies*, which form sub-eddies and so on, and there is said to be a *turbulent cascade* of energy from some *outer scale* — the scale on which bulk kinetic energy is being injected — down to some *inner scale* at which viscosity becomes important and the energy is converted to heat.

Turbulent flows are highly non-linear and complicated and are a challenge even to numerical calculations. There are however some simple scaling properties of turbulent flows that are expected to hold if the inner scale is much less than the outer scale. These scaling laws are known as the *Kolmogorov spectrum*. The range of scales $L_{\text{inner}} \lesssim L \lesssim L_{\text{outer}}$ is known as the *inertial range*.

22.4.1 Kolmogorov Spectrum

Assuming the *Mach number* to be small, the flow can be modeled as approximately incompressible. At length scale L , there will be eddies with some characteristic velocity $v(L)$ and the kinetic energy density in these eddies is $\epsilon \sim \rho v^2(L)$. These eddies turn over on time-scale $\tau(L) \sim L/v(L)$ and it is reasonable to assume that these eddies transfer their energy to smaller scale eddies on the order of the turnover time. These eddies will also be receiving energy from larger scale 'parent' eddies. In the steady state there can be no build up of energy at any particular scale, so we require that the energy density transfer rate $d\epsilon/dt \sim \epsilon/\tau$ be independent of L . This gives $\rho v^2(L) \propto L/v(L)$, or

$$v(L) \propto L^{1/3} \quad (22.5)$$

this is the Kolmogorov velocity spectrum for fully developed turbulence.

22.4.2 Passive Additives

Similar arguments can be applied to the spatial power spectrum of *passive additives*. A passive additive is a property which is conserved along the flow. Examples are the entropy of a gas, water vapor content of air, ‘creaminess’ of liquid in a just stirred cup of coffee. For the last example, defining the creaminess $c(\mathbf{r})$ to be unity in the cream and zero otherwise then after we pour the cream, the power spectrum $P_c(k)$ will be dominated by low frequencies, but the turbulent cascade will cause power to be created at higher frequencies. Now the variance of the additive $\langle c^2 \rangle$ is conserved (until we get down to scales where diffusion becomes effective) and so again, in the steady state, in which there is some agency injecting kinetic energy and creaminess fluctuations on a large scale L_{outer} and these are cascading to smaller scales on a time-scale of order the eddy turnover time. The rate at which variance is being transferred $\langle c^2 \rangle_L / \tau$ should be independent of scale. This leads to

$$\langle c^2 \rangle_L \propto \tau(L) \sim \frac{L}{v(L)} \propto L^{2/3}. \quad (22.6)$$

Since $\langle c^2 \rangle_L \sim (k^3 P_c(k))_{k \sim 1/L}$ the scaling (22.6) corresponds to a power-law power spectrum $P_c(k) \propto k^n$ with $3 + n = -2/3$ or

$$P_c(k) \propto k^{-11/3}. \quad (22.7)$$

For this type of power-law spectrum, the auto-correlation function is not well-defined (in reality, the value $\xi_c(r)$ will be determined by the outer-scale) but the structure function is well defined and has the power law form

$$S_c(r) = \langle (c(0) - c(\mathbf{r}))^2 \rangle \propto r^{2/3}. \quad (22.8)$$

22.4.3 Inner Scale

The turbulent cascade persists down to the *diffusion scale* at which viscosity acts to damp out the flows and irreversibly mix any passive additives (thus destroying additive variance).

The diffusion scale is such that the damping rate $u^{-1} \partial u / \partial t \sim \nu / L^2 \simeq 1 / \tau(L) \sim v(L) / L$ with $\nu \sim v_T \lambda$ the kinematic viscosity. But with $v(L) = v(L_{\text{outer}})(L / L_{\text{outer}})^{1/3}$ this gives

$$L_{\text{inner}} \sim L_{\text{outer}} \left(\frac{\nu}{L_{\text{outer}} v(L_{\text{outer}})} \right)^{3/4} \quad (22.9)$$

22.4.4 Atmospheric Seeing

An interesting application of turbulence theory is provided by *atmospheric seeing* which arises from turbulence driven by large scale wind shear mixing air with inhomogeneous entropy, water vapor etc. Since atmospheric turbulence is strongly sub-sonic we can assume that the air is in pressure equilibrium, so fluctuations in the entropy are reflected in fluctuations in the density and refractive index δn .

These refractive index fluctuations cause corrugation of the wavefronts from distant sources. The vertical deviation of the wavefront (or effectively the phase fluctuation) is proportional to the integral of the air density, and has the same spectral index $P_h \propto k^{-11/3}$ and the two-dimensional structure function is $S_h \propto r^{5/3}$. A realization of a wavefront after propagation through a turbulent atmosphere is shown in figure 22.2.

The structure function for the phase fluctuations is also a power law, and is conventionally parameterized in terms of the *Fried length* r_0 as

$$S_\varphi(r) = 6.88(r/r_0)^{5/3}. \quad (22.10)$$

The Fried length is therefore the scale over which the rms phase difference is $\sqrt{6.88}$, so different parts of the wave-front separated by r_0 or more will be significantly out of phase with each other.

We saw earlier that the optical transfer function of the telescope (the Fourier transform of the point spread function) is proportional to $e^{-S_\varphi(r)/2}$ with S_φ the structure function for phase

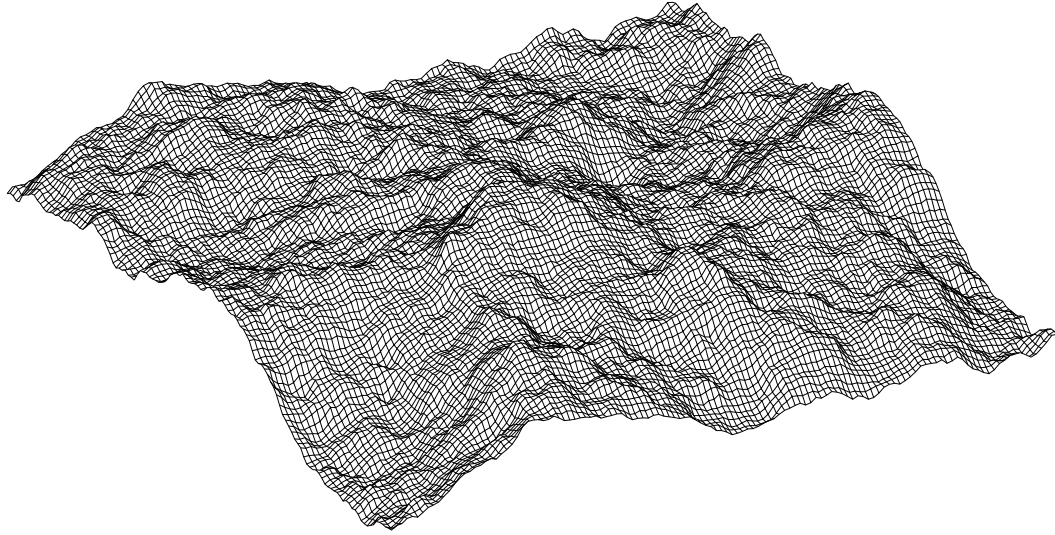


Figure 22.2: Realization of wavefront from a distant source after propagating through the atmosphere. The projected phase fluctuations were modeled as a Gaussian random field.

fluctuations. Since $S_\varphi \propto r^{5/3}$, taking the transform leads to a PSF $g(\theta)$ with a core of radius $\theta_c \sim \lambda/r_0$ and with a power-law halo or ‘aureole’ $g(\theta) \propto \theta^{-11/3}$, for $\theta \gg \theta_c$.

This is for an ideal imaging system. In real telescopes there are usually sharp edges to the pupil or input stop, and this results in wings of the PSF with $g \propto r^{-3}$. There is also usually mirror roughness which typically exceeds the atmospheric corrugation on small scales $\ll r_0$ and typically generates a PSF with a $g \propto r^{-2}$. This indicates that the spectrum of mirror errors has index $n \simeq -2$, which is 2-dimensional flicker noise. Putting these affects together suggests that a real PSF will consist of a core, surrounded by a series of power law extensions of progressively shallower slopes.

22.4.5 Stability

In the atmosphere the Reynold’s number for shearing flows on scales of order km is enormous, so one might expect turbulence to be ubiquitous. However, it is possible for the atmosphere to be stabilized against turbulence by a positive entropy gradient (entropy increasing with altitude) since a positive entropy gradient means it costs energy to create a large scale eddy. Comparing the kinetic energy in velocity shear in a region of size L to the potential energy cost for turning over a cell of this size leads to *Richardson’s criterion for stability*.

The kinetic energy is straightforward. For a cell of size L it is $E_{\text{kin}} \sim \rho L^3 (L \nabla u)^2 \sim \rho u^2 L^5$.

To compute the energy cost, recall that the specific entropy is

$$s = \frac{k}{m} \left(\frac{3}{2} \ln T - \ln n \right), \quad (22.11)$$

where m is the mass of the air molecules. Assuming adiabaticity, the density and temperature are related by

$$n = \frac{T^{3/2}}{\exp(ms/k)}. \quad (22.12)$$

Let’s say we take some parcel of gas from altitude z' at which the specific entropy is s' and raise it to level z where the entropy is s (and the density and temperature are n and T). If the density and temperature of the displaced gas are n' and T' then these satisfy

$$n' = \frac{T'^{3/2}}{\exp(ms'/k)} \quad (22.13)$$

but they must also satisfy pressure equilibrium:

$$n'T' = nT \quad (22.14)$$

so, using this to eliminate the temperature T' in (22.13) gives

$$(n')^{5/2} = \frac{n^{3/2}T^{3/2}}{\exp(ms'/k)} \quad (22.15)$$

or, using (22.12)

$$\left(\frac{n'}{n}\right)^{5/2} = \frac{\exp(ms/k)}{\exp(ms'/k)} = \exp(m(s-s')/k) \quad (22.16)$$

With $n' = n + \Delta n$ and $s' = s - \Delta s$, and assuming small fractional changes in the density, entropy etc

$$1 + \frac{5}{2} \frac{\Delta n}{n} \simeq 1 + m\Delta s/k \quad (22.17)$$

or

$$\frac{\Delta n}{n} = \frac{2}{5} m\Delta s/k \quad (22.18)$$

If we overturn a parcel of fluid of size L through its own length, the fractional density change is therefore $\Delta\rho/\rho \sim mL\nabla s/k$, and the potential energy cost is $E_{\text{pot}} \sim g(\Delta\rho L^3)L \sim g\rho mL^5\nabla s$ so the ratio is

$$\frac{E_{\text{pot}}}{E_{\text{kin}}} \sim \frac{gm\nabla s}{k(\nabla u)^2}. \quad (22.19)$$

An atmosphere will be stable if this dimensionless number exceeds unity.

22.5 Problems

22.5.1 Kolmogorov turbulence

Give an order of magnitude estimate for the specific energy ‘flux’ (from parent to daughter eddies) in a turbulent cascade in terms of the characteristic velocity U_L of eddies of size $\sim L$. Thereby obtain the Kolmogorov scaling law for fully developed turbulence in the form $U_L \propto L^\gamma$, and the corresponding law for the turnover time $\tau(L)$.

If the mean free path in air is $l = 4 \times 10^{-6}$ cm and the mean thermal velocity is $v_T = 4 \times 10^4$ cm/s, estimate the time for atoms to diffuse a distance L .

By combining the above results, estimate the ‘inner scale’ at which the bulk kinetic energy is dissipated by viscosity into heat, if the ‘outer scale’ for atmospheric turbulence is ~ 100 m and the velocity at this scale is ~ 20 m/s,

Chapter 23

Supersonic Flows and Shocks

Subsonic flows with $u \ll c_s$ are nearly incompressible while for transsonic ($u \sim c_s$) and supersonic ($u \gg c_s$) flows compression of the gas is important.

One can see this from Bernoulli's equation $u^2/2 + h = \text{constant}$, with $dh = dP/\rho$ as we will assume adiabaticity. Writing $dP = (\partial P/\partial \rho)_s d\rho = c_s^2 d\rho$ and $du^2/2 = u du$ gives

$$u^2 \frac{du}{u} + c_s^2 \frac{d\rho}{\rho} = 0. \quad (23.1)$$

or

$$\frac{\Delta \rho}{\rho} = -M^2 \frac{\Delta u}{u} \quad (23.2)$$

where the *Mach number* is defined by

$$M = u/c_s. \quad (23.3)$$

It follows that

$$\left| \frac{\Delta \rho}{\rho} \right| \left\{ \begin{array}{l} \ll \\ \gg \end{array} \right\} \left| \frac{\Delta u}{u} \right| \quad \text{for} \quad \left\{ \begin{array}{l} M \ll 1 \\ M \gg 1 \end{array} \right. \quad (23.4)$$

23.1 The de Laval Nozzle

An interesting application is the *de Laval nozzle* in which gas is accelerated to supersonic velocities. This is the basis of rocketry and the physics is relevant for astrophysical jets.

In the de Laval nozzle gas passes adiabatically through a constricting *venturi*. The continuity equation is $\rho A u = \text{constant}$, or $d\rho/\rho + du/u + dA/A = 0$. Using (23.2) this says

$$(1 - M^2) \frac{du}{u} = -\frac{dA}{A}. \quad (23.5)$$

For low Mach number $M < 1$, a constriction $dA < 0$ causes an acceleration $du > 0$ so subsonic gas entering a venturi will accelerate. For $M > 1$ though a constriction requires a deceleration. Thus, if one can arrange that the velocity just equals the sound speed in the waist of the venturi, the gas can be accelerated both entering and leaving the venturi.

This is an effective way to convert thermal energy in the gas into bulk kinetic energy.

23.2 Shock Waves

If a subsonic flow passes an obstacle, sound waves can propagate upstream and 'warn' the incoming gas, so the flow can adjust itself to accommodate the obstacle. For a supersonic flow this is impossible and obstacles lead inevitably to *shock waves*. Gas encountering a shock suffers a sudden and irreversible increase in entropy as bulk kinetic energy is converted to heat. The microscopic behavior of the shock is complicated, but important features can be obtained from the principles of conservation of mass, momentum and energy. Here we will consider collisional shocks.

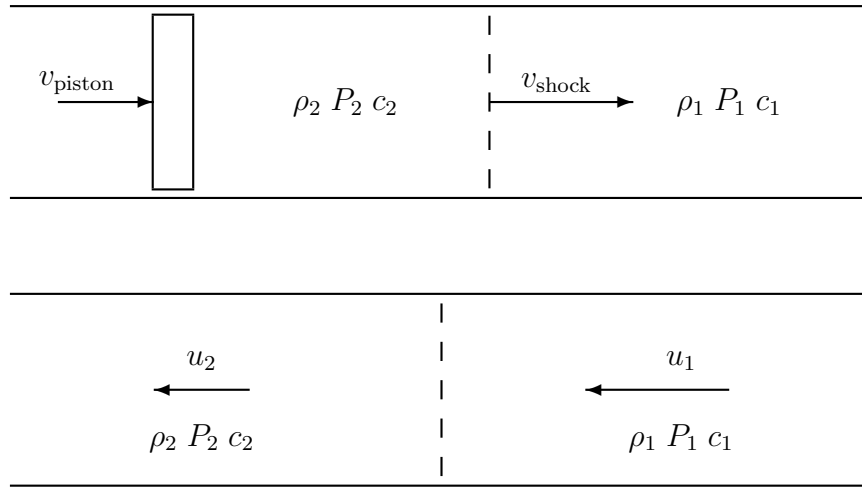


Figure 23.1: The upper sub-figure shows a piston begin driven along a tube at velocity v_{piston} . Initially the tube contains gas with sound speed $c_1 < v_{\text{piston}}$. A ‘shock-wave’ preceeds the piston and propagates into the cold gas. In the shock, collisions convert bulk kinetic energy into heat. Behind the shock the tube contains heated and compressed gas. The lower sub-figure shows the same thing from the point of view of an observer who moves along with the shock. The basic features of the shock — shock velocity, density and pressure jumps etc — can be determined from continuity of flux of mass, momentum and energy, independent of the details of the collisions happening in the shock.

23.2.1 The Shock Tube

Consider a tube containing gas with density ρ_1 , pressure P_1 and sound speed c_1 with a piston begin driven along the tube at velocity $v_p > c_1$ as illustrated in figure 23.1. The gas particles initially impacted by the piston will be strongly accelerated. They will in turn collide with particles further along the tube, and it is not unreasonable to assume that a disturbance will proceed along the tube leaving an irreversibly heated body of gas in its wake. This is known as a *collisional shock*; a narrow region within which collisions rapidly generate entropy.

To analyze this is is convenient to shift to a moving frame of reference in which the putative shock is stationary, as illustrated in the lower part of figure 23.1. In this frame, gas arrives from one side at velocity u_1 and with ‘pre-shock’ density ρ_1 and pressure P_1 , gets heated and accelerated and leaves with ‘post-shock’ velocity, density and pressure u_2 , ρ_2 and P_2 . To obtain the relation between pre- and post-shock quantities we require that in the steady state, the flux of mass, x -momentum and energy should be the same on both sides of the shock. Continuity of mass flux is trivial: we need $\rho_1 u_1 = \rho_2 u_2$. Continuity of momentum and energy is trickier since we need to take account of both bulk and micro-scopic motions.

The rate at which particles cross a unit area which is perpendicular to the flow is $dN/dtdA = \int d^3v f(\mathbf{v})v_x$. The mass flux is therefore

$$j = m \frac{dN}{dt dA} = m \int d^3v f(\mathbf{v})v_x = m \int d^3w f_0(\mathbf{w})(u + w_x) \quad (23.6)$$

where $f_0(\mathbf{w})$ is the distribution of random velocities. From the definition $u = \langle v_x \rangle$ the mean random velocity vanishes $\int d^3w f(\mathbf{w})w_x = 0$ and we have

$$j = mnu = \rho u \quad (23.7)$$

as expected.

Momentum flux is

$$\int d^3v f(\mathbf{v})v_x(mv_x) = m \int d^3w f_0(\mathbf{w})(u + w_x)^2 = \rho u^2 + \rho \langle w_x^2 \rangle = \rho u^2 + P \quad (23.8)$$

where we have assumed an isotropic distribution of random velocities so that $\langle w_x^2 \rangle = \langle w^2 \rangle / 3 = P / \rho$. Continuity of x -momentum implies continuity of $\rho u^2 + P$.

Energy flux is

$$\begin{aligned} \int d^3v f(\mathbf{v})v_x \left(\frac{1}{2} m v^2 \right) &= \frac{1}{2} m \int d^3w f_0(\mathbf{w})(u + w_x)(u^2 + 2\mathbf{u} \cdot \mathbf{w} + w^2) \\ &= \rho u \left(\frac{1}{2} u^2 + \frac{1}{2} \langle w^2 \rangle + \langle w_x^2 \rangle \right) \\ &= \rho u \left(\frac{1}{2} u^2 + \epsilon + P / \rho \right) = \rho u \left(\frac{1}{2} u^2 + h \right) \end{aligned} \quad (23.9)$$

where $h = \epsilon + P / \rho$ is the specific enthalpy. Since ρu is just the mass flux (which is continuous), continuity of energy flux therefore implies continuity of $u^2 / 2 + h$.

The *jump conditions* are then

$$\begin{aligned} \rho_1 u_1 &= \rho_2 u_2 \\ \rho_1 u_1^2 + P_1 &= \rho_2 u_2^2 + P_2 \\ \frac{1}{2} \rho_1 u_1^2 + h_1 &= \frac{1}{2} \rho_2 u_2^2 + h_2 \end{aligned} \quad (23.10)$$

These were derived here for the simple case of a monatomic gas, but they are in fact more general.

It is interesting that $h = \epsilon + P / \rho$ and not just the internal energy ϵ appears in the energy flux continuity equation. This means that the sum of the bulk and thermal kinetic energies of a parcel of gas are *not* the same after the shock as before. This is because the *volume* of this element will have changed, in fact, as we will shortly see, the gas parcel will have been compressed, and this required work to be done which came from the gas itself.

To solve (23.10) we proceed as follows:

1. Let $u_1 = jV_1$ and $u_2 = jV_2$ where j is the mass flux and $V = 1/\rho$ is the *specific volume*, ie the volume occupied by unit mass of gas.
2. Eliminating the velocities u_1, u_2 , the second of (23.10) becomes

$$P_1 + j^2 V_1 = P_2 + j^2 V_2 \quad (23.11)$$

or

$$j^2 = \frac{P_2 - P_1}{V_1 - V_2} \quad (23.12)$$

3. Eliminating the velocities u_1, u_2 from the third of (23.10) gives

$$h_1 + \frac{1}{2} j^2 V_1^2 = h_2 + \frac{1}{2} j^2 V_2^2 \quad (23.13)$$

or

$$h_1 - h_2 + \frac{1}{2} (V_1 + V_2) (P_2 - P_1) = 0 \quad (23.14)$$

or, since $h = \epsilon + P / \rho$,

$$\epsilon_1 - \epsilon_2 + \frac{1}{2} (V_1 - V_2) (P_2 + P_1) = 0. \quad (23.15)$$

If we specify P_1 and $V_1 = 1/\rho_1$ then either of equations (23.14) or (23.15), whichever is most convenient, provides a relation between P_2 and V_2 . This relation is known as the *shock adiabat*.

Once we specify one post-shock quantity such as P_2 (which is determined eventually by the energetics of the piston or explosion which is driving the shock) then all other post-shock quantities are fixed.

A shock is said to be ‘strong’ if the post shock pressure greatly exceeds the pre-shock pressure. In this case we set $P_1 = h_1 = \epsilon_1 = 0$ and it is not difficult to show that the density increases by a factor 4.

See Landau and Lifshitz “Fluid Dynamics” for discussion of the stability of collisional shocks.

23.2.2 Vorticity Generation

Kelvin's circulation theorem tells us that an ideal fluid which is initially vorticity free remains that way for ever. However, with dissipation this is no longer the case. In collisional shocks there is strong dissipation in the shock and, if the shock is oblique (ie the normal to the shock does not coincide with the flow velocity vector) vorticity can be generated.

23.2.3 Taylor-Sedov Solution

The *Taylor-Sedov shock solution* describes what happens if a point-like explosion occurs in a uniform density gas. After some short time the explosion will have 'swept up' more than the mass of the material ejected in the explosion, and subsequently one expects to find a spherical shock wave propagating radially outward, with radius as a function of time $r = r(E, \rho, t)$ where E is the energy of the explosion and ρ is the density of the ambient gas.

By dimensional analysis, if we assume that the solution has power law dependence on E , ρ and t , ie $r = E^l \rho^m t^n$ then equating powers of mass, length and time on both sides of this equation gives the power law indices $l = 1/5$, $m = -1/5$ and $n = 2/5$ so the form of the solution is

$$r(t) = \alpha E^{1/5} \rho^{-1/5} t^{2/5}. \quad (23.16)$$

One could also have reached this conclusion from energy considerations. Note that the velocity of the shock is $v = (2/5)r/t$ and the kinetic energy is $E_{\text{kin}} \sim \rho r^3 v^2$ which is independent of r and t and is on the order of the initial energy injected.

The Taylor-Sedov solution applies to supernovae explosions. At late times cooling can become effective and the shock then evolves into the 'snow-plow' phase.

23.3 Problems

23.3.1 Collisional shocks

a) Write down or, if you like, derive the equations expressing the continuity of fluxes of mass, momentum and energy for a planar collisional shock. Work in frame of reference such that the shock is stationary and denote the pre- and post-shock velocities by u_1 , u_2 respectively and assume that the flow velocity is perpendicular to the shock surface.

These equations relate the densities ρ , pressures P and enthalpies $h = \varepsilon + P/\rho$ on the two sides of the shock discontinuity.

b) Take the limit of the continuity equations for a strong shock ($P_1 \ll P_2$) and show that for a monatomic gas the density jumps by a factor 4.

c) Find the relation between the shock propagation velocity (i.e. the velocity at which the shock moves into the unshocked medium) and the post-shock sound speed.

d) A strong spherically symmetric shock resulting from an explosion of energy E is propagating into uniform cold gas of density ρ . In the absence of radiative cooling the shock radius $r(t; E, \rho)$ can only be a power-law function of its arguments:

$$r = \alpha E^l t^m \rho^n \quad (23.17)$$

where α is some dimensionless constant which we will assume to be of order unity. Use dimensional analysis to find the indices l, m, n .

e) A supernova explosion dumps energy $\sim 10^{51}$ erg in a region containing uniform cool gas at density $n = 1\text{cm}^{-3}$. Use the 'Taylor-Sedov' scaling law you have just derived to compute the radius and velocity of the shock front after 100 years.

Chapter 24

Plasma

24.1 Time and Length Scales

Much of the baryonic material in the Universe is in the form of plasma; highly ionized, and therefore electrically conductive, gas. There are a variety of important length- and time-scales for plasmas:

24.1.1 Plasma Frequency

Consider an otherwise electrically neutral plasma where we take a block of electrons and displace them sideways by an amount x as illustrated in figure 24.1. This generates a restoring force so the equation of motion is

$$\ddot{\mathbf{x}} = -\omega_p^2 \mathbf{x} \quad (24.1)$$

where the *plasma frequency* is

$$\omega_p = \sqrt{4\pi n e^2 / m_e} \simeq 5.6 \times 10^4 \left(\frac{n}{\text{cm}^{-3}} \right)^{1/2} \text{s}^{-1} \quad (24.2)$$

Thus if we physically disturb a plasma it will ‘ring’ at the frequency ω_p . Also, as we will shortly see, electromagnetic waves can only propagate through the plasma if their frequency exceeds ω_p .

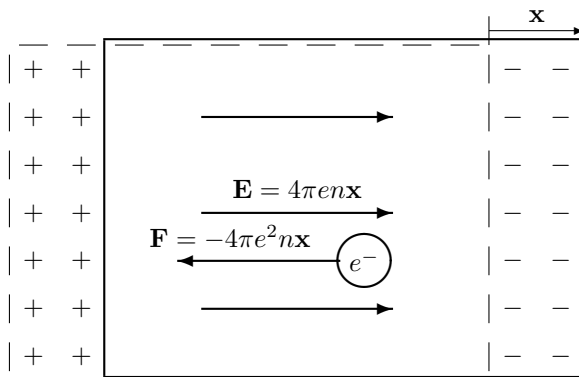


Figure 24.1: If we displace a block of electrons in a plasma from their initial location — as indicated by the dashed box — this generates an electric field which acts to try to restore neutrality. In the absence of damping, this leads to oscillation at the plasma frequency $\omega_p = \sqrt{4\pi n e^2 / m_e}$.

24.1.2 Relaxation Time

The *relaxation time* is the time-scale for collisions to establish a Maxwellian distribution of velocities. We can crudely estimate this as the time between collisions with sufficiently small impact parameter that they substantially deflect an electron. For such collisions, the potential energy at closest approach is on the order of the typical thermal kinetic energy:

$$\frac{e^2}{b} \simeq kT \quad (24.3)$$

so the cross-section for such collisions is

$$\sigma \sim b^2 \sim \frac{e^4}{(kT)^2}. \quad (24.4)$$

The mean free path is

$$l \sim \frac{1}{n\sigma} \sim \frac{k^2 T^2}{ne^4} \sim \frac{m^2 v_T^4}{ne^4} \quad (24.5)$$

where $v_T \sim \sqrt{kT/m_e}$ is the typical thermal velocity. The relaxation time is then

$$t_c \sim \frac{l}{v_T} \sim \frac{m^2 v_T^3}{ne^4} \simeq 0.92 \frac{T^{3/2}}{n} \text{ s} \quad (24.6)$$

where n is in units of cm^{-3} and T is in Kelvin.

For most astrophysical plasmas, t_c is much less than the plasma oscillation frequency, but is still short compared to most other time-scales such as the age of the system, so we expect the electrons to have relaxed to a Maxwellian velocity distribution.

24.1.3 Debye Length

Consider what happens if we introduce a positive ion into an electrically neutral plasma. The electrons will respond to the electric field of the ion, and will be attracted towards it, tending to counteract the positive charge. They will, in general, overshoot, and there will be some oscillations which radiate outwards. Once these have departed, the result will be an electrically neutral plasma with a very concentrated electron charge concentration surrounding, and just compensating, the electric field due to the ion. The time-scale for the plasma to adjust to the implanted ion is on the order of the inverse of the plasma frequency ω_p .

This is for a cold plasma. What happens if the electrons have a finite thermal velocity? Random thermal velocities will tend to smear out the electron charge concentration, but this will then reveal charge of the implanted ion, so there is a competition between these two effects. Now at a distance l from the ion, the time-scale for the velocities to act is $\tau \sim l/v_T$. Here $v_T \sim \sqrt{kT/m}$ is the typical thermal velocity. This ‘smearing time-scale’ $\tau(l)$ increases with l , so there will be some distance l_D such that $\tau(l) \sim 1/\omega_p$:

$$l_D \sim \frac{v_T}{\omega_p} \sim \sqrt{\frac{kT}{ne^2}}. \quad (24.7)$$

This length-scale (which will be defined more precisely below) is known as the *Debye length*. Its significance is as follows: at distances $l \gg l_D$ the plasma can relax towards neutrality faster than thermal velocities can smear out the electron charge concentration, so we expect that the ionic charge will be effectively screened by the electrons. On scales $l \ll l_D$, on the other hand, smearing by random motions is effective; the electron concentration is smoothed out, and the field due to the ion will not be screened.

Landau and Lifshitz give a more quantitative calculation of the screening: In equilibrium (and in the ensemble average state) we expect the surrounding electrons and ions to have a Boltzmann distribution, with the density of electrons, for example, being $n_e(r) = n_e e^{e\phi(r)/kT}$ where n_e is the

mean density and $\phi(r)$ is the potential due to the ion and the screening cloud. Similarly, the density of ions will be $n_i(r) = n_i e^{-Z_i e \phi(r)/kT}$. The total mean charge density profile will be

$$\rho(r) = Ze\delta(r) - n_e e e^{e\phi(r)/kT} + Z_i n_i e^{-Z_i e \phi(r)/kT} \quad (24.8)$$

where the potential $\phi(r)$ is related to $\rho(r)$ by Poisson's equation

$$\nabla^2 \phi = -4\pi\rho. \quad (24.9)$$

The problem here is to find a self-consistent solution to this pair of equations. In general this is difficult, but finding a solution at large radii is easier, since we then expect $e\phi(r) \ll kT$, so we can expand the exponentials to obtain

$$\rho(r) \simeq Ze\delta(r) - \phi(r)/L_D^2 \quad (24.10)$$

where we have invoked overall charge neutrality ($n_e = Z_i n_i$) and where we have defined the *Debye length* L_D such that

$$L_D^{-2} = \frac{e^2(n_e + Z_i n_i)}{kT}. \quad (24.11)$$

This is in accord with the hand-waving argument above. Poisson's equation, with the Laplacian written in spherically symmetric form, is

$$\frac{1}{r^2} \frac{dr^2 \frac{d\phi}{dr}}{dr} = L_D^{-2} \phi \quad (24.12)$$

and the boundary condition is $\phi \rightarrow Ze/r$ as $r \rightarrow 0$. This has solution

$$\phi(r) = \frac{Ze}{r} e^{-r/L_D}. \quad (24.13)$$

- The Debye length is the *screening length*; on scales $r \ll L_D$ the potential is that of the bare ion, while on larger scales the potential is exponentially suppressed.
- On scales $\gg L_D$ one can ignore the particulate nature of the plasma.
- The Debye length is on the order of the thermal velocity times the inverse of the plasma frequency. This is in accord with the idea that the plasma will tend to relax towards charge-neutrality on time-scale $1/\omega_p$ — resulting in an enhancement of electrons around the ion — but thermal motions will tend to smear this out.
- We have assumed here that the electrons are effectively collisionless.

24.2 Electromagnetic Waves in a Plasma

The mobile charges in a plasma have a profound influence on the propagation of electromagnetic waves. Here we will consider two applications; the dispersion of waves propagating through a plasma and the rotation of polarization which occurs if there is a large-scale magnetic field present. Both of these effects can be used as diagnostics to probe the properties of the medium through which the waves are propagating.

24.2.1 Dispersion in a Cold Plasma

If we consider waves with period much less than the relaxation time we can ignore collisions. Any electric field \mathbf{E} associated with the wave will drive a current in the electrons. The equation of motion for an electron is $\ddot{\mathbf{x}} = e\mathbf{E}/m$, so the current generated by the field obeys

$$\frac{d\mathbf{j}}{dt} = \frac{ne^2\mathbf{E}}{m}. \quad (24.14)$$

This is valid provided the period of the waves is much less than the relaxation time, as is usually the case.

If we assume a wave-like solution in which all properties vary as $e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ then the current is

$$\mathbf{j} = \frac{ine^2}{\omega m} \mathbf{E} = \frac{i\omega_p^2}{4\pi\omega} \mathbf{E}. \quad (24.15)$$

The third and fourth of Maxwell's equations become

$$\begin{aligned} i\mathbf{k} \times \mathbf{E} &= \frac{i\omega\mathbf{B}}{c} \\ i\mathbf{k} \times \mathbf{B} &= \frac{4\pi\mathbf{j}}{c} - \frac{i\omega\mathbf{E}}{c} = -i\frac{\omega^2 - \omega_p^2}{\omega c} \mathbf{E}. \end{aligned} \quad (24.16)$$

If we let the wave propagate along the z -axis, then on combining these equations we obtain the dispersion relation

$$\omega^2 = c^2 k^2 + \omega_p^2. \quad (24.17)$$

This is the same as the dispersion relation for a Klein-Gordon field with mass $m = \hbar\omega_p/c^2$. The presence of the modile sea of electron charge density has evidently endowed the electromagnetic waves with a non-zero 'effective mass'.

- For $\omega < \omega_p$ the wave-number becomes imaginary, and instead of traveling waves we have an evanescent decaying behavior $\mathbf{E} \propto e^{-kr}$.
- The current and electric field are 90 degrees out of phase, so $\langle \mathbf{j} \cdot \mathbf{E} \rangle = 0$. This means that no work is done by the wave on the plasma and so the energy is reflected.
- For $\omega > \omega_p$ we have traveling wave solutions, but the waves are is dispersive (see appendix D).
- The *group velocity* is $v_{\text{group}} = d\omega/dk = c^2 k/\omega = c\sqrt{1 - \omega_p^2/\omega^2}$. This is the speed at which information propagates, and tends to zero as ω approaches ω_p .
- If a source (a pulsar perhaps) emits pulses, then the pulses will be dispersed and we receive a 'chirp' of decreasing frequency (see problem).
- Measurement of the time of arrival vs frequency provides the *dispersion measure* $D \equiv \int dl n_e$.

24.2.2 Faraday Rotation

Plasmas are highly conductive, and so cannot sustain any large-scale electric fields. They are, however, often threaded by large scale magnetic fields. The presence of a magnetic field introduces a new frequency in the problem: the electrons will gyrate around the field lines at the *gyro-frequency*

$$\omega_G = \frac{eB}{mc} = 1.67 \times 10^7 \left(\frac{B}{\text{gauss}} \right) \text{s}^{-1}. \quad (24.18)$$

Just as with bulk oscillation of the plasma at the plasma frequency, we might expect the internal oscillations associated with the magnetic field to affect the oscillations of the electromagnetic field and thereby further modify the dispersion relation. This is indeed the case. We shall see that the presence of a magnetic field causes circularly polarized waves to propagate at different velocities, depending on their helicity. An important consequence of this is that the plane of polarization for linearly polarized waves rotates as it propagates through a magnetized plasma. This phenomenon is known as *Faraday rotation*, and provides an important diagnostic for magnetic fields.

Calculation of the dispersion relation is complicated in general since this will depend on the wave direction and on the polarization. Here we will consider for simplicity only waves propagating parallel to the field lines, which we shall take to lie along the z -axis.

Consider a circularly polarized wave propagating along the z -axis, such that the electric field lies in the $x - y$ plane and is

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = E_0 \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix}. \quad (24.19)$$

In the absence of any large-scale magnetic field, and assuming electron velocities $v \ll c$ so we can neglect the effect on the electron of the magnetic field of the wave, the acceleration of an electron is

$$\ddot{\mathbf{x}} = e\mathbf{E}/m \quad (24.20)$$

as before. This acceleration vector is the time derivative of a velocity

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{eE_0}{\omega m} \begin{bmatrix} \sin(\omega t) \\ -\cos(\omega t) \end{bmatrix} \quad (24.21)$$

which in turn is the derivative of a displacement

$$\begin{bmatrix} x \\ y \end{bmatrix} = -\frac{eE_0}{\omega^2 m} \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix}. \quad (24.22)$$

Thus each electron moves in a circular orbit of radius $r_0 = eE_0/\omega^2 m$ at velocity $v_0 = eE_0/\omega m$.

Now add a static magnetic field parallel to the z -axis. There will now be an additional $e\mathbf{v} \times \mathbf{B}/(mc)$ component to the acceleration. If we guess that the velocity is

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = v_0 \begin{bmatrix} \sin(\omega t) \\ -\cos(\omega t) \end{bmatrix} \quad (24.23)$$

(with v_0 to be determined) the extra component of the acceleration is

$$\frac{e\mathbf{v} \times \mathbf{B}}{mc} = \frac{eB}{mc} \begin{bmatrix} v_y \\ -v_x \end{bmatrix} = -\frac{ev_0 B}{mc} \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix}. \quad (24.24)$$

The magnetic force is therefore anti-parallel to the electric force. The equation of motion is

$$\ddot{\mathbf{x}} = \frac{e}{m} [\mathbf{E} + \mathbf{v} \times \mathbf{B}/c] \quad (24.25)$$

or

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \end{bmatrix} = \omega v_0 \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix} = \frac{e}{m} [E_0 - v_0 B/c] \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix}. \quad (24.26)$$

Our guess is then indeed a solution, provided

$$v_0 \left(\omega + \frac{eB}{mc} \right) = \frac{e}{m} E_0 \quad (24.27)$$

or equivalently

$$v_0 = \frac{eE_0}{m(\omega + \omega_G)}. \quad (24.28)$$

What is happening is that the magnetic force is opposing, and therefore hindering, the electric centripetal force, resulting in a velocity which is smaller by a factor $1/(1 + \omega_G/\omega)$ (see figure 24.2). For a magnetic field which is anti-parallel to the wave-vector, or equivalently for a wave of the opposite helicity (with the field, velocity and displacement rotating clockwise rather than anti-clockwise), the magnetic force adds to the centripetal acceleration due to the electric field, and the velocity is then larger by a factor $1/(1 - \omega_G/\omega)$.

The propagation of an electromagnetic wave of a certain frequency ω through a magnetized plasma is, for the case of \mathbf{B} parallel or anti-parallel to the wave-vector \mathbf{k} , identical to that for a non-magnetized plasma consisting of particles with the same number density and charge but with a slightly different mass ratio $m' = m(1 \pm \omega_G/\omega)$, or equivalently to a wave in a neutral plasma with a slightly different plasma frequency

$$\omega_p'^2 = \frac{\omega_p^2}{1 \pm \omega_G/\omega}. \quad (24.29)$$

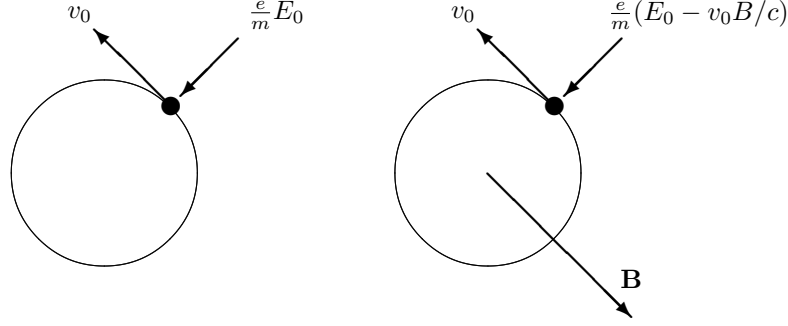


Figure 24.2: For a circularly polarized wave, the electrons confined to circular orbits by the centripetal acceleration eE_0/m and therefore move with speed $v_0 = eE_0/m\omega$ (left panel). If there is a magnetic field parallel (anti-parallel) to the wave vector the centripetal acceleration is partially annulled (enhanced) by a magnetic force, and the velocity is reduced (increased) by a factor $1/(1 \pm \omega_G/\omega)$.

The dispersion relation is therefore

$$c^2 k_{\pm}^2 = \omega^2 - \frac{\omega_p^2}{1 \pm \omega_G/\omega}. \quad (24.30)$$

This is the dispersion relation for circularly polarized waves. These are the ‘propagation eigenstates’ of a magnetized plasma. Here we are more interested in the propagation of radiation which is initially linearly polarized. We can obtain evolution as follows: We can decompose an initially linearly polarized wave into the sum of two circular waves of opposite helicity. These then propagate independently through some path length of plasma and we can then recombine them to obtain the final field.

For example, consider the wave

$$\begin{bmatrix} E_x(z, t) \\ E_y(z, t) \end{bmatrix} = \frac{E_0}{2} \begin{bmatrix} \cos(\omega t - k_+ z) \\ \sin(\omega t - k_+ z) \end{bmatrix} + \frac{E_0}{2} \begin{bmatrix} \cos(\omega t - k_- z) \\ -\sin(\omega t - k_- z) \end{bmatrix} \quad (24.31)$$

i.e. the sum of two circularly polarized waves. At $t = 0$, $z = 0$ the field is $(E_x, E_y) = (E_0, 0)$. Now write $k_{\pm} = \bar{k} \pm \Delta k$, to obtain

$$\begin{bmatrix} E_x(z, t) \\ E_y(z, t) \end{bmatrix} = E_0 \cos(\omega t - \bar{k}z) \begin{bmatrix} \cos z \Delta k \\ z \sin \Delta k \end{bmatrix} = E_0 \cos(\omega t - \bar{k}z) R(z \Delta k) \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (24.32)$$

Where $R(\theta)$ denotes the 2-D rotation matrix. For $\Delta k \ll \bar{k}$ this is a linearly polarized wave whose plane of polarization rotates progressively with distance $\theta(z) = z \Delta k$. More generally, if the plasma density and/or magnetic field (and therefore $\Delta k = (k_+ - k_-)/2$) varies with position the rotation angle is

$$\theta(z) = \frac{1}{2} \int dz (k_+ - k_-) \simeq \frac{1}{2c\omega^2} \int dz \omega_p^2 \omega_G \quad (24.33)$$

or equivalently

$$\theta = \frac{2\pi e^3}{m^2 c^2 \omega^2} \int dl nB. \quad (24.34)$$

The line integral appearing here is called the *rotation measure*. Now of course we usually do not know what the initial electric field orientation was, but the rotation angle depends on frequency, so by measuring the polarization angle as a function of frequency we can infer the rotation measure. Combining the rotation measure and the dispersion measure provides independent estimates of the large-scale field B and also the integral of the electron density.

24.3 Problems

24.3.1 Dispersion Measure

A distant pulsar emits short pulses which propagate to us through an intervening ionized medium.

- a) Sketch the ‘periodogram’ (power as a function of frequency and time) for the received signal.
- b) Show that the signal arrival time is related to frequency as

$$t \simeq \frac{d}{c} + (2c\omega^2)^{-1} \int dl \, \omega_p^2 \quad (24.35)$$

where ω_p is the plasma frequency.

- c) Under what conditions is the approximation above valid?

Part V

Gravity

Chapter 25

The Laws of Gravity

25.1 General Relativity

The best theory for gravity is Einstein's *general theory of relativity*. This is a classical theory of *geometrodynamics*, and has not yet been accommodated within a proper quantum mechanical framework.

In general relativity matter causes curvature of space-time, which in turn influences the orbits of particles etc. As Wheeler has colorfully put it, “space tells matter how to move — matter tells space how to curve”. Central to general relativity is the *metric tensor* $g_{\mu\nu}$ which is the generalization to curved space-times of the Minkowski metric $\eta_{\mu\nu}$ from special relativity. The basic equation of general relativity is an identity between the *curvature tensor* $G_{\mu\nu}$, which is constructed from second derivatives of the metric, and the *stress-energy tensor* $T_{\mu\nu}$ describing the matter.

In the *weak-field limit* of GR the metric is $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ where $h_{\mu\nu}$ is a small perturbation. If, in addition, the matter configuration is slowly varying (velocity of massive particles $\ll c$) then GR becomes identical to *Newtonian gravity*. Newtonian gravity provides a good description of most weak field phenomena with the exception of gravitational waves (these are usually weak fields, but require rapidly moving sources to be efficiently generated).

The bending of light by weak gravitational fields can also be ‘fudged’ simply by multiplying the Newtonian result for a test particle of velocity c by a factor two.

25.2 Newtonian Gravity

In Newton's theory the acceleration of a particle is the sum over all other particles of G times the mass times the inverse square of the distance.

$$\ddot{\mathbf{x}}_j = \sum_{i \neq j} \frac{G m_i (\mathbf{x}_i - \mathbf{x}_j)}{|\mathbf{x}_i - \mathbf{x}_j|^3} \quad (25.1)$$

where

$$G \simeq 6.67 \times 10^{-8} \text{cm}^3 \text{g}^{-1} \text{s}^{-2} \quad (25.2)$$

For a continuous density distribution this is

$$\ddot{\mathbf{x}} = \mathbf{g}(\mathbf{x}) = G \int d^3x' \rho(\mathbf{x}') \frac{(\mathbf{x}' - \mathbf{x})}{|\mathbf{x}' - \mathbf{x}|^3} \quad (25.3)$$

The gravity \mathbf{g} can be written as the gradient of the *gravitational potential* $\mathbf{g} = -\nabla\Phi$ where

$$\Phi(\mathbf{x}) = -G \int d^3x' \frac{\rho(\mathbf{x}')}{|\mathbf{x}' - \mathbf{x}|}. \quad (25.4)$$

Taking the gradient of $\Phi(\mathbf{x})$ one recovers (25.3).

In addition to the explicit formula for the potential as an spatial integral (25.4) there is also an equivalent local relationship between the Laplacian of the potential $\nabla^2\Phi$ and the density ρ

$$\nabla^2\Phi = 4\pi G\rho \quad (25.5)$$

which is *Poisson's equation*.

One way to establish the equivalence of (25.4) and (25.5) is to take the divergence of the gravity (25.3)

$$\nabla_{\mathbf{x}} \cdot \mathbf{g}(\mathbf{x}) = -\nabla^2\Phi = G \int d^3x' \rho(\mathbf{x}') \nabla_{\mathbf{x}} \cdot \left(\frac{\mathbf{x}' - \mathbf{x}}{|\mathbf{x}' - \mathbf{x}|^3} \right) \quad (25.6)$$

now the divergence $\nabla \cdot (\mathbf{x}/x^3)$ appearing here is readily computed and is found to vanish for $\mathbf{x} \neq 0$, so it must therefore be proportional to the Dirac δ -function. To obtain the constant of proportionality integrate $\nabla \cdot (\mathbf{x}/x^3)$ over a small sphere and use the divergence theorem to obtain $\int d^3x \nabla \cdot (\mathbf{x}/x^3) = \int \mathbf{dS} \cdot \mathbf{x}/x^3 = 4\pi$ so we have $\nabla \cdot (\mathbf{x}/x^3) = 4\pi\delta(\mathbf{x})$ and using this in (25.6) yields Poisson's equation (25.5).

Integrating (25.5) over some region gives

$$4\pi GM \equiv 4\pi G \int d^3x \rho(\mathbf{x}) = \int d^3x \nabla^2\Phi = \int \mathbf{dS} \cdot \nabla\Phi \quad (25.7)$$

so the integral of the normal component of the gravity over some closed surface is equal to $4\pi G$ times the mass enclosed. This is *Gauss' law*.

For a system of point masses the *gravitational binding energy* is defined as

$$W = \frac{1}{2} \sum_i \sum_{j \neq i} \frac{Gm_i m_j}{|\mathbf{x}_i - \mathbf{x}_j|} \quad (25.8)$$

the factor 1/2 arising because each pair is counted twice in this sum.

For a continuous distribution

$$W = \frac{1}{2} \int d^3x \rho(\mathbf{x}) \Phi(\mathbf{x}). \quad (25.9)$$

Note that 'continuous distribution' results can be used for point masses with $\rho(\mathbf{r}) \rightarrow \sum_i \delta(\mathbf{r} - \mathbf{r}_i)$.

25.3 Spherical Systems

25.3.1 Newton's Theorems

Newton found that

- $\mathbf{g} = 0$ inside a spherical shell of mass.
- The gravity outside such a shell is the same as for an equivalent mass at the origin.

These can be proved geometrically (see Binney and Tremaine), and they also follow directly from Gauss' law and spherical symmetry.

These theorems imply that the gravity $\mathbf{g}(\mathbf{r})$ for an arbitrary spherical system with cumulative mass profile $M(r)$ is

$$\mathbf{g} = -\frac{GM(r)}{r^2} \hat{\mathbf{r}}. \quad (25.10)$$

25.3.2 Circular and Escape Speed

The speed of a particle on a circular orbit satisfies

$$\frac{dv}{dt} = \frac{v^2}{r} = \frac{GM(r)}{r^2} \quad \rightarrow \quad v_{\text{circ}} = \sqrt{GM(r)/r} \quad (25.11)$$

The *escape speed* is

$$v_{\text{esc}} = \sqrt{2\Phi(r)} \quad (25.12)$$

where Φ is measured relative to its value at spatial infinity.

25.3.3 Useful Spherical Models

Point Mass

For a point mass $M(r) = m$

- The potential is $\Phi = -Gm/r$.
- The circular velocity is $v_{\text{circ}} = \sqrt{Gm/r}$.
- The escape velocity is $v_{\text{esc}} = \sqrt{2}v_{\text{circ}}$.
- The $v_{\text{circ}} \propto r^{-1/2}$ circular speed profile is usually referred to as a *Keplerian profile*.

Uniform Density Sphere

For a static uniform sphere of density ρ

- The gravity is

$$g = \frac{GM(r)}{r^2} = \frac{4\pi}{3}G\rho r. \quad (25.13)$$

- The circular speed is

$$v_{\text{circ}} = \sqrt{\frac{4\pi G\rho}{3}}r \quad (25.14)$$

- The *orbital period* is

$$t_{\text{orbit}} = \frac{2\pi r}{v_{\text{circ}}} = \sqrt{\frac{3\pi}{G\rho}} \quad (25.15)$$

which is independent of the radius of the orbit.

- The potential, measured with respect to the origin, is a parabola and the equation of motion for test particles within the sphere is

$$\ddot{\mathbf{r}} = -\frac{4\pi G\rho}{3}\mathbf{r}. \quad (25.16)$$

The period of *any* orbit in this potential is the same as that for a circular orbit.

- The *dynamical time* corresponding to density ρ is variously defined as the orbital time, the collapse time etc, but is always on the order of $t_{\text{dym}} = 1/\sqrt{G\rho}$.

All of the above properties are independent of the radius a of the sphere, and the dynamical and other times-scales are well defined in the limit that $a \rightarrow \infty$. The potential with respect to spatial infinity depends on the radius and is given by

$$\Phi(r) = \begin{cases} -2\pi G\rho \left(a^2 - \frac{1}{3}r^2\right) \\ -\frac{4\pi}{3}\frac{G\rho a^3}{r} \end{cases} \quad \text{for} \quad \begin{cases} r < a \\ r > a \end{cases} \quad (25.17)$$

Power Law Density Profile

A *power law density profile* $\rho(r) = \rho_0(r/r_0)^{-\alpha}$ has

- Mass $M(r) \propto r^{3-\alpha}$.
- We need $\alpha < 3$ if the mass at the origin is to be finite.
- The density cusp at the origin can be ‘softened’ as in the NFW models.
- A *flat rotation curve* results for $\alpha = 2$ and is referred to as a *singular isothermal sphere* profile.

Hernquist and NFW Models

Mass condensations that grow in cosmological simulations have been found to be quite well described by double power-law models.

The NFW model is

$$\rho(r) \propto \frac{1}{r(r^2 + r_c^2)} \quad (25.18)$$

which has asymptotic forms

$$\rho \propto \begin{cases} r^{-1} \\ r^{-3} \end{cases} \quad \text{for} \quad \begin{cases} r \ll r_c \\ r \gg r_c \end{cases} \quad (25.19)$$

Chapter 26

Collisionless Systems

26.1 Relaxation Time

The statistical mechanics of gravitating systems is very different from collisional gases. In the latter, long range electrostatic forces are screened by the neutrality of matter, whereas for the former the acceleration of particles is usually dominated by long range forces. For example, for a random distribution of point masses of mass m and number density n , the typical gravity from the nearest particle is $g \sim Gm/n^{2/3}$ while the overall gravity of the system is $g \sim GM_{\text{tot}}/R^2 \sim GmnR$ which is larger than the short-range force by a factor $n^{1/3}R \sim N^{1/3}$ where N is the number of particles in the system.

In many systems, the short-range gravitational accelerations are almost completely negligible, and the state of the system depends entirely on the way it was initially assembled.

To estimate the effect of *graininess* of the mass distribution on particle motions consider a collision with impact parameter b . This will give a transverse impulse on the order of the acceleration times the duration of the impact or

$$\delta v_{\perp} \sim \frac{Gm}{b^2} \frac{b}{v} = \frac{Gm}{bv}. \quad (26.1)$$

In crossing the system once the mean number of collisions with impact parameter in the interval b to $b + db$ is $dn \sim (N/R^2)bdb$. Now the mean vector sum of the impulses vanishes, but the *mean square* impulse accumulates and integrating over impact parameters gives the mean square impulse for one crossing of the system

$$\langle \Delta v_{\perp}^2 \rangle \sim \frac{N}{R^2} \int db b \left(\frac{Gm}{bv} \right)^2 \sim \left(\frac{Gm}{Rv} \right)^2 N \ln \Lambda \quad (26.2)$$

with $\Lambda = R/b_{\text{min}}$. The minimum impact parameter is $b_{\text{min}} = Gm/v^2$, and is the impact parameter that results in a large deflection $\delta v_{\perp} \sim v$.

Now $v^2 \sim GmN/R$ so $R/b_{\text{min}} \sim Rv^2/Gm = N$ and therefore in 1 crossing we have

$$\frac{\langle \Delta v_{\perp}^2 \rangle}{v^2} \simeq \frac{\ln N}{N} \quad (26.3)$$

which is typically much less than unity and the number of crossings required for the effect of short-range collisions to become important is $\sim N/\ln N$ and the *relaxation time* is

$$t_{\text{relax}} \sim \frac{N}{\ln N} t_{\text{orbit}}. \quad (26.4)$$

- In **galaxies** with $N_{\text{stars}} \sim 10^{11}$ and dynamical, or orbital, time $t_{\text{dyn}} \sim 10^8 \text{yr}$ the relaxation time is on the order of $t_{\text{relax}} \sim 10^{19} \text{yr}$ which greatly exceeds the age of the Universe which is $t \sim H^{-1} \sim 10^{10} \text{yr}$. This allows that the shapes of elliptical galaxies may be supported by anisotropic pressure, rather than by rotation, as was thought in olden times.

- In **globular clusters** with $N \sim 10^5$ and $t_{\text{dyn}} \sim 10^5 \text{yr}$ and the relaxation time is on the order of the Hubble time, so relaxations effects are important for such systems.
- In **galaxy clusters** with $N \sim 10^2$ and $t_{\text{dyn}} \sim 10^9 \text{yr}$ and the relaxation time is somewhat larger than the Hubble time, so relaxations effects are marginally effective for such systems.

26.2 Jeans Equations

On time-scales less than the relaxation time, a gravitating system of point masses is described by the collisionless Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{r}} f - \nabla \Phi \cdot \nabla_{\mathbf{v}} f = 0 \quad (26.5)$$

Taking the zeroth and first moments of this equation yields the equation of continuity

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad (26.6)$$

with $\mathbf{u} \equiv \langle \mathbf{v} \rangle$, and the Euler equation

$$\frac{\partial u_i}{\partial t} + (\mathbf{u} \cdot \nabla) u_i = -\frac{\partial \Phi}{\partial x_i} - \frac{1}{\rho} \frac{\partial (\rho \sigma_{ij}^2)}{\partial x_j} = 0 \quad (26.7)$$

where $\sigma_{ij}^2 \equiv \langle (v_i - u_i)(v_j - u_j) \rangle$ is the *velocity dispersion tensor*. In the steady state, the left hand side is $(\mathbf{u} \cdot \nabla) \mathbf{u}$ and is the centrifugal acceleration. Note that we do not split the *pressure tensor* $\rho \sigma_{ij}^2$ into isotropic and anisotropic parts as we did for a collisional gas.

Together with Poisson's equation (25.5) to relate the potential Φ to the mass density, (26.6) and (26.7) provide 5 equations, known as *Jeans' equations*. However, one has in general 10 unknowns (the density ρ , the three components of the streaming velocity \mathbf{u} and the six components of the symmetric 3×3 tensor σ_{ij}^2). In order to make progress it is necessary to make some assumption about the velocity dispersion tensor; common choices are to model σ_{ij}^2 as isotropic or, for a spherical system, to introduce an anisotropy parameter specifying the ratio of radial to tangential components of σ_{ij}^2 .

26.3 The Virial Theorem

Consider the moment of inertia of a system of point masses $I \equiv \sum m r^2$. The time derivative is $\dot{I} = 2 \sum m \mathbf{r} \cdot \dot{\mathbf{r}}$ and taking a further time derivative gives

$$\frac{1}{2} \ddot{I} = \sum m \dot{r}^2 + \sum m \mathbf{r} \cdot \ddot{\mathbf{r}}. \quad (26.8)$$

Requiring that \ddot{I} vanish for a stable system and expressing the acceleration $\ddot{\mathbf{r}}$ as a sum of the gravity from all the other particles gives

$$2T + \sum_{\mathbf{r}} m \mathbf{r} \cdot \sum_{\mathbf{r}' \neq \mathbf{r}} G m' \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} = 0 \quad (26.9)$$

with T the kinetic energy of the particles.

Now switching $\mathbf{r} \leftrightarrow \mathbf{r}'$ in the last term simply changes the sign, so we can write this as

$$\begin{aligned} & \sum_{\mathbf{r}} \sum_{\mathbf{r}' \neq \mathbf{r}} G m m' \mathbf{r} \cdot \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} \\ &= \frac{1}{2} \left[\sum_{\mathbf{r}} \sum_{\mathbf{r}' \neq \mathbf{r}} G m m' \mathbf{r} \cdot \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} - \sum_{\mathbf{r}} \sum_{\mathbf{r}' \neq \mathbf{r}} G m m' \mathbf{r}' \cdot \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} \right] \\ &= -\frac{1}{2} \sum_{\mathbf{r}} \sum_{\mathbf{r}' \neq \mathbf{r}} \frac{G m m'}{|\mathbf{r}' - \mathbf{r}|} = W \end{aligned} \quad (26.10)$$

and we therefore have the *virial theorem*

$$2T + W = 0. \quad (26.11)$$

The virial theorem provides a useful way to estimate the mass of bound systems. For example, if we assume equal mass particles, then the particle mass m is given by

$$m = \frac{2}{G} \frac{\sum \dot{r}^2}{\sum_{\text{pairs}} 1/r_{12}}. \quad (26.12)$$

For a roughly spherical system it is reasonable to assume that the 3-dimensional velocity dispersion in the numerator is 3 times the observed line of sight velocity dispersion. Similarly, the *mean harmonic radius* which appears in the denominator can be estimated from the observed distribution of projected separations, and this provides a useful way to determine the mass of a gravitating system.

The virial theorem gives the correct answer if the luminous particles trace the mass, but will fail if, for example, the dark matter has a different profile from the luminous particles.

26.4 Applications of the Virial Theorem

26.4.1 Spherical Collapse Model

Consider a uniform static sphere of dust of mass M and radius R_i . A perfectly symmetrical sphere will collapse to form a black hole, but this requires an enormous collapse factor, and any sensible amount of asphericity or initial angular momentum will cause the system to instead oscillate and eventually settle into some virialized final state. The initial energy is all potential $E_i = W_i \sim -GM^2/R$ and energy is conserved so we have $T_f + W_f = E_f = E_i$. But the virial theorem tells us that $2T_f + W_f = 0 \rightarrow T_f = -W_f/2$ and therefore $T_f + W_f = W_f/2 = W_i$ which means that the sphere must collapse by about a factor 2 in radius.

26.4.2 Galaxy Cluster Mass to Light Ratios

Photometric observations provide the *surface brightness* Σ_l of a cluster. On the other hand, observations of the velocity dispersion σ_v^2 together with the virial theorem give $\sigma_v^2 \simeq \frac{W}{M} \simeq \frac{GM}{R} \simeq G\Sigma_m R$ where Σ_m is the projected *mass density*. If the distance D to the cluster is known from its redshift then the *mass to light ratio* can be estimated as

$$\frac{M}{L} = \frac{\Sigma_m}{\Sigma_l} = \frac{\sigma_v^2}{GD\theta\Sigma_l} \quad (26.13)$$

where θ is the angular size of the cluster.

Applying this technique, Zwicky found that clusters have $M/L \sim 300M_\odot/L_\odot$ where subscript \odot indicates solar values.

26.4.3 Flat Rotation Curve Halos

More accurate masses are obtained for disk galaxies if the rotation curve can be measured (say from HI radio measurements). Once the measured velocity width has been corrected for inclination, the mass is given by

$$\frac{GM(r)}{r} = v_c^2(r). \quad (26.14)$$

Spiral galaxies are found to have rather flat rotation curves extending to at least a few tens of kpc (well beyond the radius where most of the visible stars reside) and this indicates that these galaxies have *dark halos* with $M \propto r$.

26.5 Masses from Kinematic Tracers

The virial theorem is exact, but requires that the light traces the mass. This is not a very good assumption. The Euler equation (26.7) can be used to constrain the gravitational potential of a system using kinematic data for particles which need not necessarily trace the mass.

For a static non-rotating system and assuming isotropic velocity dispersion temperature the Euler equation is

$$\nabla(n\sigma_{1D}^2) = -n\nabla\Phi \quad (26.15)$$

with n the number density of particles. This is just like the equation of hydrostatic equilibrium for a collisional system.

Now let's assume that the observed velocity dispersion happens to be *isothermal*, ie independent of radius $\partial\sigma^2/\partial r = 0$, and so

$$\sigma_{1D}^2 = \frac{nd\Phi/dr}{dn/dr} = \text{constant}. \quad (26.16)$$

If we further assume that the kinematic tracer density is a power law $n \propto r^{-\gamma}$ then this is consistent with a mass profile

$$M(r) = M_0(r/r_0) \quad (26.17)$$

since we then have $\nabla\Phi = GM/r^2 = GM_0/(r_0 r) \propto 1/r$.

With these assumptions, the observed velocity dispersion is related to the circular velocity $v_c^2 = GM_0/r_0$ by

$$\sigma_{1D}^2 = v_c^2 \left(-\frac{\partial \ln n}{\partial \ln r} \right)^{-1} \quad (26.18)$$

- If the tracers have the same profile as the mass ($n \propto 1/r^2$) then $v_c^2 = 2\sigma_{1D}^2$.
- If the tracers have a steeper (shallower) profile then the observed velocity dispersion will be lower (higher). This is not surprising since for shallower profiles the test particle orbits take them further up the potential.
- For gas and galaxies with similar profiles residing in the same potential well we expect the specific energy to be the same, or

$$\frac{1}{2}\sigma_{3D}^2 = \frac{(3/2)kT}{\mu m_p} \quad (26.19)$$

where μ is the mean molecular weight ($\mu = 1/2$ for fully ionized hydrogen for instance). This is a testable prediction which seems to be quite well obeyed. This result places constraints on possible long range interactions between *dark matter* particles since these would affect the galaxies but not the gas.

This was for a power-law tracer density profile. Another interesting model is a population of tracers with finite extent residing in a flat rotation curve extended halo. The velocity dispersion can then be found much as in the derivation of the virial theorem by considering the second derivative of the moment of inertia:

$$0 = \frac{1}{2}\ddot{I} = \sum \dot{r}^2 + \sum \mathbf{r} \cdot \ddot{\mathbf{r}} \quad (26.20)$$

but now with $\ddot{\mathbf{r}} = -GM_0/(rr_0)$ this gives $\sum \dot{r}^2 = (GM_0/r_0) \sum 1$ or equivalently

$$\sigma_{3D}^2 = \langle \dot{r}^2 \rangle = v_c^2 \quad (26.21)$$

as compared to $\sigma_{3D}^2 = (3/2)v_c^2$ for particles with $n \propto 1/r^2$ profile.

26.6 The Oort Limit

The Jeans equations can also be applied to a flattened disk geometry, in which case the vertical mass profile can be determined from measurements of densities and velocities of kinematic tracers oscillating up and down through the plane.

We start with the collisionless Boltzmann equation. For a planar geometry we have $\nabla_{\mathbf{r}} f = \hat{\mathbf{z}} \partial f / \partial z$ and $\nabla_{\mathbf{r}} \Phi = \hat{\mathbf{z}} \partial \Phi / \partial z$. Multiplying by v_z and integrating (with $\partial f / \partial t = 0$ as appropriate for a steady state solution) gives

$$\int d^3v v_z^2 \frac{\partial f}{\partial z} - \frac{\partial \Phi}{\partial z} \int d^3v v_z \frac{\partial f}{\partial z} = 0 \quad (26.22)$$

Integrating the second term by parts and dividing through by the number density n gives

$$\frac{1}{n} \frac{dn \langle v_z^2 \rangle}{\partial z} = - \frac{\partial \Phi}{\partial z} \quad (26.23)$$

but $\nabla^2 \Phi = \partial^2 \Phi / \partial z^2 = 4\pi G \rho$ and so

$$\rho = \frac{1}{4\pi G} \frac{\partial \frac{1}{n} \frac{dn \langle v_z^2 \rangle}{\partial z}}{\partial z} \quad (26.24)$$

where the quantities on the right hand side, which is, in essence, the second derivative of the kinetic pressure of the tracers, is observable given a collection of stars with well determined distances.

Applying this to stars in the solar neighborhood, Oort found

$$\rho \simeq 0.15 M_{\odot} \text{pc}^{-3} \quad (26.25)$$

which is known as the *Oort limit*.

Integrating the density gives the surface mass density

$$\Sigma(z) = \int_{-z}^z dz \rho(z) = \frac{1}{2\pi G n} \frac{dn \langle v_z^2 \rangle}{\partial z} \quad (26.26)$$

or, for $z = 700 \text{pc}$,

$$\Sigma(700 \text{pc}) \simeq 90 M_{\odot} \text{pc}^{-2}. \quad (26.27)$$

Comparing this to the surface luminosity density Σ_l is of some interest, since it can tell us if there is dark matter in the disk. The value of Σ_l has been hotly debated, but the estimates of the fraction of *missing matter* in the disk range from zero (Gilmore) to about 50% (Bahcall).

26.7 Problems

26.7.1 Two-body Relaxation.

Consider a virialised self-gravitating system of size R consisting of N identical particles of mass m (picture).

1. Give order of magnitude estimates of
 - (a) The long-range gravitational acceleration due to the system as a whole.
 - (b) The typical distance b from a particle to its nearest neighbor.
 - (c) The short-range gravitational acceleration due to the nearest neighbor.
 - (d) The ratio of long- to short-range force (in terms of N).
 - (e) The typical velocity v of a particle.

2. What is the typical transverse impulse Δv suffered by a particle moving at velocity v as it passes a neighboring particle at impact parameter b ?
3. How many such encounters does a particle suffer as it traverses the whole system once?
4. The mean impulse from many encounters averages to zero, but the mean square impulse accululates.
 - (a) What is the accululated mean square impulse in one crossing?
 - (b) How large is this compared to the mean square orbital velocity?
 - (c) At this rate, how many orbits or crossing times will it take for the particle's velocity to be significantly affected by nearest neighbor collisions.
5. How does this 'relaxation time' compare with the age of the Universe for
 - (a) A galaxy consisting of $\sim 10^{10}$ stars with a dynamical time of $\sim 10^8$ years?
 - (b) A globular cluster with $\sim 10^5$ stars and a dynamical time of $\sim 10^5$ years?

Chapter 27

Evolution of Gravitating Systems

27.1 Negative Specific Heats

A curious property of bound stable gravitating systems is that they have *negative specific heats*. The total energy is $E = K + W$ (with K here the kinetic energy) and the virial theorem is $W = -2K$ so the total energy is

$$E = -K. \quad (27.1)$$

The kinetic energy scales with the kinetic temperature T defined such that $K = \frac{1}{2} \sum mv^2 = \frac{3}{2} NkT$ and so the *specific heat* is

$$C = \frac{dE}{dT} = -\frac{3}{2} Nk \quad (27.2)$$

which is negative.

This leads to instability if gravitating systems are allowed to interact with other systems. As an example, consider two virialized clusters, both consisting of N equal mass particles, and the first with radius R_1 , kinetic energy K_1 , potential energy W_1 and similarly for cluster 2. The virial theorem tells us that each cluster satisfies $W = -2K$.

The total entropy for one of these clusters is $S = Nk(\ln T^{3/2} - \ln n)$ where n is the density. Now $n \propto 1/R^3 \propto (-W)^3$ and $T \propto K$, so the entropy for a cloud is

$$S = Nk \left(\frac{3}{2} \ln K - 3 \ln(-W) \right) + \text{constant}. \quad (27.3)$$

Now the virial theorem tells us that if the energy of the cluster changes, the changes in logarithms of the energies here are related by $\Delta \ln K = \Delta \ln(-W)$ and therefore the change in the entropy if the kinetic energy changes by ΔK is

$$\Delta S = -\frac{3}{2} Nk \frac{\Delta K}{K}. \quad (27.4)$$

This says that if a system gets hotter ($\Delta K > 0$) it loses entropy.

Now if some energy passes between the two systems such that $\Delta K_1 = \Delta K$ and $\Delta K_2 = -\Delta K$, then the change in the total entropy is

$$\Delta S_{\text{total}} = -\frac{3}{2} Nk \Delta K \left(\frac{1}{K_1} - \frac{1}{K_2} \right). \quad (27.5)$$

Thus if $K_1 > K_2$ — so system 1 has a higher kinetic temperature than system 2 — then for the change in the total entropy to be positive requires $\Delta K > 0$, which, since the total energy $E = -K$ is negative, corresponds, as usual, to a transfer of energy from the hotter system to the cooler, and, due to the negative specific heat, this results in the hotter system becoming still hotter and *vice versa*.

We have been rather vague about how the energy exchange actually takes place.

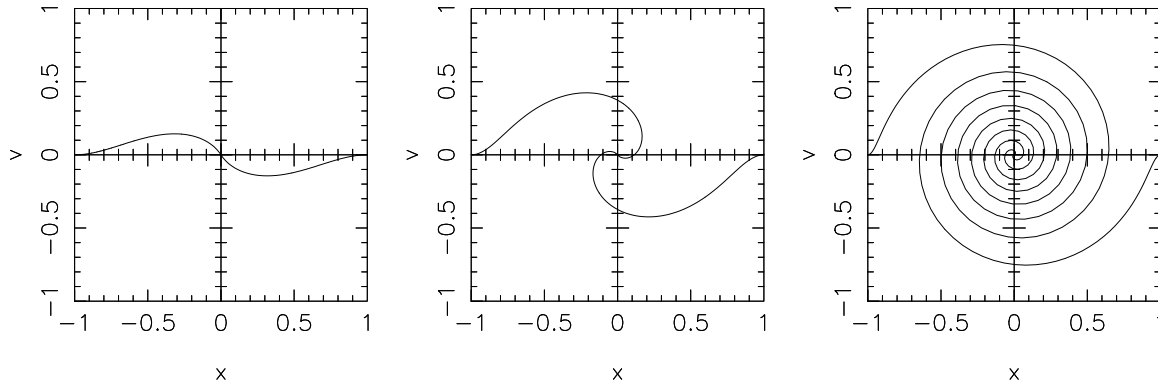


Figure 27.1: Illustration of phase-mixing for an initially cold gravitating system (such as cold-dark matter). The three panels illustrate the folding of the phase sheet as the system evolves.

- One manifestation of this instability is the possibility of runaway interactions between binary star systems and single stars within globular clusters. Here, if the orbital motion of the binaries exceeds the typical velocity of stars in the cluster then energy will be transferred from the binaries to the general star population, causing the binaries to become still hotter etc.
- A similar interaction is between stars of different mass within a cluster. It is not unreasonable to assume that initially the stars all have similar velocities. However, that means that the energy per particle, and therefore also the kinetic temperature, scales with the mass, and so collisions will tend to transfer energy from the more massive to the lighter stars. This leads to concentration of massive stars in the core and the lighter stars to be expelled, or scattered to large radius orbits.
- Another example is a *hierarchical* system with bound gravitating sub-units orbiting within a large system (much like globular clusters in a galaxy say, or galaxies within a cluster of galaxies), where the energy transfer is mediated by tidal forces and inelasticity of collisions between the clusters. Entropic considerations here say that interactions which increase entropy are those where energy is transferred from hotter to colder bodies, where temperature is defined so that kT is the kinetic energy of a ‘meta-particle’. This usually means transfer of energy from large-scales to smaller scales (for example, a globular cluster orbiting in a galaxy has much more energy than a single star in a globular cluster). This causes the ‘hot’ system (the cluster orbits within the halo) to give energy to the sub-units (the stars in the clusters themselves). This causes the hot system to heat up and the cold system to cool further until eventually the clusters will become unbound.

Negative specific heats lead to a *gravo-thermal catastrophe*. There is no hope of finding a true equilibrium solution like the Maxwellian for a collisional gas.

27.2 Phase Mixing

Liouville’s theorem tells us that the phase-space density is constant along particle orbits. If the dark matter is initially ‘cold’ (ie it has negligible thermal or random velocities) then it occupies only a 3-dimensional subspace $v_x = v_y = v_z = 0$ of phase-space, and it will remain that way forever. Initially, this *phase sheet* is flat, but as structures start to form the particles will accelerate and the velocity at a given position will deviate from zero. Eventually, particles will fall into potential wells and oscillate, with the result that the phase sheet gets wrapped up, as illustrated in 1-spatial dimension is figure 27.1.

Once the phase-sheets has folded over on itself, there will no longer be a single velocity at each point in space, rather there will be a set of discrete velocities which are populated. Our galaxy is

about 50 dynamical times old, so one might expect there to be about 50 distinct ‘streams’ of dark matter passing through our neighborhood.

While the true phase space density does not evolve, this wrapping of phase sheets will cause the *coarse-grained phase-space density* to decrease.

27.3 Violent Relaxation

In a stable potential particles orbit at fixed energy and $dE/dt = 0$. When a gravitating system is initially forming (perhaps from the merger of sub-units) the potential will be rapidly fluctuating, and the particle energy will change as $dE/dt = m\partial\Phi/\partial t$. This process will tend to generate a distribution in which the velocities of particles are independent of the particle mass, as was originally pointed out by Lynden-Bell.

27.4 Dynamical Friction

A heavy object orbiting in a system consisting of predominantly lighter particles will tend to sink towards the center because of *dynamical friction*. The calculation of the dynamical friction force is similar to the calculation of the relaxation time and is also very similar indeed to the calculation of bremsstrahlung radiation.

There are two ways of looking at this. One picture is that the massive particle will induce a wake in the lighter particles, and the excess density in the wake exerts a gravitational pull on the heavy article. Another view is that a massive particle passing by causes an impulse in the ‘background’ particles, and transfers to them kinetic energy which must come eventually from the kinetic energy of the massive particle. We will follow the second route.

Let a heavy particle of mass M move at velocity V through a background sea of light particles of mass m . As it passes by, a background particle at impact parameter b is given a velocity impulse $\delta v = (GM/b^2) \times (b/V)$ and a corresponding kinetic energy $\Delta E = m(\delta v)^2/2 \sim m(GM/bV)^2$. Note the rate of such collisions is $dN/dt = nVb db$ with n the density of background particles. The deceleration of the massive particle is then

$$\frac{dV}{dt} = \frac{1}{MV} \frac{dE}{dt} = \frac{m}{MV} \int db bnV \left(\frac{GM}{bV} \right)^2. \quad (27.6)$$

As before, this gives a logarithmic integral and we have

$$\frac{dV}{dt} \simeq \frac{G^2 M m n}{V^2} \ln \Lambda \quad (27.7)$$

where as before the factor $\ln \Lambda$ is the ratio of the size of the system to the minimum impact parameter b_{\min} , and is rather insensitive to the precise value of b_{\min} .

- The acceleration is proportional to the mass of the heavy particle, so the force is proportional to M^2 . This can also be obtained from considering the wake — the amplitude of the density fluctuation in the wake is proportional to M .
- The acceleration is inversely proportional to the inverse square of the massive particle velocity.
- The analysis here is oversimplified in that the velocity of the background particles was not taken into account. A more careful analysis shows, not surprisingly, that the energy transferred to rapidly moving background particles is less than if they are stationary.
- Applications include decay of globular cluster orbits in a massive galactic halo. For $M \sim 10^6 M_\odot$ and $V \sim 250 \text{ km/s}$ the dynamical friction time becomes comparable to the age of the universe at about 1kpc. The Magellanic clouds are more distant ($\sim 50 \text{ kpc}$) but much more massive ($M \sim 10^{10} M_\odot$) and again the friction time-scale is on the order of the Hubble time.
- Dynamical friction is may play a role in building giant ‘cD’ galaxies in the centers of clusters.

27.5 Collisions Between Galaxies

Galaxies will occasionally collide in clusters. However, since the orbital velocity in the cluster is typically on the order of 1000 km/s which is considerably greater than the internal rotation velocity (typically 200 km/s) these collisions have little effect on the galaxies since the galaxies just feel a short impulsive force and have no time to respond.

In poor clusters and groups the relative collision velocity is close to the galaxy internal velocities and such collisions would be inelastic and would lead to merging of galaxies.

The question whether elliptical galaxies form from merging of spirals has been hotly debated. On one side, there are certainly examples of tidal tails indicating ongoing merging systems. On the other, there is evidence that ellipticals are very old and can be found in a mature state at redshifts $z \sim 1$, so if they did form by merging they did so a long time ago.

27.6 Tidal Stripping

Another evolutionary effect in gravitating systems is tidal stripping. The outer parts of galaxies in clusters may be stripped in this way. The *tidal radius* is the radius at which the density is equal to the mean density of the parent body.

Part VI

Cosmology

Chapter 28

Friedmann-Robertson-Walker Models

28.1 Newtonian Cosmology

Consider a small uniform density uniformly expanding sphere of pressure-free dust of radius a and density ρ . Gauss' law tells us that the acceleration of particles at the edge of the sphere is $\ddot{a} = -GM/a^2$, or, with $M = \frac{4}{3}\pi\rho a^3$,

$$\ddot{a} = -\frac{4}{3}\pi G\rho a. \quad (28.1)$$

This *acceleration equation* can be integrated by multiplying by $2\dot{a}$, which gives

$$2\dot{a}\ddot{a} = \frac{d\dot{a}^2}{dt} = -\frac{2GM\dot{a}}{a^2} \quad (28.2)$$

so

$$\dot{a}^2 = -\int dt \frac{2GM\dot{a}}{a^2} = -\int da \frac{2GM}{a^2} = \frac{2GM}{a} + \text{constant}. \quad (28.3)$$

With $M = \frac{4}{3}\pi\rho a^3$ again, this gives the *energy equation*

$$\dot{a}^2 = \frac{8}{3}\pi G\rho a^2 + 2E_0 \quad (28.4)$$

where E_0 is constant. This equation expresses conservation of energy, the term on the left hand side being proportional to the kinetic energy and the first term on the right hand side being the proportional to potential energy. More precisely, E_0 is the total energy per unit mass — the specific energy, that is — of test particles on the boundary of the sphere.

The equivalence of (28.1) and (28.4) could have been established by differentiating the latter. This gives

$$2\dot{a}\ddot{a} = \frac{8}{3}\pi G(\dot{\rho}a^2 + 2\rho\dot{a}a) \quad (28.5)$$

but with the continuity equation, which in this context is $\rho a^3 = \text{constant}$, or

$$\dot{\rho} = -3\frac{\dot{a}}{a}\rho \quad (28.6)$$

we recover (28.1).

Note that both sides of (28.1) are linear in the dust-ball radius a , so if $a(t)$ is a solution then so is $a' = \alpha a$ for an arbitrary constant α . This means that, if the density is initially uniform, all of the shells of the dust cloud evolve in the same way, and the density will remain uniform.

It is interesting to contrast the properties of a uniform density gravitating cloud to a sphere of uniform *electrical charge density*. In the latter case, the electric field at the surface of the sphere

grows linearly with the size of the sphere. Since the electric field is an observable quantity, this means that there is no sensible solution as $a \rightarrow \infty$, it also means that an observer within a finite sphere can always determine the direction to the center of the sphere from local measurements of \mathbf{E} . The analogous quantity in the gravitating sphere is the *gravity* vector \mathbf{g} , which also diverges linearly with the sphere radius etc. However, the gravity itself is not directly observable — this is fundamentally because of the *equivalence principle*; all particles accelerate alike in a gravitational field — and the only directly observable property of the gravity is its spatial gradient, which is the *tidal field tensor* $\partial^2\Phi/\partial x_i\partial x_j$. Since the gravity is proportional to radius this means that the tide is spatially uniform and perfectly regular in the limit $a \rightarrow \infty$. Also it means that an observer cannot determine the direction to the center from local measurements; each observer simply sees locally isotropic expansion. Newtonian gravity can therefore model an expanding pressure-free cosmology of arbitrary size.

Newton, Laplace and contemporaries were of course unaware that we live in a galaxy surrounded by a seemingly infinite sea of other galaxies which, on large scales, are apparently uniformly distributed. Nor were they aware that these galaxies are receding from us according to Hubble's law. Had they been privy to this information, they would have had no difficulty concocting the physical model above which is used by all practicing cosmologists today in interpreting their observations.

The system of equations above are also valid in general relativity, as originally shown by Friedmann and by Robertson and Walker. This is because Newtonian gravity provides the correct description of a finite sphere in the limit $a \rightarrow 0$, and because of *Birchoff's theorem*, which is the relativistic equivalent of Gauss' law, and which says that for a spherically symmetric mass distribution, the gravity within some shell is independent of the matter distribution outside. We will refer to these models as the FRW models.

28.2 Solution of the Energy Equation

There is an analytic parametric solution of the energy equation:

$$\begin{aligned} a(\eta) &= A(1 - \cos \eta) \\ t(\eta) &= B(\eta - \sin \eta) \end{aligned} \quad (28.7)$$

from which it follows that the expansion velocity is

$$\dot{a} \equiv \frac{da}{dt} = \frac{da/d\eta}{dt/d\eta} = \frac{A}{B} \frac{\sin \eta}{1 - \cos \eta} \quad (28.8)$$

Substituting these in (28.4) one finds that this equation is satisfied if we choose the constants A , B to be

$$\begin{aligned} A &= GM/2|E_0| \\ B &= GM/(2|E_0|)^{3/2} \end{aligned} \quad (28.9)$$

where $M = \frac{4}{3}\pi\rho a^3$.

These are the solutions for energy constant $E_0 < 0$, corresponding to a gravitationally bound system which first expands but then turns around and collapses back to zero radius. The total time from big-bang to big-crunch in these models is $2\pi B$. A family of such solutions are shown as the lower solid curves in figures 28.1 and 28.2.

There are analogous solutions for the case of $E_0 > 0$ where A and B are still given by (28.9) but where the trigonometric functions are replaced by their hyperbolic equivalents:

$$\begin{aligned} a(\eta) &= A(\cosh \eta - 1) \\ t(\eta) &= B(\sinh \eta - \eta) \end{aligned} \quad (28.10)$$

Such solutions are unbound and expand forever. The constant B in this case is the time at which the kinetic and potential energies become comparable to the total energy. For $t \ll B$ the kinetic and potential energies are much larger, in modulus, than the total energy, while for $t \gg B$ the potential energy becomes negligible and the solutions become freely expanding with $a \propto t$.

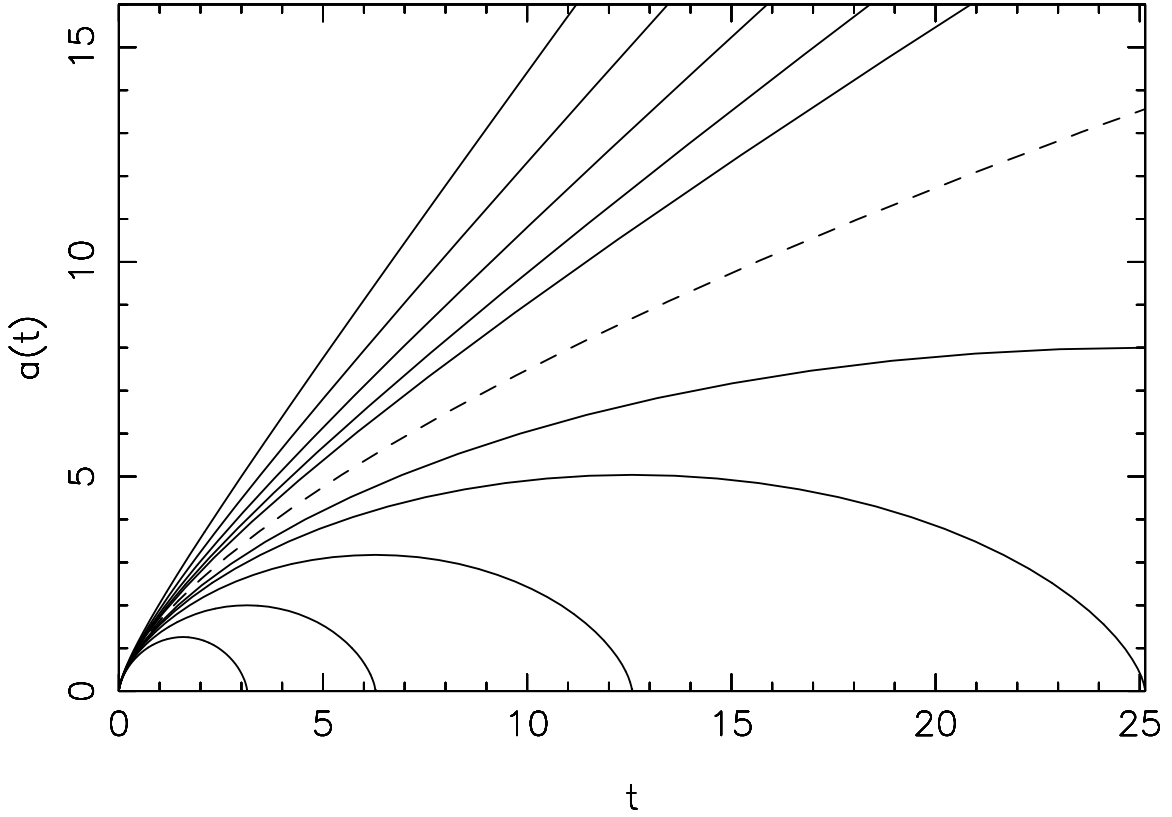


Figure 28.1: Scale factor a vs time for a family of FRW models. The quantities plotted in the lower set of solid curves are $t = B(\eta - \sin \eta)$ and $a = (B)^{2/3}(1 - \cos \eta)$ for $B = 0.5, 1.0, 2.0, 4.0, 8.0$. The quantities plotted in the upper set of solid curves are $t = B(\sinh \eta - \eta)$ and $a = (B)^{2/3}(\cosh \eta - 1)$ for the same set of B values. These curves are the time evolution of the radii of spheres of the same mass for various energies. The lower/upper curves have negative/positive total energy. All of these solutions become identical at early times (this is because the kinetic and potential energies are nearly equal but opposite and much larger in modulus than the total energy). The dashed curve is the marginally bound case with zero total energy. This can be considered to be the limiting case of either open or closed models as $B \rightarrow \infty$.

A family of such solutions are shown as the upper set of solid curves in figures 28.1 and 28.2. The negative/positive total energy solutions are referred to as ‘open’ and ‘closed’ respectively. Here we see that the open models are open-ended in time. The real reason for this choice of terminology will become apparent later where we will see that the spatial geometries corresponding to these models are open and closed respectively. The marginally bound solution is commonly referred to as the *Einstein - de Sitter model*.

Some useful *cosmological parameters* are

- The *Hubble parameter* or *expansion rate* is defined to be

$$H = \dot{a}/a \quad (28.11)$$

and has units of inverse time. Its current value is measurable locally from the ratio of recession velocity to distance and is $H_0 \simeq 75 \text{ km/s/Mpc}$. It is generally very similar to the inverse age of the Universe.

- The *critical density* is defined as

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G}. \quad (28.12)$$

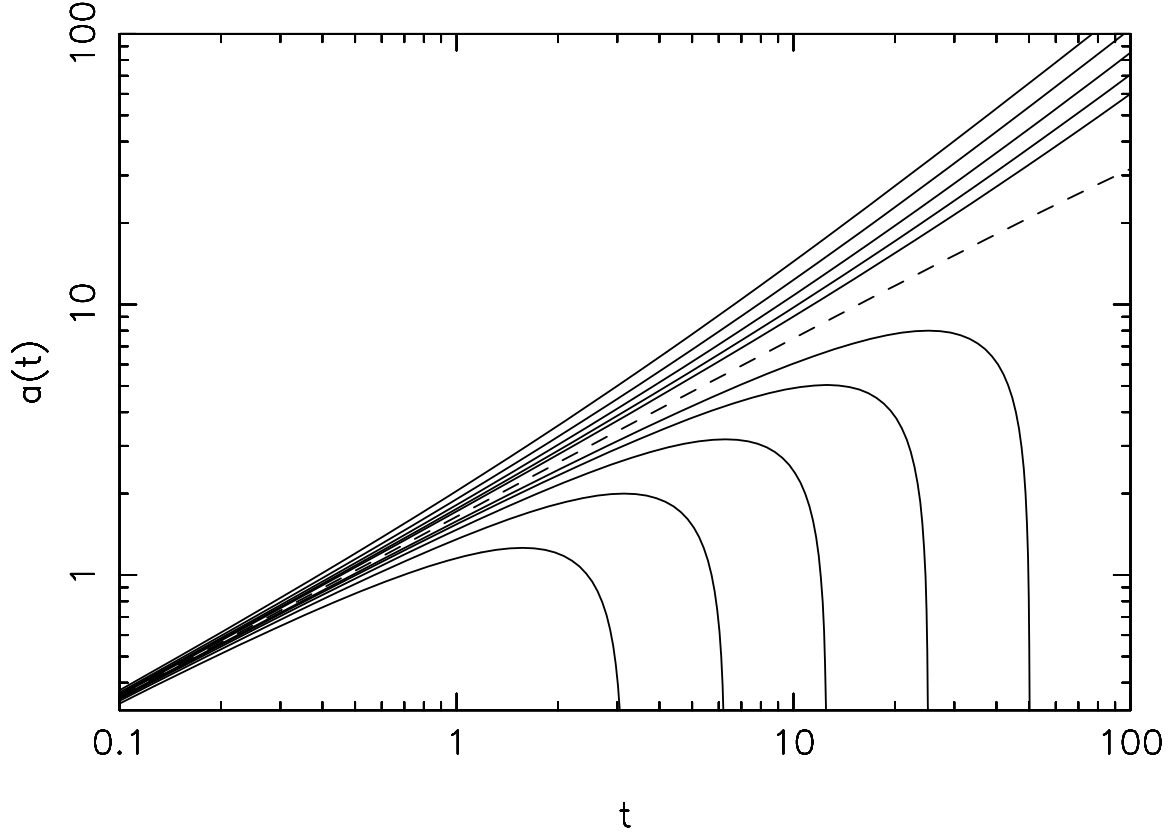


Figure 28.2: As figure 28.1 but plotted on logarithmic scales. This makes clearer that all models have the same power-law form, $a \propto t^{2/3}$, for early times $t \ll B$. The open model solutions become linear, $a \propto t$ for $t \gg B$.

For a given expansion rate H , (28.12) gives the density required so that the potential energy just balances the kinetic energy. Dividing (28.4) by $8\pi G a^2/3$ gives

$$\rho_{\text{crit}} = \rho + 3E_0/4\pi G a^2 \quad (28.13)$$

so if $\rho > \rho_{\text{crit}}$ we must have $E_0 < 0$ and the universe must be bound and *vice versa*.

- The *cosmological density parameter* is

$$\Omega = \frac{\rho}{\rho_{\text{crit}}}. \quad (28.14)$$

The density parameter is close to unity at early times in all models (and also in the *big crunch* for closed models) but tends to diverge strongly from the critical density solution if Ω is not exactly unity (see figure 28.3).

28.3 Asymptotic Behavior

In the energy equation (28.4) the potential energy term scales as $1/a$ which means that the kinetic energy must also scale as $1/a$. Regardless of the current value of the energy constant E_0 , if we go back to sufficiently early times the two time varying terms will completely dominate and the constant of energy will be negligible. Taking $E_0 \rightarrow 0$ the energy equation becomes

$$\dot{a}^2 = 2GM/a. \quad (28.15)$$

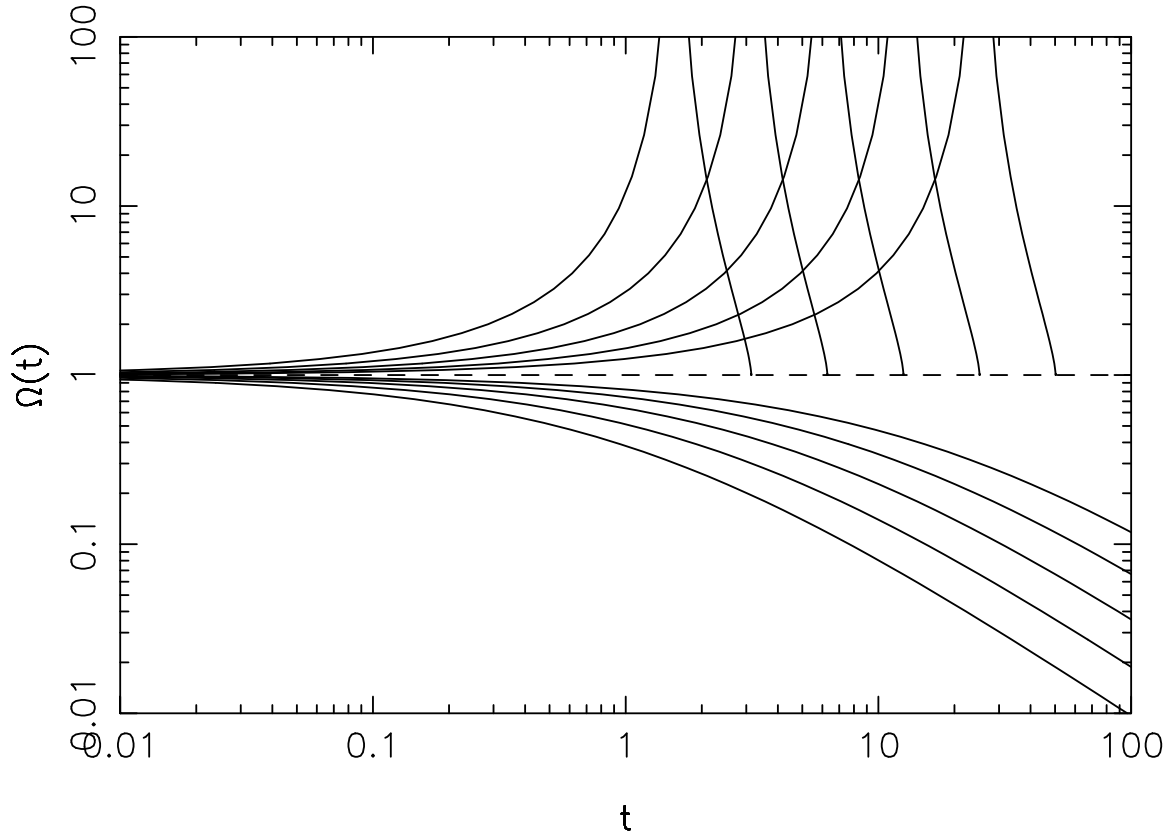


Figure 28.3: Evolution of the cosmological density parameter Ω vs time in matter dominated FRW models. As before, $t = B(\eta - \sin \eta)$ for the upper curves (closed models) for $B = 0.5, 1.0, 2.0, 4.0, 8.0$, and $t = B(\sinh \eta - \eta)$ for the lower curves (open models). The density parameter is $\Omega = 2(1 - \cos \eta)/\sin^2 \eta$ for the closed models and $\Omega = 2(\cosh \eta - 1)/\sinh^2 \eta$ for the open models.

If we postulate a power-law solution $a = a_0(t/t_0)^\alpha$ then $\dot{a} = \alpha a/t$, so the energy equation says $\alpha^2 a^2/t^2 = 2GM/a \propto 1/a$, or

$$a \propto t^{2/3}. \quad (28.16)$$

This behavior can also be inferred from the parametric solutions in the limit $\eta \ll 1$. At early times $\rho \rightarrow \rho_{\text{crit}}$ and therefore $\Omega \rightarrow 1$.

The future asymptotic behavior depends critically on the energy constant. For $E_0 < 0$ the Universe will reach a maximum size and will then re-collapse and the late time behavior will be the reverse of the initial expansion. If $E_0 > 0$ then at some point the kinetic and potential terms will become comparable to E_0 , at which point Ω will start to deviate appreciably from unity. For late times $\dot{a}^2 \rightarrow 2E_0$, corresponding to undecelerated expansion. The critical density then scales as $\rho_{\text{crit}} \propto H^2 = (\dot{a}/a)^2 \propto 1/a^2$ while the actual density scales as $\rho \propto 1/a^3$ and so $\Omega \rightarrow 0$ as $\Omega \propto 1/t$ in the limit $t \rightarrow \infty$.

28.4 The Density Parameter

Estimates of the local expansion rate H_0 tell us the current critical density ρ_{crit} . Redshift surveys tell us the local *luminosity density* of the Universe \mathcal{L} . If we multiply this by the mass-to-light ratio M/L determined from virial analysis of clusters of galaxies we find that the density of matter is around 0.2 – 0.3 times critical, and therefore that the density parameter $\Omega_{\text{crit}} \simeq 0.2 - 0.3$.

There is of course considerable slop in the dynamical mass measurements, and also in the implicit

assumption that the M/L of clusters is representative, but there is a growing consensus that this is about the correct value, and almost unanimous agreement that $\Omega < 1$. Also, estimates of the *acceleration* of the universe from supernovae studies also indicate a low matter density. This result is remarkable for two reasons: Ω is very close to unity and yet it is apparently not exactly equal to unity. If this estimate is correct it says that we live at a rather special time when matter has just stopped dominating the expansion of the Universe. Why this should be remains a mystery.

28.5 The Cosmological Redshift

Consider a photon which is emitted by the observer at the center of an expanding sphere of mass M and radius a and received by an observer on the surface. Let us assume that the radius of the sphere is small enough that $\dot{a} \ll c$. The received wavelength will then be greater than that emitted because of the Doppler effect, and we therefore have a redshift

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = 1 + \dot{a}/c. \quad (28.17)$$

The fractional change in wavelength is $\Delta\lambda/\lambda = \dot{a}/c$. Any gravitational redshift effect is $\Delta\lambda/\lambda \sim GM/ac^2$ and is negligible compared to the Doppler effect since $GM/ac^2 \sim (\dot{a}/c)^2 \ll \dot{a}/c$.

Now consider the change in the size of the sphere in the time $\Delta t = a/c$ that it takes for the photon to make this trip. The ratio of sphere sizes is

$$\frac{a_{\text{obs}}}{a_{\text{em}}} = \frac{a + \dot{a}\Delta t}{a} = 1 + \frac{\dot{a}\Delta t}{a} = 1 + \dot{a}/c. \quad (28.18)$$

Thus, for a small sphere, the fractional change in the wavelength of light is equal to the fractional change in the sphere radius. If we have a photon which travels a large distance through a FRW model then we can compute the net change in the wavelength by multiplying the effect due to a series of infinitesimal steps. The result is that

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{a_{\text{obs}}}{a_{\text{em}}} \quad (28.19)$$

i.e. the wavelength grows in proportional to the scale factor.

28.6 The Horizon Problem

The FRW equations provide a self-consistent mathematical model for a large, perhaps infinite, uniform density expanding Universe. However, it has a rather peculiar feature; at early times the distance over which light, and therefore any other causal influence, can propagate shrinks to zero, and moreover it shrinks faster than the Universe itself.

To analyze this it is useful to introduce the concept of *comoving coordinates* \mathbf{r} which are related to *physical coordinates* by

$$\mathbf{x} = a(t)\mathbf{r} \quad (28.20)$$

Thus the surface of our dust sphere has comoving radius $r = 1$, and the dust particles do not move in comoving coordinates. Now the physical distance that a photon can travel in time interval dt is just $dx = cdt$, corresponding to a change in comoving coordinate $dr = cdt/a(t)$. The integrated comoving distance that a photon can travel since $t = 0$ is called the *horizon* r_h and is

$$r_h(t) = c \int_0^t \frac{dt}{a(t)} \propto \int_0^t \frac{dt}{t^{2/3}} \propto t^{1/3} \quad (28.21)$$

which tends to zero for $t \rightarrow 0$.

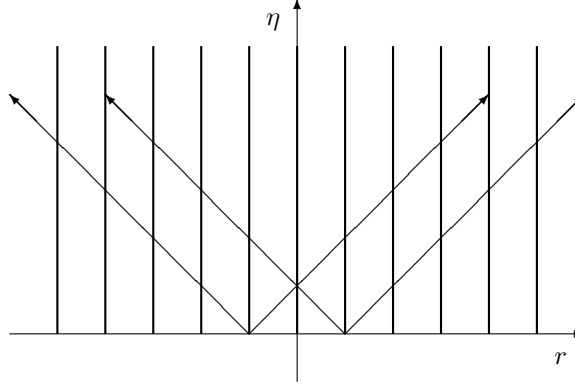


Figure 28.4: The causal structure of FRW models is most clearly shown if we plot particle world-lines in $\eta-r$ coordinates (i.e. conformal time *vs* comoving spatial coordinate). The vertical lines represent comoving observers, while the diagonal lines are light rays. Causal influences are constrained to lie within the light-cones. Since conformal time $\eta \rightarrow 0$ at the big-bang, this means that the initial singularity (the line $\eta = 0$) is acausal, in the sense that comoving observers are initially out of causal contact with each other. If, however, the Universe *accelerates* at early times, the initial singularity is pushed back to $\eta = -\infty$ in this plot, and the big-bang is then causal.

A useful way to visualize the causal structure of the FRW model is to work in *conformal time* η which satisfies $d\eta = cdt/a(t)$. Since photons have $dr = cdt/a(t)$ this means that if we plot world-lines in $\eta-r$ space then comoving observers are lines of constant r while photons are lines at ± 45 degrees (hence the terminology ‘conformal time’). This is shown in figure 28.4. This shows that comoving observers which can now exchange information and physically influence each other were, at early times, causally disconnected.

The problematic nature of the horizon is most clearly appreciated if we consider the photons of the microwave background radiation. These photons have been propagating freely since they were last scattered at a redshift of $z_{\text{dec}} \simeq 10^3$. At that time, the horizon was smaller than the present value by a factor $(t_0/t_{\text{dec}})^{1/3} = (a_0/a_{\text{dec}})^{1/2} = z_{\text{dec}}^{1/2} \simeq 30$. Consider photons arriving from opposite directions (see figure 28.5). The regions which last scattered the radiation now arriving from these directions were not able to causally influence each other at the time the radiation was scattered — indeed they are not causally connected even today — yet the radiation we see is, aside from the dipole anisotropy due to our motion, isotropic to a few parts in 10^5 . The standard FRW model provides no explanation for how this degree of isotropy was established; it must be put in as initial conditions.

The horizon problem is fundamentally due to the fact that the universal expansion is decelerating: $\ddot{a} < 0$. If we consider a pair of neighboring observers who are currently receding from each other with some velocity $v \ll c$, then at earlier times this recession velocity was larger and, at some sufficiently early time, exceeded the speed of light. Before that time these observers were causally disconnected. If one had instead a power-law expansion $a \propto t^\alpha$ with $\alpha > 1$ then $\ddot{a} > 0$ and the comoving horizon scale diverges as $t \rightarrow \infty$ and there is no horizon problem.

28.7 Cosmology with Pressure

So far we have considered pressure-free dust. What happens if this dust is actually a dust of bombs all primed to explode at a certain time? At that instant, the equation of state changes from $P = 0$ to $P \neq 0$ as the shrapnel, radiation etc from the explosions will have positive pressure.

Naively, one might imagine that pressure would tend to counteract gravity and would cause the universal expansion to decelerate less rapidly, but this is incorrect. It is true that for a *finite* sphere, such an explosion would accelerate the outer layers, and this acceleration would work its way in, but this acceleration is caused by pressure *gradients* which only arise by virtue of there being an edge to

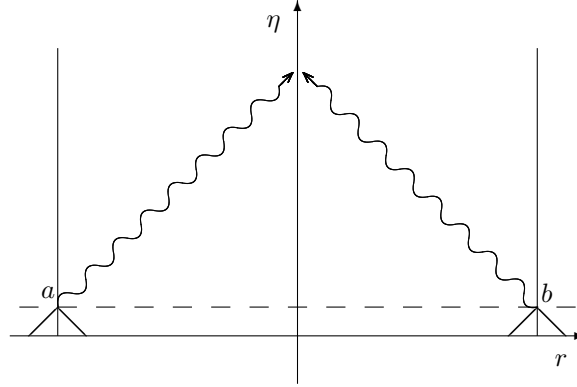


Figure 28.5: The wiggly diagonal lines in this conformal space-time plot show the world lines of two photons which we now see arriving from opposite directions, and which were last scattered at points a , b . The horizontal dashed line indicates the moment of recombination of the plasma. Also shown are the past light cones of the events a , b . The regions which can have causally influenced the photons prior to their departure are very small compared to the separation of the scattering events.

the sphere. Deep within a very large sphere, there is no pressure gradient (at least for short times after the explosions), so there are no ∇P forces acting.

Recall that for $P = 0$ the equations governing the scale factor of the universe and the density parameter are the acceleration equation

$$\ddot{a} = -\frac{4}{3}\pi G\rho a \quad (28.22)$$

the continuity equation

$$\dot{\rho} = -3\frac{\dot{a}}{a}\rho \quad (28.23)$$

and the energy equation

$$\dot{a}^2 = \frac{8}{3}\pi G\rho a^2 + 2E_0 \quad (28.24)$$

with E_0 constant. These equations are not independent; any one can be derived from the other two. They provide two independent equations which determine the two functions $a(t)$ and $\rho(t)$, the initial conditions being set by the energy constant E_0 and also the time of the big-bang.

One consequence of non-zero pressure is to modify the continuity equation. For $P = 0$ this is the statement that the mass within a comoving sphere is constant: $\rho a^3 = \text{constant}$. If $P \neq 0$ this is not correct. With $P > 0$, any volume of the Universe does work on its surroundings as it expands, and since mass and energy are equivalent according to Einstein the mass is not conserved. Equating c^2 times the rate of change of the mass to the PdV work for a sphere of radius a gives

$$d(4\pi\rho a^3 c^2/3) = -PdV = -4\pi P a^2 \dot{a} dt \quad (28.25)$$

and hence

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{P}{c^2}\right) \quad (28.26)$$

which must be used in place of (28.23). Note that this equation is a simple consequence of conservation of energy — the first law of thermodynamics — and the identification of energy and mass $E = Mc^2$.

A positive pressure therefore causes a reduction in the mass enclosed within a comoving sphere. Now Newtonian gravity theory can readily describe the motion of particles on radial orbits in the potential generated by a time varying mass. The result is that the acceleration is still given by

(28.22), but that the term E_0 in the energy equation is no longer constant in time. Differentiating (28.24) with respect to time gives

$$\dot{E}_0 = \dot{a}\ddot{a} - \frac{4}{3}\pi G\dot{\rho}a^2 - \frac{8}{3}\pi G\rho a\dot{a} \quad (28.27)$$

or, using (28.22) and (28.23),

$$\dot{E}_0 = 4\pi G\rho Pa\dot{a}/c^2. \quad (28.28)$$

In this respect, however, Newtonian theory is wrong. General relativity tells us that the term E_0 in the energy equation actually remains constant even with $P \neq 0$. This means that the Newtonian acceleration equation (28.22) cannot be correct. Differentiating the energy equation (with $E_0 = \text{constant}$) and using the equation of continuity gives the modified acceleration equation

$$\ddot{a} = -\frac{4}{3}\pi G \left(\rho + \frac{3P}{c^2} \right) a. \quad (28.29)$$

Thus we see that a positive pressure actually *increases* the deceleration of the Universal expansion.

Equation (28.29) is an important and surprising result; it says that *pressure gravitates in general relativity*. We should be clear what is meant by the density ρ here, and, in particular, draw a distinction between total mass-energy density and the proper mass density. Naively, one might imagine the pressure term in (28.29) as being some $\mathcal{O}(v^2/c^2)$ correction of the proper mass density for the kinetic energy of random motions. This term has the right magnitude at least. This is not correct; the density ρ here is the total mass-energy density including kinetic energy. To reinforce this, note that the proper mass density is not conserved in the transition from $P = 0$ to $P > 0$, since some of the rest-mass of the bombs is consumed in creating the kinetic energy of shrapnel and any energy in radiation. Since \dot{a} and E_0 should not change discontinuously in this transition it is the total mass energy density — which is conserved through this transition — which must appear here.

At first sight it may appear unphysical to identify the source of the gravitational field with $\rho + 3P/c^2$. This is because this quantity does *not* obey any fundamental conservation law; we can change the pressure, within limits, at will by exploding bombs etc. The bombs example gives an irreversible change in the pressure, but this is not an essential requirement. Consider instead a universe consisting of a dust of baseball players who, at some predetermined instant after the big bang, start practicing their pitching and lob balls to their neighbors. Since the players are receding from each other they receive balls with less energy than which they are thrown, and they must therefore be steadily consuming rest-mass in the process. If the players subsequently stop practicing, the pressure returns to zero. While contrived, this example shows that we can in principle switch the pressure on and off as we like.

To see why this is unsettling, consider what happens if we have a large region containing a dust of bombs primed to explode — would this then suddenly change the gravitational acceleration, and therefore the orbits, of satellites outside? It is pretty obvious that they cannot change, at least instantaneously, since this would violate causality. To extend the example, let us enclose the bombs in an initially slack balloon. The steady state after the bombs explode is to have the internal pressure balanced by tension in the enclosing membrane. Again, it does not seem reasonable that the gravitating mass as seen by an external observer (measuring the Keplerian orbits of satellites say) would change. How can the internal micro-physics in the balloon affect the total mass and energy seen by an external observer? The resolution, in this case, is that the extra gravitational attraction due to the positive pressure in the interior is canceled by a *repulsion* arising from the tension (a kind of negative pressure) in the balloon skin. Inside the balloon, however, the acceleration of test particles is correctly given by (28.29) rather than the Newtonian law (28.22).

Note that energy is *not* conserved locally in an expanding Universe with non-zero pressure, since each region does work on its surroundings. Recall that, in mechanics, energy is conserved only for a Lagrangian which has no explicit time dependence. The Lagrangian density for fields or particles in an expanding Universe, when written in comoving coordinates, contains the time-dependent scale factor $a(t)$ as a parameter, so energy is not conserved. For a finite sphere, bounded by a tense membrane to provide appropriate boundary conditions the total energy (including the potential

energy stored in the membrane) is conserved. Thus, if you want the energy budget to balance with $P \neq 0$, just imagine the Universe to be contained within some large balloon.

An important feature that emerges from the fact that (28.4) applies equally with or without pressure is that changing the equation of state of the universe can have no influence on the sign of the energy term; if the universe is initially unbound then it stays that way forever and *vice versa*. A related consequence is that, for open models, the late-time asymptotic velocity of a sphere $\dot{a}(t \rightarrow \infty)$ is also independent of the thermal history of the Universe.

28.8 Radiation Dominated Universe

Another important constituent of the Universe is the thermal *microwave background radiation* (MBR), which is a relic of the big bang. Currently, the density of the MBR is around 10^{-4} times the critical density, and therefore also much less than that of the matter, but photons redshift and lose energy as the universe expands (a consequence of the work done by their pressure) and the radiation density scales as $\rho_{\text{rad}} \propto a^{-4}$ while $\rho_{\text{matter}} \propto a^{-3}$ so when the Universe was about 10^{-4} times smaller than today (ie at a *redshift of equality* $1 + z_{\text{eq}} = 10^4$) the matter and radiation density would have been equal. Since the expansion law in the matter dominated era is $a \propto t^{2/3}$, and the present age of the Universe is $t_0 \sim 10^{10}\text{yr}$, this transition from radiation to matter domination at $t_{\text{eq}} \simeq (1 + z_{\text{eq}})^{-3/2} t_0 \simeq 10^4\text{yr}$. At earlier times, the Universe was radiation dominated.

In the radiation dominated era it is a very good approximation to neglect the constant $2E_0$ in the energy equation (28.4), and with $\rho \propto a^{-4}$ the solution of this equation is again a power law $a \propto t^\alpha$ but now with $\alpha = 1/2$.

- In both radiation and matter dominated eras, the age of the Universe is roughly the inverse of the expansion rate, though the constant of proportionality depends on the details of the equation of state.
- The Universe decelerates more with non-zero positive pressure.
- The horizon problem persists in the radiation dominated era since $r_h = c \int dt/a(t) \propto t^{1/2}$. As discussed, there is a ‘horizon problem’ in any universe which decelerates.

28.9 Number of Quanta per Horizon Volume

Assuming that the universe is radiation dominated at early times it is interesting to ask: what is the number of quanta per horizon volume (i.e. physical volume $V = c^3 t^3$).

For thermal radiation most of the energy is in photons with $E \sim kT$; if we imagine the Universe to be a periodic box of side L , the spacing of wave-numbers is $\delta k = 2\pi/L$ so, with $\hbar k \sim E/c$ the number of photons in volume L^3 is $N \sim (k/\delta k)^3 \sim (E/\hbar c)^3 L^3$ and therefore the number density $n \sim E^3/(\hbar^3 c^3)$. The density is $\rho \sim nE/c^2 \sim E^4/(\hbar^3 c^5)$ — the Stefan-Boltzmann law — and this is related to the age of the universe by $t \sim 1/\sqrt{G\rho}$ and therefore the relation between the age of the universe and the characteristic energy of the quanta is

$$t \sim \frac{h^{3/2} c^{5/2}}{G^{1/2}} \frac{1}{E^2}. \quad (28.30)$$

We are ignoring here the fact that at very high energies there are more particle species that are in equilibrium *via* number changing reactions. This is because number changing reactions for particles of mass m are ‘frozen out’ when the temperature falls below $T \sim mc^2/k$. More rigorously we should include a factor for the number of degrees of freedom, but this only introduces a correction factor of order unity.

The physical horizon size is $V_h = (ct)^3$, so with $n \sim E^3/(\hbar^3 c^3)$ the number of quanta per horizon volume is $N_h = nV_h \sim E^3 t^3/\hbar^3$ or, with (28.30)

$$N_h \sim \left(\frac{E^3}{\hbar^3}\right) \left(\frac{h^{9/2} c^{15/2}}{G^{3/2} E^6}\right) \sim \left(\frac{h^{1/2} c^{5/2}}{G^{1/2} E}\right)^3 = \left(\frac{E}{E_{\text{pl}}}\right)^{-3} \quad (28.31)$$

where we have defined the *Planck energy*

$$E_{\text{pl}} \equiv \sqrt{\frac{\hbar c^5}{G}} \simeq 10^{19} \text{GeV}. \quad (28.32)$$

It is also useful to define the *Planck mass*

$$m_{\text{pl}} \equiv \frac{E_{\text{pl}}}{c^2} = \sqrt{\frac{\hbar c}{G}} \simeq 10^{-5} \text{g}. \quad (28.33)$$

One can also define the *Planck time*

$$t_{\text{pl}} \equiv \hbar/E_{\text{pl}} \simeq 10^{-43} \text{s}. \quad (28.34)$$

Thus as we go back in time, the number of quanta per horizon volume decreases. Only for Universes much older than the Planck-time is it reasonable to assume a nearly uniform energy density. At *Planck-scale* energies, the fluctuations in the fields are expected to produce strong gravitational effects.

28.10 Curvature of Space-Time

So far we have concentrated on local properties of the universe — expansion rate, density, pressure etc — and have thus avoided any discussion of the global *curvature of space-time*, which, according to general relativity is caused by matter.

Space-time curvature is described by the metric tensor $g_{\mu\nu}$. This is the curved-space generalization of the Minkowski metric $\eta_{\mu\nu}$ and is defined such that the proper separation of two events with coordinate separation \vec{dx} is $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$. The metric, or *line element* for a spatially homogeneous and isotropic world model can be written as

$$ds^2 = -d\tau^2 + a^2(\tau)(d\omega^2 + S_k^2(\omega)d\sigma^2) \quad (28.35)$$

where $d\sigma^2 = d\theta^2 + \sin^2\theta d\phi^2$ is the usual solid angle.

- The radial coordinate ω is a comoving coordinate — comoving observers have constant ω , θ , ϕ .
- The time coordinate τ is the proper time since the big-bang for a comoving observer.
- The line element (28.35) follows from geometrical considerations alone (see Gunn's Saas Fee lectures), and is independent of any specific theory of gravity.
- The function $S_k(\omega)$ takes one of three forms, corresponding to the value of the *curvature eigenvalue* $k = 0, \pm 1$:

$$S_k(\omega) = \begin{cases} \sin \omega \\ \omega \\ \sinh \omega \end{cases} \quad \text{for } k = \begin{cases} 1 \\ 0 \\ -1 \end{cases} \quad (28.36)$$

- For $k = 0$, the constant τ surfaces are *spatially flat*.
- For $k = +1$, the constant τ surfaces are unbounded but spatially finite (or *closed*) 3-dimensional analogs of the 2-sphere. The scale factor a is then the curvature radius.
- For $k = -1$, the constant τ surfaces are unbounded and spatially infinite. They are analogous to a saddle in two dimensions.
- A common alternative representation of the line element is to use radial coordinate $r = S_k(\omega)$, in terms of which the line element is

$$ds^2 = -d\tau^2 + a^2(\tau) \left(\frac{dr^2}{1 - kr^2} + r^2 d\sigma^2 \right). \quad (28.37)$$

- It is common to define the *conformal time* η such that $d\eta = d\tau/a(\tau)$ and the line element is then

$$ds^2 = a^2(\tau)(-d\eta^2 + d\omega^2 + S_k^2(\omega)d\sigma^2). \quad (28.38)$$

Radial photons move along 45 degree diagonals in $\eta - \omega$ space.

So far this is pure geometry plus the assumptions of homogeneity and isotropy. The field equations of general relativity provide a relation between the curvature of space-time and the matter content. Here, this boils down to the equation

$$\left(\frac{da}{d\tau}\right)^2 = \frac{8\pi}{3}G\rho a^2 - c^2k \quad (28.39)$$

with $k = \pm 1$ the curvature eigenvalue. This is equivalent to the energy equation (28.4), with the energy constant $2E_0 \rightarrow -c^2k$, and with an appropriate scaling of a . In the Newtonian analysis, the scale factor is arbitrary. In general relativity, we take the scale factor to be the curvature scale.

This tells us that spatially closed and open models correspond to bound and unbound cosmologies respectively. The borderline case is spatially flat, and can be obtained as the limiting case of either closed or open models.

Note that the curvature scale is a comoving scale, and is therefore fixed. Changing the equation of state can have no influence on the geometry of the Universe. However, in inflation, the curvature scale is *stretched* and becomes exponentially large, and in this way the Universe can be prepared in a state which is effectively spatially flat.

Finally, one should not confuse space-time curvature and the curvature of the $\tau = \text{constant}$ surfaces. In the $k = 0$ solution, the constant τ surfaces are spatially flat, but the mass density does not vanish so the *space-time* is still curved. In open models, the density, and therefore space-time curvature tend to zero at late times, but the constant τ surfaces are still hyperbolic, saddle-like surfaces; the metric in that limit (with $a(\tau) \propto \tau$) is simply a re-parameterization of the flat space-time Minkowski metric.

The spatial metric of the unit 2-sphere is $dl^2 = d\theta^2 + \sin^2\theta d\phi^2$. The length of a line element perpendicular to the radial direction (i.e. $d\theta = 0$) is $dl = \sin\theta d\phi$ and the circumference of a circle is $l = 2\pi \sin\theta$. This increases linearly with θ for $\theta \ll 1$, reaches a maximum of 2π at $\theta = \pi/2$, and then shrinks to zero at the antipodal point $\theta = \pi$. Similarly, the spatial metric of the closed model is a 3-sphere, with $dl^2 = d\omega^2 + \sin^2\omega d\sigma^2$. The area of a spherical surface which is perpendicular to the radius vector (i.e. $d\omega = 0$) is $dA = a^2 \sin^2\omega d\sigma^2$ with $d\sigma^2 = d\theta^2 + \sin^2\theta d\phi^2$. The total area of a sphere of radius ω is $A = 4\pi a(t)^2 \sin^2\omega$. Just as in the 2-dimensional case, this peaks at $\omega = \pi/2$ and shrinks back to zero at the antipodal point $\omega = \pi$.

The closed model is finite, yet has no boundary. However, at least if we restrict attention to zero-pressure equation of state, we are free to take only a finite part of the total solution $\omega < \omega_{\max}$. This is a spherically symmetric mass configuration, and so should match onto the Schwarzschild solution for a point mass m , for which the space-time metric is (in units such that $c = G = 1$)

$$-(d\tau)^2 = -\left(1 - \frac{2m}{r}\right)(dt)^2 + \left(1 - \frac{2m}{r}\right)^{-1}(dr)^2 + r^2((d\theta)^2 + \sin^2\theta(d\phi)^2). \quad (28.40)$$

Comparing the angular part of the metric it is apparent that the Schwarzschild radial coordinate r and the FRW ‘development angle’ ω are related by $r = a \sin\omega$. Now a particle at the edge of the FRW model can equally be considered to be a radially moving test particle in the Schwarzschild geometry. We found in problem (???) that such particles orbits satisfy

$$\left(\frac{dr}{d\tau}\right)^2 = \frac{2Gm}{r} + \text{constant}. \quad (28.41)$$

Compare this with the energy equation

$$\left(\frac{da}{d\tau}\right)^2 = \frac{2GM}{a} + \text{constant}. \quad (28.42)$$

where we have defined the mass parameter $M = 4\pi\rho a^3/3$. With $r = a \sin \omega$ this implies that the Schwarzschild mass parameter is

$$m = M \sin^3 \omega \quad (28.43)$$

This is interesting. For $\omega \ll 1$, the mass increases as ω^3 as expected. However, the mass is maximized for a model with a development angle of $\pi/2$, or half of the complete closed model. If we take a larger development angle, and therefore include more proper-mass, the Schwarzschild mass parameter decreases. To the outside world, this positive addition of proper mass has negative total energy. This means that the negative gravitational potential energy outweighs the rest-mass energy. The gravitating mass shrinks to zero as $\omega \rightarrow \pi$. Evidently a nearly complete closed model with $\omega = \pi - \epsilon$ looks, to the outside world, like a very low mass, that of a much smaller closed model section with $\omega = \epsilon$.

The total energy of a complete closed universe is therefore zero. Zel'dovich, and many others subsequently, have argued that this is therefore a natural choice of world model if, for instance, one imagines that the Universe is created by some kind of quantum mechanical tunneling event. To be consistent with the apparent flatness of the Universe today one would need to assume that the curvature scale has been inflated to be much larger than the present apparent horizon size.

It is interesting to compare the external gravitational mass with the total proper mass. The volume element of the parallelepiped with legs $d\omega$, $d\theta$, $d\phi$ is

$$d^3x = (a d\omega) \times (a \sin \omega d\theta) \times (a \sin \omega \sin \theta d\phi), \quad (28.44)$$

so the total mass interior to ω is

$$M_{\text{proper}} = \rho a^3 \int_0^\omega d\omega \sin^2 \omega \int d\theta \sin \theta \int d\phi = \frac{3}{2} M \left[\omega - \frac{\sin 2\omega}{2} \right] \quad (28.45)$$

The gravitational mass (28.43) and proper mass (28.45) are shown in figure 28.6.

These partial closed FRW models start from a singularity of infinite density and then expand, passing through the Schwarzschild radius $r = 2Gm/c^2$. With $r = a \sin \omega$, $m = M \sin^3 \omega$, and $a = M(1 - \cos \eta)$, particles on the exterior cross the Schwarzschild radius at conformal time η when $1 - \cos \eta = 2 \sin^2 \omega$. For $\omega \ll 1$, this occurs when $\eta = 2\omega$. Such solutions spend the great majority of their time outside the Schwarzschild radius. For the case $\omega = \pi/2$ — i.e. half of the complete solution — the exterior particles just reach the Schwarzschild radius. It may seem strange that the matter in these models can expand from within the Schwarzschild radius, but this is indeed the case. If one considers only the collapsing phase of these models then one has the classic model for black-hole formation as developed by Oppenheimer and Snyder. The spherical mass collapses to a point, and photons leaving the surface can only escape to infinity if they embark on their journey while the radius exceeds the Schwarzschild radius. The expanding phase of these models is just the time reverse of such models; what we have is a ‘white-hole’ solution. The initial singularity is visible to the outside world (eventually) just as photons from the outside can fall in to the final singularity.

One can visualize the geometry solution in a 2-dimensional analog (figure 28.7). The interior is part of a 2-sphere, which matches smoothly onto a exterior solution much like a depressed rubber sheet. The closed solutions with $\omega > \pi/2$ are slightly peculiar in the sense that, at the edge, the radius r is decreasing with increasing ω . This means that the FRW geometry matches onto a ‘throat’ like exterior geometry.

In the limit of small development angle $\omega_{\text{max}} \ll 1$ both the proper mass (28.45) and the gravitational mass (28.43) are given, to lowest order in ω_{max} , by

$$M_{\text{proper}} = M_{\text{grav}} = M \omega_{\text{max}}^3 = \frac{4\pi}{3} \rho a^3 \omega_{\text{max}}^3 \quad (28.46)$$

by the gravitational mass. However, keeping terms of next higher non vanishing order one obtains for the ratio of proper to gravitational mass

$$\frac{M_{\text{proper}}}{M_{\text{grav}}} = 1 + \frac{2}{5} \omega_{\text{max}}^2 + \mathcal{O}(\omega_{\text{max}}^4). \quad (28.47)$$

This result will prove useful below.

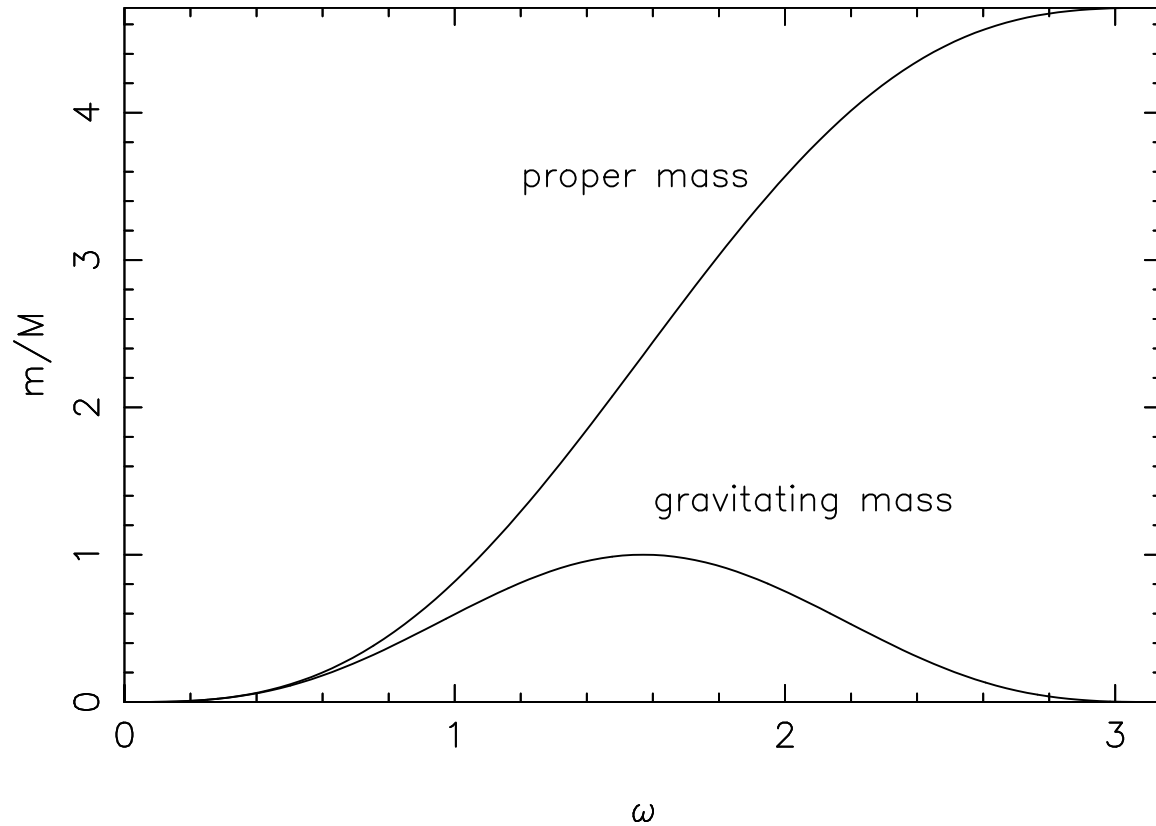


Figure 28.6: The proper-mass and gravitational mass for a partial closed FRW cosmology are plotted against the development angle ω .

28.11 Problems

28.11.1 Energy of a Uniform Expanding Sphere

Compute the total kinetic energy and the total gravitational binding energy for a uniformly expanding constant density sphere of radius a , density ρ and surface velocity \dot{a} .

28.11.2 Solution of the FRW Energy Equation

Verify that a parametric solution of the energy equation

$$\dot{a}^2 = \frac{8}{3}\pi G\rho a^2 + 2E_0 \quad (28.48)$$

is

$$\begin{aligned} a(\eta) &= A(1 - \cos \eta) \\ t(\eta) &= B(\eta - \sin \eta) \end{aligned} \quad (28.49)$$

and find the relation between constants A , B and the energy constant E_0 .

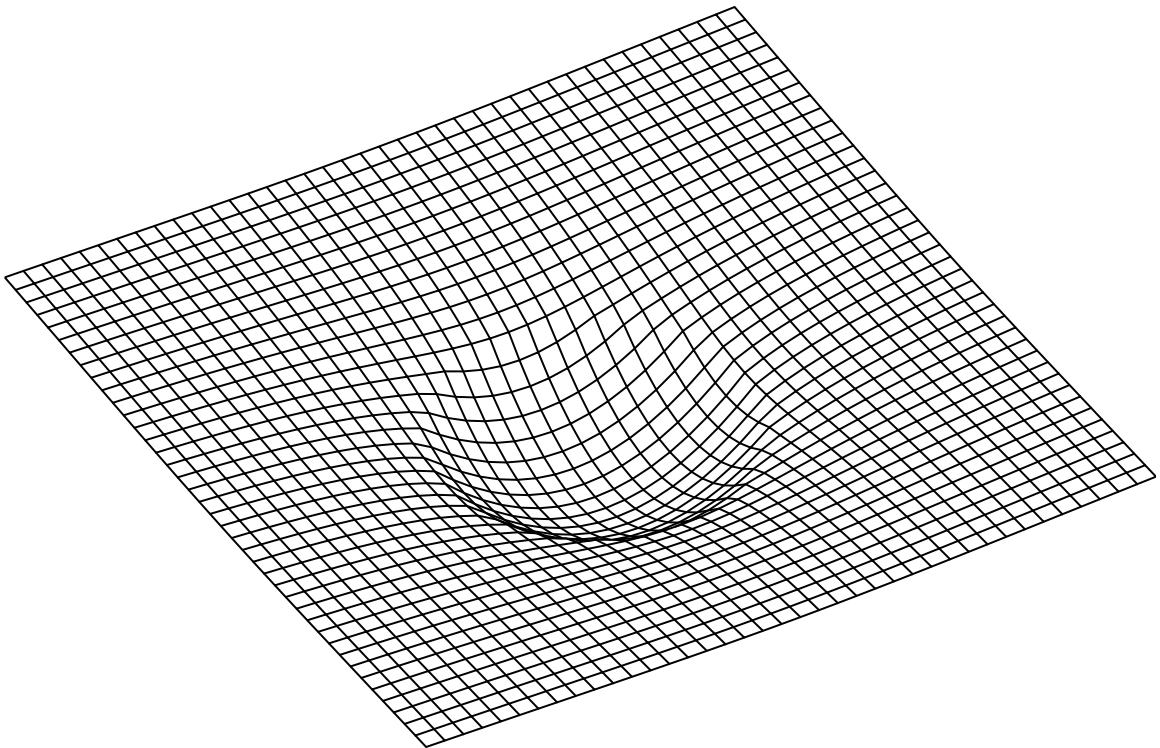


Figure 28.7: A partial closed FRW model consists of part of a 3-sphere which matches smoothly onto the horn-like Schwarzschild geometry. The 2-dimensional analog of this composite geometry can be visualized as part of a parabolic sphere connecting smoothly onto a 2-dimensional horn.

Chapter 29

Inflation

29.1 Problems with the FRW Models

We have already discussed two problematic features of the matter or radiation dominated FRW models. One of these is the *horizon problem*, which says that the remarkable uniformity we see on large scales today must be imposed a-causally. The other is the *flatness problem*, which is that the observational fact that Ω is currently not very different from unity requires it to have been astonishingly close to unity in the distant past. Both of these problems render the models unattractive since the basic properties of flatness and homogeneity are not really explained by the theory, rather they must be imposed as finely tuned initial conditions.

To these problems we can add the *monopole problem*. In *grand unified theories* (GUTs) massive magnetic monopoles are predicted to exist. In the hot big bang model, at the time of GUT symmetry breaking (as the Universe cools through the GUT temperature of around 10^{16}GeV) these monopoles appear as topological defects, with a number density on the order of one per horizon size. These objects are a definite prediction of GUTs, yet their existence in anything like this abundance would be a disaster for cosmology, as they would have a density today hugely in excess of that observed.

There are also a number of additional unsettling features of the hot big bang model. One might ask what happened before the initial singularity? What are the seeds of the structure that we see in the Universe? What explains the *baryon asymmetry* of the Universe? There are now $\sim 10^8$ photons per baryon, which seems to imply that there was initially a slight asymmetry between baryons and anti-baryons at the one part in 10^8 level.

29.2 The Inflationary Scenario

In the *inflationary scenario* many of these problems appear to be solved or at least ameliorated. The essence of inflation is to assume that at early times the Universe passed through a phase with a strongly negative pressure (i.e. positive tension).

Let us start with the horizon problem. As we have already discussed, this can be traced directly to the deceleration of the expansion Universe; if $\ddot{a} < 0$ then the velocity difference between any two observers, which is proportional to \dot{a} , decreases with time. Therefore, going back in time, the relative velocity inexorably increases and at some finite time in the past reaches the speed of light c , and before that the two observers cannot exchange information or causal influences.

The only way to avoid this is for the Universe to have undergone an *accelerating phase* with $\ddot{a} > 0$ in its early history. From the acceleration equation (28.29) this requires $\rho + 3P/c^2 < 0$, or a strong negative pressure $P < -\rho c^2/3$. This is the strange, and somewhat counterintuitive, feature of the general relativistic expansion law; just as a positive pressure augments the gravitational deceleration, a sufficiently strong negative pressure can cause the expansion to accelerate.

At first sight it is hard to see how a negative pressure can arise. Certainly, for a gas of particles interacting through localized collisions, the pressure cannot be negative. As shown in Weinberg's

book, for instance, a relativistic ideal gas must have pressure in the range $0 \leq P \leq \rho c^2/3$. Also, we can consider pressure to be the flux of momentum. A particle moving in the positive x direction carries a positive x -component of momentum, and therefore the flux of x -momentum passing in the positive x -direction through a surface must be positive.

However, if we consider *fields*, rather than particles, then the possibility of negative pressure is not at all unreasonable. After all, the most commonplace field that we can feel macroscopically is the magnetic field. Anyone who has played with a pair of bar-magnets or pulled magnets off a fridge knows that such fields have strong tension. However, such fields do not have *isotropic* tension; there is tension along the field lines — you have to do work to stretch the field out and create more of it — but in the transverse directions the opposite is true; as we see from images of the field produced with iron filings the field between a pair of magnets clearly wants to burst out sideways. This follows directly from energetic considerations, along with flux conservation. Imagine you try to confine the field to pass through a smaller area. Flux conservation means that the field strength must increase inversely with the area, but the energy density scales as the square of the field strength, so the total energy is larger the smaller the cross-sectional area. A static magnetic field then has negative pressure along the field lines but positive pressure in the transverse directions.

This anisotropy of the pressure for a macroscopic quasi-static magnetic field is associated with the fact that electromagnetism is a *vector field*. Now in grand unified theories, there are also *scalar fields* such as the Higgs field. The leap of inspiration which led to the concept of inflation was the realization that a field like this can have a pressure which is isotropic — this is natural enough since a scalar field cannot point in any particular direction — and may have the negative pressure required to make the universal expansion accelerate. As discussed in §18.4, a possible solution of the field equations for a *massive scalar field* at early times is a spatially and temporally constant field value. The energy density and pressure (i.e. diagonal components of the stress-energy tensor) for a scalar field are

$$\rho c^2 = \frac{1}{2} \dot{\phi}^2 + \frac{c^2}{2} (\nabla \phi)^2 + V(\phi) \quad (29.1)$$

$$P = \frac{1}{2} \dot{\phi}^2 - \frac{c^2}{6} (\nabla \phi)^2 - V(\phi). \quad (29.2)$$

We are following the usual convention in cosmology that any mass term $m^2 c^4 \phi^2 / \hbar^2$ is considered part of the potential function $V(\phi)$. A static and homogeneous field configuration — one with $\dot{\phi} = 0$ and $\nabla \phi = 0$ — therefore has density $\rho = V(\phi)/c^2$ and pressure $P = -V(\phi) = -\rho c^2$. Such a potential dominated field configuration therefore amply satisfies the inequality $P < -\rho c^2/3$.

With $P = -\rho c^2$, the continuity equation tells us that $\dot{\rho} = 0$. The cosmological expansion does work against the tension of the field at just the rate required to keep the density constant; for this reason the inflationary universe has been dubbed the *ultimate free lunch*. The acceleration equation is then

$$\ddot{a} = -\frac{4}{3} \pi G (\rho + 3P/c^2) = \frac{8}{3} \pi G \rho, \quad (29.3)$$

with $\rho = \text{constant}$. The general solution of this equation is

$$a(t) = a_+ e^{+Ht} + a_- e^{-Ht} \quad (29.4)$$

with

$$H = \sqrt{8\pi G \rho / 3} = \text{constant}. \quad (29.5)$$

For generic initial conditions, a potential dominated universe, will tend towards an exponentially expanding solution $a \propto e^{Ht}$.

During inflation, the comoving horizon size — defined here as the comoving distance that a photon can travel per expansion time — is $r_h \sim cH^{-1}/a$. Since H is constant this decreases exponentially as $r_h \propto e^{-Ht}$. At early times during inflation photons can travel great comoving distances but this decreases as time goes on. In a viable inflationary model, inflation cannot continue forever, but must end, with conversion of the energy density — all stored in the scalar field — into ‘ordinary’ matter with $P = \rho c^2/3$, i.e. we must make a transition from a scalar field dominated universe to

a radiation dominated hot-big bang model. Side-stepping, for the moment, the issue of exactly how this so-called ‘re-heating’ occurs, the overall behavior of the comoving horizon scale (as we have defined it above) is shown as the solid line in figure 29.1. This allows the possibility that the entire Universe was initially in causal contact. Let’s look at this from the point of view of a pair of comoving observers. These have a constant comoving separation, as indicated by the horizontal dashed line say. During inflation, the velocity difference between these observers increases as they accelerate apart, and a pair of observers with initial relative velocity $v < c$ will at some time lose causal contact with each other once their relative velocity reaches the speed of light. If the universe later becomes radiation dominated, the relative velocity will subsequently fall and these observers can regain causal contact. For those who feel uneasy with the somewhat hand-waving definition of the horizon size as the distance light can travel in an expansion time, consider instead the rigorous definition of the comoving distance to a distant source as a function of the ‘look-back time’ $\tau = t_0 - t$

$$r(\tau) = c \int_0^\tau \frac{d\tau}{a(t_0 - \tau)}. \quad (29.6)$$

In the matter dominated era this increases with decreasing τ at first, but tends towards a limiting asymptote. Back in the inflationary era, however, this integral grows exponentially and becomes arbitrarily large.

What about the flatness problem? Recall that departure from flatness is an indication of an imbalance between the kinetic and potential energy terms in the energy equation

$$\dot{a}^2 = \frac{8}{3} \pi G \rho a^2 - k c^2. \quad (29.7)$$

For an exactly flat universe $k = 0$ and these two terms are exactly equal. Now consider a universe which is initially open say, with $k = -1$. During an inflationary phase, the expansion accelerates, \dot{a} increases, as must the potential energy term on the right hand side. Inflation acts to increase, exponentially, the kinetic and potential terms in the energy equation. Thus even if there is a non-zero initial energy constant, it will tend to become exponentially small at the end of inflation. Inflation therefore drives the universe towards flatness; the $\Omega = 1$ state becomes an attractor rather than an unstable state. Another way to look on this is to realize that in FRW models — and the inflationary universe is an FRW model, just one with a weird equation of state — the curvature scale is a comoving scale. Relative to the horizon scale, the curvature scale is stretched exponentially. Thus it might be that our universe is open or closed, but that the curvature scale has been stretched to be enormously larger than the currently observable region of the universe.

What about the monopole problem? These are topological defects of a field. Inflation allows this field to be coherent over very large scales; up to the initial comoving horizon scale. Provided the universe re-heats to a temperature less than the GUT scale, monopoles — which have a mass around the GUT energy scale — will not be effectively created.

It is interesting to ask, how many e -foldings of inflationary expansion are required in order to establish causality over the region of the universe (size $l \sim c/H_0$) that we can currently observe? The answer depends on the temperature at which reheating occurs. If this reheating temperature is around the energy scale of *grand unification*, or $T \sim T_{\text{GUT}} \sim 10^{16} \text{GeV}$, then the temperature falls by about a factor 10^{25} before the Universe becomes matter dominated at a temperature of about 10eV . During that period the comoving horizon grows as $r_h \propto t^{1/2} \propto a \propto 1/T$, or by about 15 orders of magnitude. Once the universe becomes matter dominated the horizon grows as $r_h \propto t^{1/3} \propto a^{1/2}$ or by about another factor of 100. The current horizon is therefore about $10^{27} \simeq e^{62}$ times larger now than at the reheating time, so we need about 60 e -foldings of inflation.

29.3 Chaotic Inflation

Originally, it was imagined that the field driving inflation, the *inflaton field*, had a w-shaped potential of the kind involved in spontaneous symmetry breaking with the Higgs field. For reasons we shall

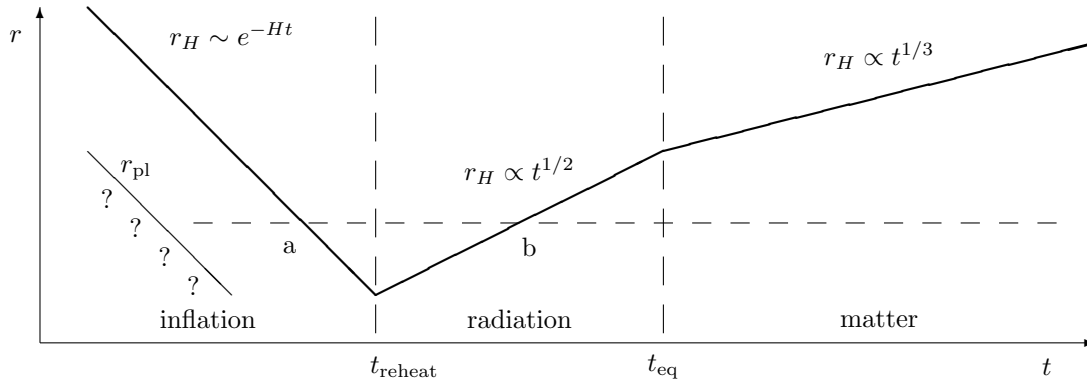


Figure 29.1: The evolution of the comoving horizon scale (heavy solid line) in a Universe which passes through three phases; an inflationary stage followed by a radiation dominated and then a matter dominated era. The ordinate is logarithmic and the abscissa is linear in the inflationary era and logarithmic thereafter. The diagonal line labeled r_{pl} indicates the Planck length. The horizontal dashed line indicates the comoving separation of a pair of observers. This length scale appears at the Planck scale at some time during the inflationary era. We have no adequate description of physics below and to the left of the Planck scale line. The observers then accelerate away from one another. If they were to exchange signals they would perceive an increasing redshift. The separation between the observers reaches the horizon scale at the point labeled ‘a’. At that time their relative velocity reaches the speed of light and their relative redshift becomes infinite. Subsequently the observers are unable to exchange signals. At the reheating epoch the Universe starts to decelerate, and at point ‘b’ the recession velocity falls below the speed of light. The observers then re-appear on each other’s horizon; they can exchange signals which are received with steadily decreasing redshift. The separation chosen here is such that it re-enters the horizon during the radiation dominated era. Larger separations enter the horizon at later times. The current horizon scale is $ct_0 \sim c/H_0 \simeq 3000\text{Mpc}$. Since the comoving horizon scale is proportional to $t^{1/3}$ in the matter era, the horizon scale at t_{eq} is smaller than the current horizon by a factor ~ 100 , or about $\sim 30\text{Mpc}$, the scale of super-clusters. The horizontal dashed line might represent the size of a region encompassing the matter now comprising a galaxy say.

not go into here, such models have fallen out of favor. Instead, most attention is currently focused on so-called *chaotic inflation* models in which the field has a potential function as sketched in figure 29.2. It is assumed that the field starts out at some point far from the origin, and then evolves to smaller values much as a ball rolling down a hill. In this section we shall explore what is required in order to obtain a viable inflationary scenario, i.e. one in which there are sufficiently many e -foldings.

For concreteness, consider a field with Lagrangian density

$$\mathcal{L} = \frac{1}{2c^2} \dot{\phi}^2 - \frac{1}{2} (\nabla \phi)^2 - \frac{\lambda}{\hbar c} \phi^4. \quad (29.8)$$

This is a massless field with a self-interaction term parameterized by the dimensionless constant λ . Note that the Lagrangian density has units of energy density $[\mathcal{L}] = \text{ML}^{-1}\text{T}^{-2}$ so the field has dimensions $[\phi] = \text{M}^{1/2}\text{L}^{1/2}\text{T}^{-1}$ (in natural units $\text{L} = 1/\text{M}$, $\text{T} = 1/\text{M}$ the field has units of mass). Assuming the field to be spatially uniform, the equation of motion is

$$\ddot{\phi} + 3H\dot{\phi} + \frac{4\lambda c}{\hbar} \phi^3 = 0, \quad (29.9)$$

where the last term is the potential gradient and the second term is the damping due to the cosmo-

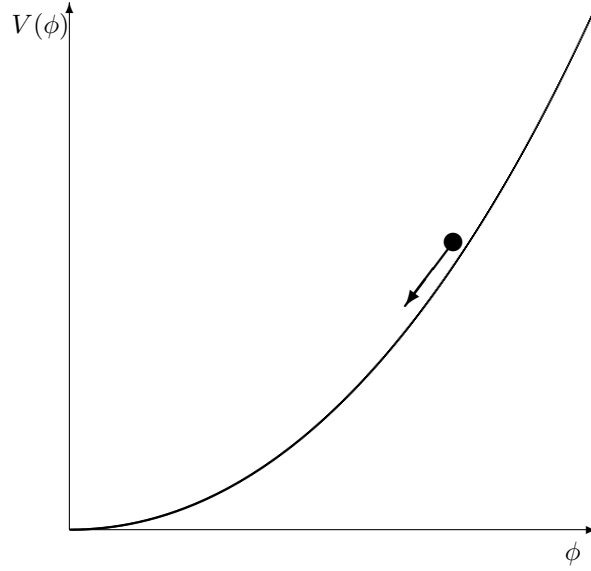


Figure 29.2: In the *chaotic inflation* scenario, the potential function $V(\phi)$ is assumed to be a monotonically increasing function with $V(0) = 0$. The potential could simply be a mass term $V \propto \phi^2$ or perhaps $V \propto \phi^4$.

logical expansion. The energy density and pressure are

$$\rho c^2 = \frac{1}{2c^2} \dot{\phi}^2 + \frac{\lambda}{\hbar c} \phi^4 \quad (29.10)$$

$$P = \frac{1}{2c^2} \dot{\phi}^2 - \frac{\lambda}{\hbar c} \phi^4 \quad (29.11)$$

and the expansion rate is given by

$$H^2 = \frac{8}{3} \pi G \rho = \frac{8}{3} \pi G \left(\frac{1}{2c^4} \dot{\phi}^2 + \frac{\lambda}{\hbar c^3} \phi^4 \right). \quad (29.12)$$

Equation (29.9) is like that of a ball rolling down a hill with a frictional force, the coefficient of friction H being dependent on the field and the field velocity through (29.12). For such a system there are two limiting types of behavior, depending on the value of the field. In one, the friction is negligible and the field is in free-fall with $\ddot{\phi}$ equal to the potential gradient. In the other, the friction is important, the first term in (29.9) is negligible compared to the other terms and the field moves at a ‘terminal velocity’ such that the friction force just balances the potential gradient.

Let’s assume, for the moment, that the former is the case. The effective equation of motion is then

$$\ddot{\phi} + \frac{4\lambda c}{\hbar} \phi^3 = 0. \quad (29.13)$$

The time-scale for changes in the field velocity is

$$t_{\text{accel}} \sim \sqrt{\frac{\phi}{\ddot{\phi}}} \sim \sqrt{\frac{\hbar}{\lambda c}} \phi^{-1}. \quad (29.14)$$

After one acceleration time-scale the field will acquire a velocity

$$\dot{\phi} \sim \ddot{\phi} t_{\text{accel}} \sim \sqrt{\frac{\lambda c}{\hbar}} \phi^2. \quad (29.15)$$

The kinetic energy term in the density is

$$\frac{1}{c^4}\dot{\phi}^2 \sim \frac{\lambda c}{\hbar}\phi^4 \sim \frac{V}{c^2} \quad (29.16)$$

so the potential and kinetic energy terms in the density are comparable. Now the condition that the friction should still be negligible is $3H\dot{\phi} \ll (\lambda c/\hbar)\phi^3$. Using (29.12), this inequality becomes

$$3H\dot{\phi} \simeq 3\sqrt{\frac{8\pi G\lambda\phi^4}{3\hbar c^3}}\sqrt{\frac{\lambda c}{\hbar}}\phi^2 \ll \frac{\lambda c}{\hbar}\phi^3. \quad (29.17)$$

The dimensionless interaction strength factors out of this inequality, so the condition that the friction be negligible is simply that the field be sufficiently weak ($\phi \ll \sqrt{c^4/G}$).

Conversely, the condition that the friction should dominate — the *slow-roll condition* — is

$$\phi \gg \sqrt{\frac{c^4}{G}}. \quad (29.18)$$

We can write this inequality in a rather more revealing manner if we note that the quantity $\sqrt{\hbar/c^3}\phi$ has dimensions of mass, so we need

$$\sqrt{\frac{\hbar}{c^3}}\phi \gg \sqrt{\frac{\hbar c}{G}} = m_{\text{Pl}} \quad (29.19)$$

where m_{Pl} is the Planck-mass. In natural units ($\hbar = c = 1$) this says that if the field is much greater than the Planck-mass then the friction will dominate and the field will be unable to roll freely down the potential, rather it will roll slowly down the hill at the terminal velocity

$$\dot{\phi} = -\frac{4\lambda c}{3\hbar H}\phi^3. \quad (29.20)$$

Assuming that the inequality (29.18) holds, what is the equation of state, or equivalently how large is the positive kinetic energy term $\sim \dot{\phi}^2/c^2$ in the pressure as compared to the potential term $V = \lambda\phi^4/\hbar c$? Squaring the terminal velocity (29.20) and using the inequality $H^2 \geq (8/3)\pi GV/c^2$ yields

$$\frac{1}{2c^2}\dot{\phi}^2 = \frac{16\lambda^2\phi^6}{9\hbar^2 H^2} \leq \frac{16\lambda^2\phi^6}{9\hbar^2} \left(\frac{8\pi G\lambda\phi^4}{3\hbar c}\right)^{-1} \simeq \left(\frac{\lambda\phi^4}{\hbar c}\right) \left(\frac{c^4}{G\phi^2}\right). \quad (29.21)$$

The first factor here is just the potential and, under the slow-roll condition (29.18), the second factor is much less than unity, so this says that

$$\frac{1}{2c^2}\dot{\phi}^2 \ll V(\phi). \quad (29.22)$$

The kinetic energy terms in the pressure and density are therefore much less than the potential terms and we therefore have $P \simeq -\rho c^2$ as required for inflation to proceed.

As already mentioned, in a viable model, inflation must be sustained for many e -foldings in order to solve the flatness, horizon problems. In one e -folding, the field will move a distance $\Delta\phi \sim \dot{\phi}/H$. For GUT scale inflation, where we need ~ 60 e -foldings, we need

$$\frac{\Delta\phi}{\phi} = \frac{\dot{\phi}}{H\phi} \lesssim \frac{1}{60} \equiv \epsilon. \quad (29.23)$$

Using (29.20) and $H^2 \sim GV/c^2 \sim G\lambda\phi^4/\hbar c^3$ this becomes

$$\phi \geq \sqrt{\frac{c^4}{\epsilon G}}. \quad (29.24)$$

Thus, the field needs to exceed the Planck mass by at least a factor $\epsilon^{-1/2} \sim 8$ in order to achieve sufficiently many e -foldings of inflation.

In this model, the field rolls slowly — very slowly at first — down the potential and the universe inflates. The expansion rate H does not remain precisely constant, but decreases slowly with time. Eventually the field reaches the value $(\hbar/c^3)^{1/2}\phi \simeq m_{\text{pl}}$, at which point the friction term $H\dot{\phi}$ in the equation of motion is no longer effective and the field starts to oscillate freely about the potential minimum. As discussed previously, the field amplitude will damp adiabatically, and the density will fall as $\rho \propto 1/a^3$; just as in a matter dominated universe. In the particle description, a macroscopic coherent field with $\nabla\phi = 0$ is a huge number of zero momentum particles — a ‘Bose condensate’ if you will — and therefore behaves just like non-relativistic particles. In more detail, one finds that the pressure for such a system oscillates about zero, being positive when $\phi = 0$ and $\dot{\phi}$ is maximized at the bottom of the potential and being negative at the limits of the field excursion when $\dot{\phi} = 0$. This is not what we want, which is a transition to a radiation dominated cosmology. If, however, there are other fields χ in the Lagrangian, and if these are coupled to the inflaton field ϕ , then once the ϕ field starts oscillating, it can decay into χ -ons, provided these have a mass which is lower than that of the ϕ -field. The details of this re-heating process depends on the details of the interactions between the field. In figure (29.1) we have assumed that re-heating happens promptly once the inflaton starts to oscillate. It is presumably possible, if the interactions are sufficiently weak, the universe might pass through a period of matter domination after the end of inflation before re-heating occurs.

We have considered a rather specific model above, with a $\lambda\phi^4$ potential. The main results are not specific to this choice. Had we instead assumed $V = (m^2 c^2/\hbar^2)\phi^2$ — i.e. a non-interacting, but massive, field, then we again find that the ‘slow-roll’ condition is simply that $\phi \gg \sqrt{c^4/G}$.

29.4 Discussion

The ‘inflationary scenario’ described above provides a plausible mechanism for preparing the universe in a flat and spatially homogeneous state on all scales that we can observe, starting from fairly generic initial conditions (rather than absurdly finely tuned ones). All we require is that the field start off at a sufficiently high value; the initial value of the field velocity $\dot{\phi}$ is largely irrelevant, since the cosmic drag term rapidly reduces $\dot{\phi}$ to the terminal velocity. The empirical evidence strongly encourages one to suspect that our universe has passed through such a phase. The situation facing cosmologists is rather like that facing a policeman who, walking down the high street, finds a jewelry shop with a broken window. Now it could be that this is a random accident, but the circumstantial evidence that it is a jewelry shop rather than a grocery or shoe shop, strongly encourages one to believe that this not an accident but a crime.

As we shall discuss later, the inflationary scenario also creates density fluctuations which can seed the structures we see in the distribution of galaxies and in the cosmic microwave background. The amplitude of these fluctuations is strongly model dependent, but the prediction is for fluctuations with dependence on wavelength very much like that which seem to be required.

These results make the inflationary model highly attractive. On the down side, one has to invoke a new field, the inflaton, precisely to obtain these desirable results. Initially, the development of this field of research was strongly linked to developments in fundamental particle physics — spontaneous symmetry breaking etc. — but the subject has now taken on a life of its own. While we have used GUT-scale inflation in order to derive e.g. the number of e -foldings, there is really no need to assume this (though reheating to super-GUT temperatures would be problematic). Indeed, studies of the expansion rate using supernovae have suggested that the universal expansion is now accelerating; it would seem that we are entering another inflationary phase. The ideas described above can readily be re-cycled to describe late-time inflation by choosing appropriate parameters (specifically, this requires that the fields be very light). The inflaton field must be coupled to other fields in order to allow re-heating, and in principle this allows empirical tests of the theory. However, the requirements on the form and strength of the interaction are not very specific, and the energies required to make GUT-scale inflatons is beyond the reach of terrestrial particle accelerators. Aside from the ‘predictions’ of flatness, homogeneity and density fluctuations — all of which were observed before inflation was invented — it is hard to find testable predictions. One hope is that the inflaton field and its potential will emerge as the low-energy some more fundamental theory which unifies

all of the forces, including gravity. This is an area of much activity at present, but hopefully will explain why there is an inflaton; why it has the potential it needs; why the minimum of the potential is at zero energy density and so on.

There is another rather unsettling aspect of the inflationary scenario, which is that we had to assume that the field is highly homogeneous. Many discussions of the subject argue that any inhomogeneity will be stretched to super-horizon scale in order to justify this assumption. This seems overly complacent. Recall that in figure (29.1) the boundary of the domain which we can describe without a theory of quantum gravity is not a fixed time $t = t_{\text{pl}}$, rather the time at which a region starts to be describable classically depends on the size of the region. Each time the universe doubles in size, each Planck-scale region gets replaced by eight new Planck-scale volumes. Predicting the ‘initial’ state of such regions requires a quantum theory of gravity, but it is commonly imagined that the classical universe emerges from some chaotic space-time foam. Now even if this process were to generate quite small occupation numbers for these Planck-scale modes, this would give a positive contribution to the pressure which would stop inflation taking place. If we want to invoke inflation then we must assume that this quantum-gravitational process produce a very rare vacuum.

There is one other peculiar feature of a potential dominated medium that needs mention. We have developed the theory here simply as we did for the FRW models, with the sole modification being the adoption of the equation of state $P = -\rho c^2$. There is, however, an important distinction to be drawn between such a medium and a fluid with $P = \rho c^3/3$ or $P = 0$ say. In the latter cases there are a preferred set of observers — the ‘comoving observers’ — for whom the stress energy tensor takes the symmetric form $T^{\mu\nu} = \text{diag}(\rho c^2, P, P, P)$. At each point in space, this zero momentum density condition picks out a unique velocity, and this gives us a unique ‘congruence’ of comoving observers. We can clearly determine unambiguously whether the universe is expanding or contracting by having such observers exchange light signals and measure red-shifts, for instance. In the inflationary phase, in contrast, when $P = -\rho c^2$ to high accuracy, there is no such unique congruence of comoving observers, since with $P = -\rho c^2$ the stress energy tensor has the same form in all inertial frames. One can construct a set of test particle world-lines which are exponentially expanding, as we have done here, and these observers would say that mass-energy is being created spontaneously by the universal expansion. However, one can also find a set of test particles whose world-lines are initially converging (the acceleration equation only tells us that $\ddot{a} > 0$, and one can have test particles with $\dot{a} < 0$ initially). Such observers would not agree that mass-energy is being created. One can also construct a congruence of world-lines which are tilted (i.e. in a state of motion) with respect to our comoving observers, and they would also see vanishing momentum density for the scalar field. The usual response to this is to argue that the pressure is not precisely $P = -\rho c^2$, rather there will be a small correction, either due to the field velocity $\dot{\phi}$ or due to the presence of other matter fields, which will break the exact invariance of $T^{\mu\nu}$ under Lorentz boosts.

Finally, in the introduction, we raised the question ‘what happened before the initial singularity?’. In the standard hot big bang, the initial singularity is unavoidable, and such a question is probably best answered by saying it is meaningless. With the inflationary equation of state, we have seen that the general solution for the expansion factor is the sum of exponentially growing and decaying terms (29.4). Generically one would imagine that the coefficients of both terms would be non-zero, in which case, at very early times, the negative exponential term would come to dominate and the universe would be collapsing rather than expanding. One might therefore claim that inflationary models one can have an initially collapsing universe which contracts to a minimum size and then bounces, and one could similarly argue for the possibility of repeated cycles of expansion and contraction. In the context of the models developed here, however, this is a misconception. There is a quantitative difference in the behavior of a scalar field in a contracting universe as compared the the expanding model considered above. In the latter case, H is positive and the term $3H\dot{\phi}$ in the equation of motion is a friction, and the evolution of the field will relax towards the slowly rolling terminal velocity solution. In a collapsing phase H is negative, so we have negative friction. In this case the slowly rolling solution — while possible, since the system as a whole is time symmetric — is an unstable one. For generic initial conditions going into a ‘big-crunch’ we do not expect the field to become potential dominated, and so the inflationary equation of state will not arise.

29.5 Problems

29.5.1 Inflation

Derive the general form of the solution for the scale factor $a(t)$ in an inflationary universe — i.e. one in which $P = -\rho c^2$.

Chapter 30

Observations in FRW Cosmologies

30.1 Distances in FRW Cosmologies

The distance to an object is determined by two quantities: the current curvature radius a_0 , which sets the overall scale, and the comoving distance ω . Observationally the more accessible quantities are the expansion rate H_0 and the redshift z , the latter being rather accurately measurable from the shift of lines in the spectrum. In the next two subsections we show how these two sets of variables are related.

30.1.1 Scale Factor vs Hubble Parameter

The energy equation tells us that

$$H^2 = \frac{8}{3}\pi G\rho \pm \frac{c^2}{a^2} \quad (30.1)$$

with the positive sign for open cosmologies and *vice versa*. With the definition of the critical density and the density parameter Ω , this provides a connection between the scale factor and the Hubble parameter. In what follows, I shall assume an open cosmology (from which the Einstein-de Sitter results are obtained as the limiting case). The results for closed cosmologies are almost identical, save for appropriate sign changes and replacing hyperbolic trigonometric functions by their regular counterparts.

If we specify the present matter density ρ_0 and expansion rate H_0 , or equivalently the expansion rate and density parameter Ω , then the present value of the scale factor is

$$a_0 = \frac{c/H_0}{\sqrt{1-\Omega_0}}. \quad (30.2)$$

Note that $a_0 \rightarrow \infty$ for $\Omega_0 \rightarrow 1$; we will return to this presently.

30.1.2 Redshift vs Comoving Distance

Now consider the redshift z . As already discussed, the observed wavelength for a photon which left a comoving source at time t is equal to the wavelength at emission times the factor by which the scale factor has grown. Defining the *redshift* as $1+z = \lambda_{\text{obs}}/\lambda_{\text{em}}$ we have

$$1+z = \frac{a_0}{a(t_{\text{em}})}. \quad (30.3)$$

This gives $z = z(t_{\text{em}})$ if the expansion history is known. The comoving distance ω of the source can also be related to the time of emission, and hence to the redshift, since

$$\omega = c \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)}. \quad (30.4)$$

To make this connection between ω and z it is most convenient to work in *conformal time* η defined such that $d\eta = cdt/a$, and for which $d\omega = -d\eta$. The energy equation, written in terms of $a' = da/d\eta$ rather than $\dot{a} = da/dt$ is

$$a' = \sqrt{\frac{8\pi G\rho a^4}{3c^2} + a^2}. \quad (30.5)$$

To evolve this backwards in time we need to know how the density ρ changes with time. This is given by the equation of continuity: $\dot{\rho} = -3(\dot{a}/a)(\rho + P/c^2)$ so $\rho' = d\rho/d\eta$ is

$$\rho' = -3\frac{a'}{a}(\rho + P/c^2). \quad (30.6)$$

Equations (30.5) and (30.6), together with a specification of the ‘equation of state’ for the matter content, can be integrated backwards in time to give $a(\eta)$ and $\rho(\eta)$. Specifically, one might specify the current expansion rate H_0 and the current density parameters Ω_m , Ω_r , Ω_Λ etc for the different constituents (which have $P = 0$, $P = \rho c^2/3$, $P = -\rho c^2$ respectively). The final scale factor is then given by (30.2), with $\Omega_0 = \Omega_m + \dots$ the total density parameter. Having solved these equations — in generally this must be done numerically, though this is quite straightforward — for $a(\eta) = a(-\omega)$ (taking $\eta_0 = 0$) gives the redshift $z(\omega) = a_0/a(-\omega) - 1$ and inverting this gives ω , the comoving distance, as a function of redshift.

While these integrations cannot generally be performed analytically, there are several simple and illuminating special cases, and the more general behavior can be qualitatively understood by interpolating between these cases.

First, consider a very low density universe: $\Omega \ll 1$. In this case we can neglect the density entirely and (30.5) says $a' = a$, or $da/a = d\eta$. Integrating this up gives $\log(a_0/a) = -\eta$, or, with $\omega = -\eta$ and using the definition of the redshift, this is $\omega(z) = \log(1+z)$. As a quick sanity check note that for $z \ll 1$ this gives $\omega \simeq z$, whereas $a_0 \simeq H_0/c$ for $\Omega_0 \rightarrow 0$ and hence the local distance-redshift relation is $dl = a_0\omega \simeq H_0 z/c$. For high redshift the comoving distance increases without limit in this model, but only logarithmically. However, this is somewhat academic, since Ω_m is almost certainly not much less than 0.05, so the Universe was almost certainly matter dominated at some not too distant (i.e. logarithmically recent) time in the past.

Another case of much interest is the flat Universe, in which case we can ignore the second term in the square root sign in (30.5). Consider the case $\Omega = \Omega_m = 1$: The density is then $\rho = \rho_0(a_0/a)^3$ and (30.5) becomes $da/\sqrt{a} = (H_0 a_0^{3/2}/c)d\eta$. Integrating this and dividing by $\sqrt{a_0}$ gives $2(1 - \sqrt{a/a_0}) = H_0 a_0 \omega/c$. This is somewhat awkward since, as already mentioned, $a_0 \rightarrow \infty$ for a flat universe. However, in this case, where $S_k(\omega) = \omega$, the metric is

$$ds^2 = a^2(\tau)(-d\eta^2 + d\omega^2 + \omega^2 d\sigma^2). \quad (30.7)$$

However, this is invariant if we re-scale $a \rightarrow a' = \alpha a$, with α some constant, provided we also re-scale comoving distance $\omega \rightarrow \omega' = \omega/\alpha$, and similarly re-scale the conformal time η . If we choose α such that $a_0 = 2c/H_0$ and drop the prime we have

$$\omega = 1 - \frac{1}{\sqrt{1+z}}. \quad (30.8)$$

Note that for $z \ll 1$ we have $\omega \simeq z/2$ so the local distance-redshift relation is $dl = a_0\omega \simeq H_0 z/c$ just as before (the normalization of the present scale factor $a_0 = 2c/H_0 \simeq 8000\text{Mpc}$ was chosen for this reason). Note also that the comoving distance tends to a finite limit $\omega = 1$ as $z \rightarrow \infty$. This is the distance to the horizon.

Finally, consider the case $\Omega = \Omega_\Lambda = 1$. In this case the density is constant and (30.5) becomes $a' = H_0 a^2/c$ or $da/a^2 = (H_0/c)d\eta$. Integrating this and with $a_0 = 2c/H_0$ we have

$$\omega = z/2. \quad (30.9)$$

This again has the same local distance-redshift relation $dl = H_0 z$ as the other cases and here the comoving distance increases without limit as $z \rightarrow \infty$ reflecting the fact that the Λ -dominated

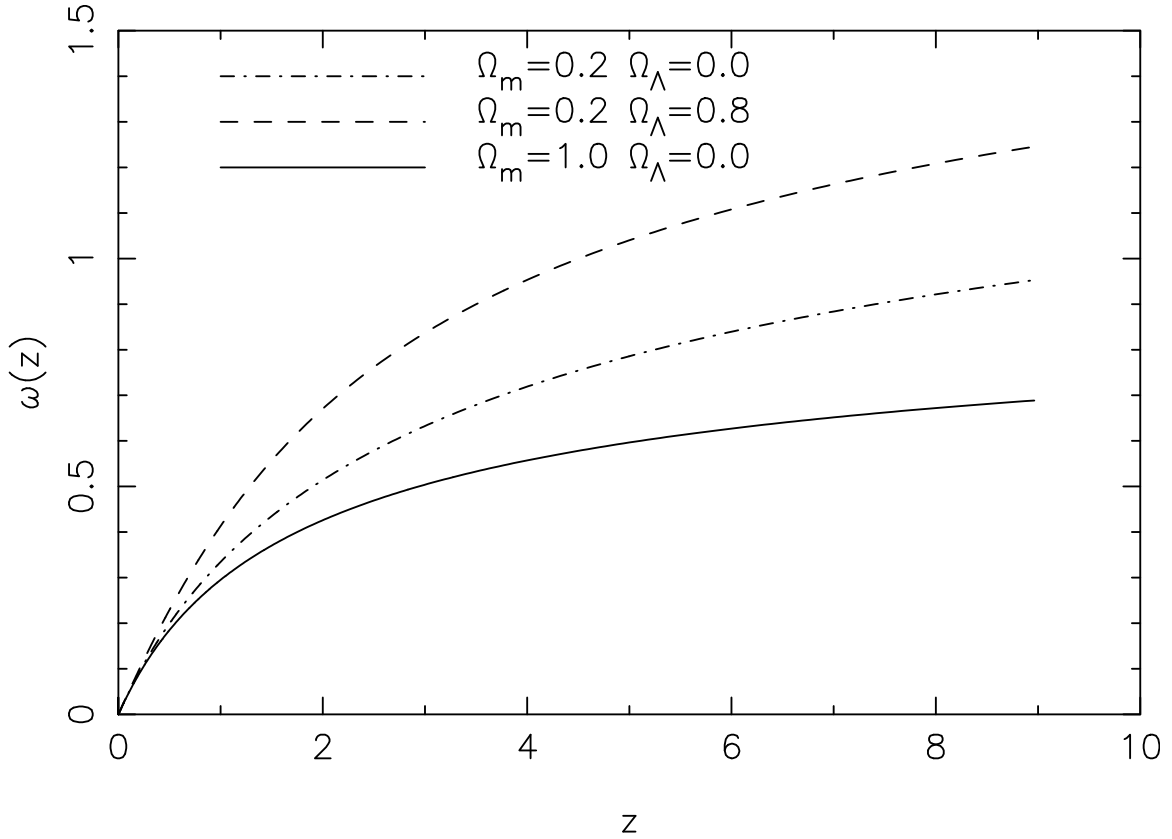


Figure 30.1: Comoving distance is plotted versus redshift for three illustrative models.

universe has no horizon. Again, this is somewhat academic since we know that the matter density is not negligible. Figure 30.1 shows the comoving distance for three plausible choices of cosmological parameters.

30.2 Angular Diameter and Luminosity Distances

If we place identical copies of some extended spherical source at various distances then it will have an angular size $\Delta\theta$ and apparent luminosity (i.e. bolometric flux density) F which are functions of the redshift. For small redshifts the angular size scales inversely with distance D and the luminosity scales as $1/D^2$. It is useful to define an *angular diameter distance* $D_a(z)$ such that $\Delta\theta \propto 1/D_a(z)$ for all z and a *luminosity distance* D_l such that $F \propto 1/D_l^2$.

To obtain D_a , consider a small spherical object of physical radius Δr (figure 30.2). The metric (28.35) tells us that the size is related to the angular radius by $\Delta r = a(z) \sinh \omega \Delta\theta$. The angular diameter distance is then defined such that $\Delta r = D_a \Delta\theta$, or

$$D_a(z) = \frac{\Delta r}{\Delta\theta} = a(z) \sinh \omega(z). \quad (30.10)$$

For low redshift $a(z) \simeq a_0$ and $D_a \simeq a_0 \sinh \omega$, and this increases linearly with redshift. However, for $z \gtrsim 1$ the scale factor becomes important. For the Einstein - de Sitter model $a \propto \eta^2 = (1 - \omega)^2$ so $D_a \propto (1 - \omega)^2 \omega$. This is maximized for $\omega = 1/3$, corresponding to $z = 5/4$, and for larger redshift objects of a given physical size become larger with increasing distance. For the Λ -dominated model $\omega \propto z$, while $a \propto 1/(1 + z)$ so in this case D_a tends asymptotically to a constant value as $z \rightarrow \infty$.

A simple way to establish the luminosity distance D_l is to consider a black-body source with rest-frame temperature T_{em} . The intrinsic bolometric luminosity is proportional to the area times

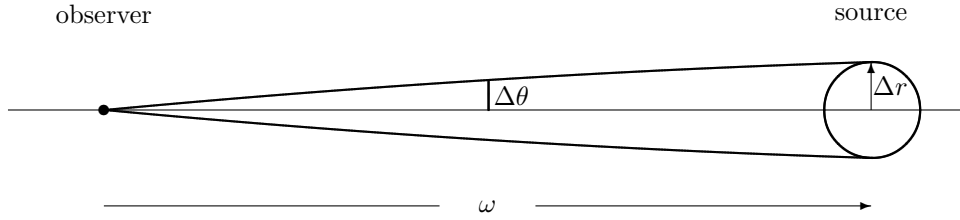


Figure 30.2: A source of physical radius Δr lies at a comoving distance ω from an observer. According to the metric (28.35), the angle and object size are related by $\Delta r = a \sinh \omega \Delta \theta$ where a is the scale factor at the time the light left the source. The angular diameter distance is therefore $D_a = a(z) \sinh \omega(z)$.

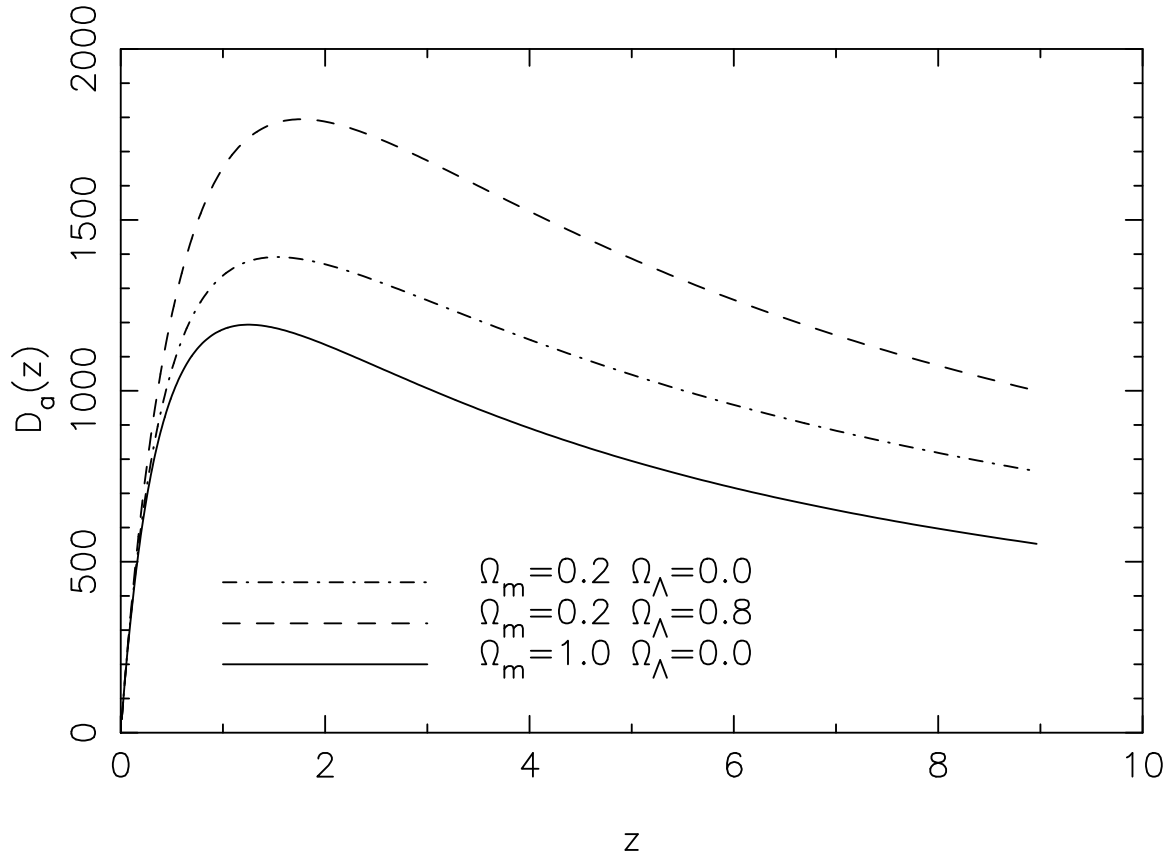


Figure 30.3: Angular diameter distance is plotted versus redshift for three illustrative models.

the fourth power of the temperature $L \propto \Delta r^2 T_{\text{em}}^4$ (the Stefan-Boltzmann law). The observed temperature is lower than T_{em} because of the redshift: $T_{\text{obs}} = T_{\text{em}}/(1+z)$, and the bolometric flux density is proportional to the product of the solid angle $\Delta \Omega$ times T_{obs}^4 . But $\Omega = \pi \Delta r^2 / D_a^2$, so the flux density is $F \propto L(1+z)^{-4} / D_a^2$ or $F \propto L / D_l^2$ with

$$D_l(z) = (1+z)^2 D_a(z). \quad (30.11)$$

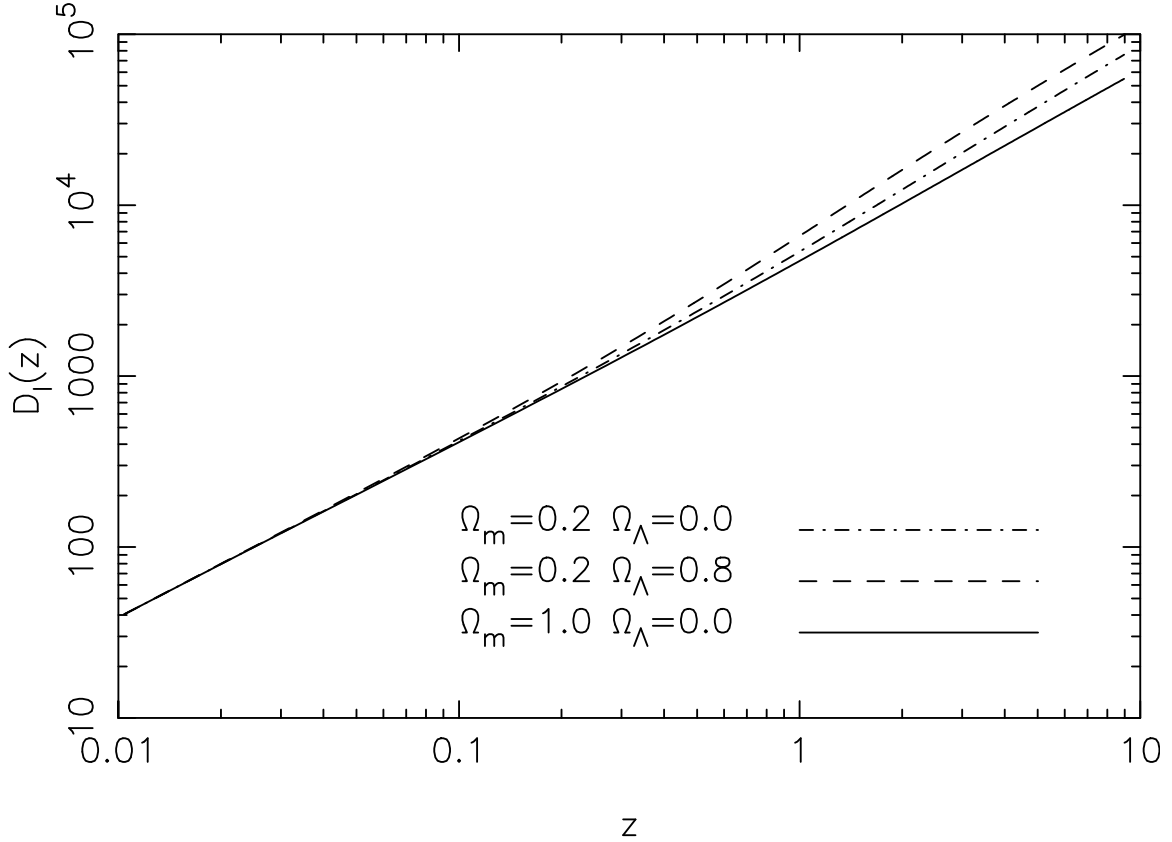


Figure 30.4: Luminosity distance is plotted versus redshift for three illustrative models.

More generally, from measurement of the monochromatic flux density F_ν at observer frequency ν we can infer the luminosity $L_{\nu'}$ at rest-frame frequency $\nu' = (1+z)\nu$ as follows: First, note that an observer cosmologically close to the source — i.e at a distance $r \ll c/H$ — will see brightness

$$I_{\nu'} = \frac{L_{\nu'}}{A\Delta\Omega} = \frac{L_{\nu'}}{(4\pi r^2)(\pi\Delta r^2/r^2)} = \frac{L_{\nu'}}{4\pi^2\Delta r^2}. \quad (30.12)$$

Now the transformation law for the surface brightness is $I_\nu/\nu^3 = \text{constant}$, or $I_\nu = I_{\nu'}(\nu/\nu')^3 = I_{\nu'}/(1+z)^3$. The observed flux density is then

$$F_\nu = \int d\Omega I_\nu = \pi \left(\frac{\Delta r}{D_a} \right)^2 I_\nu = \frac{L_{\nu'=(1+z)\nu}}{4\pi D_a(z)^2(1+z)^3}. \quad (30.13)$$

Turning this around gives

$$L_{\nu'} = 4\pi D_a^2(1+z)^3 F_{\nu'/(1+z)}. \quad (30.14)$$

Note that $d\nu = d\nu'/(1+z)$ so

$$F_\nu d\nu = \frac{L_{\nu'} d\nu'}{4\pi D_a(z)^2(1+z)^4} = \frac{L_{\nu'} d\nu'}{4\pi D_l(z)^2} \quad (30.15)$$

and integrating over all frequency gives

$$F = \int d\nu F_\nu = \frac{L}{4\pi D_l(z)^2} \quad (30.16)$$

in accord with the result obtained for a black-body emitter.

30.3 Magnitudes and Distance Moduli

Astronomers like to quote flux density in the *magnitude scale* defined such that $m = -2.5 \log_{10} F + \text{constant}$. Usually, fluxes are measured through some standard filter. The flux in filter band b is then

$$F_b = \frac{\int d\nu F_\nu R_b(\nu)}{\int d\nu R_b(\nu)}. \quad (30.17)$$

where $R_b(\nu)$ is the dimensionless *transmission function* describing the throughput of the filter. The reason for the factor 2.5 is historical, but for rough calculations is a happy coincidence that $2.5 \log_{10}$ is very close to the natural logarithm. Operationally, the magnitude scale is conveniently calibrated by measuring the magnitude relative to some standard object. In the *Vega magnitude scale*, the star Vega has magnitude m_b identically zero in all passbands. Traditionally, the *apparent magnitude* of an object is then -2.5 times the log of its flux density relative to that of Vega:

$$m_b = -2.5 \log_{10} \left[\frac{\int d\nu F(\nu) R_b(\nu)}{\int d\nu F_{\text{Vega}}(\nu) R_b(\nu)} \right]. \quad (30.18)$$

For nearby objects, and in the absence of absorption, the flux varies inversely with the square of the distance D . One can then quote the distance to an object in magnitudes also. A star identical to Vega, but ten times farther away will be 100 times fainter, so it will have an apparent magnitude of $+5$. The *distance modulus* is defined as

$$\text{DM} = 5 \log_{10}(D/10\text{pc}) \quad (30.19)$$

where one *parsec* is $1\text{pc} \simeq 3.0 \times 10^{18}\text{cm}$.

30.4 K-Corrections

For cosmologically nearby objects, and in the absence of absorption, the *absolute magnitude* M_b in some band b is related to distance and apparent magnitude m_b by

$$M_b = m_b - \text{DM}. \quad (30.20)$$

For cosmologically distant objects things are more complicated since if we observe in a filter with central wavelength λ then this receives photons which were emitted around wavelength $\lambda' = \lambda/(1+z)$, so the best we can really hope to obtain is the absolute magnitude in a filter b' with transmission function $R_{b'}(\nu) = R_b(\nu(1+z))$.

There is an additional, often overlooked, subtlety which is worth mentioning here. Many detectors such as CCDs better approximate photon counting systems than they do energy measuring devices. A photon counting system cannot measure the integrated energy appearing in (30.18). What they can measure is the flux of *photons*. Since the photon flux n_ν is equal to the energy flux divided by the energy per photon $n_\nu = F_\nu/\hbar\nu$. For very narrow band filters this is a negligible effect, but for the broad-band filters often used (because they are more efficient) the distinction can be important. We can, however, use the traditional definition of magnitudes, if we simply replace $R_b(\nu)$ by $R'_b(\nu) \equiv R_b(\nu)/\nu$. We will assume that this substitution has been made, and in what follows we will drop the prime for clarity. With $F(\nu)d\nu = L(\nu' = (1+z)\nu)d\nu'/4\pi D_1^2(z)$ the apparent magnitude is then

$$m_b = -2.5 \log_{10} \left[\left(\frac{10\text{pc}}{D_1(z)} \right)^2 \frac{\int d\nu L((1+z)\nu) R_b(\nu)}{\int d\nu L_{\text{Vega}}(\nu) R_b(\nu)} \right] \quad (30.21)$$

Changing integration variable in the upper integral to $\nu' = (1+z)\nu$ and dropping the prime gives

$$m_b = \text{DM} - 2.5 \log_{10} \left[\frac{\int d\nu L(\nu) R_b(\nu/(1+z))}{(1+z) \int d\nu L_{\text{Vega}}(\nu) R_b(\nu)} \right] \quad (30.22)$$

or as

$$m_b = \text{DM} - 2.5 \log_{10} \left[\frac{\int d\nu L(\nu) R_b(\nu)}{\int d\nu L_{\text{Vega}}(\nu) R_b(\nu)} \right] - 2.5 \log_{10} \left[\frac{\int d\nu L(\nu) R_b(\nu/(1+z))}{(1+z) \int d\nu L(\nu) R_b(\nu)} \right]. \quad (30.23)$$

The center term on the right hand side is the absolute magnitude M_b , so this is

$$M_b = m_b - \text{DM}(z) + k_b(z) \quad (30.24)$$

where the so called *k-correction* is

$$k_b(z) \equiv -2.5 \log_{10} \left[\frac{\int d\nu L(\nu) R_b(\nu/(1+z))}{(1+z) \int d\nu L(\nu) R_b(\nu)} \right]. \quad (30.25)$$

The k-correction depends on the filter response function and on the *spectral energy distribution* (SED) of the source, and accounts for the red-shifting of the source spectrum.

These correction functions have been computed for various types of galaxies so, if one has a good idea of what type of galaxy one is dealing with, one can obtain a rough estimate of the absolute luminosity in this way. One can do rather better than this. Imagine one observes a galaxy at redshift $z = 0.6$ in the *I*-band. The central wavelength is $\lambda_I \simeq 8000\text{\AA}$, so in the rest-frame this corresponds to $\lambda \simeq 5000\text{\AA}$ which is close to the central wavelength in the *V*-band. Thus *I*-band observations of galaxies at this redshift provide one with the rest-frame *V*-band absolute magnitude with very small galaxy type dependent corrections. Generalizing this, with $I \rightarrow b$ and $V \rightarrow b'$ we have

$$M_{b'} = m_b - \text{DM}(z) + k_{bb'}(z) \quad (30.26)$$

where the *generalized k-correction* is

$$k_{bb'}(z) \equiv -2.5 \log_{10} \left[\frac{\int d\nu L_{\text{Vega}}(\nu) R_{b'}(\nu)}{\int d\nu L_{\text{Vega}}(\nu) R_b(\nu)} \right] - 2.5 \log_{10} \left[\frac{\int d\nu L(\nu) R_b(\nu/(1+z))}{(1+z) \int d\nu L(\nu) R_{b'}(\nu)} \right]. \quad (30.27)$$

Chapter 31

Linear Cosmological Perturbation Theory

Having explored the idealized perfectly homogeneous FRW models we now explore departures from perfectly uniform density and expansion using linear perturbation theory. We first consider perturbations of the zero pressure models in §31.1. These are applicable at late times when the pressure is negligible. These models are also valid at early times for sufficiently long wavelength perturbations which are ‘outside the horizon’, and for which pressure gradients are negligible. In §31.2 we consider the effects of pressure, which is important at early times for sub-horizon scale perturbations. We discuss the different perturbation modes which are present when there are multiple coupled fluids (such as the radiation and plasma) and we also discuss diffusive damping.

Having laid some of the theoretical groundwork we describe in §31.3 how ideas about the nature of the matter perturbation have evolved over the years.

In the following chapter we will discuss the generation of cosmological perturbations. Specifically, we consider three spontaneous generation of perturbations; quantum fluctuations in inflation, and self ordering fields.

31.1 Perturbations of Zero-Pressure Models

31.1.1 The Spherical ‘Top-Hat’ Perturbation

In a matter dominated background cosmology the simplest way to construct a perturbation is to excise a sphere of matter and replace it with a smaller sphere of the same gravitational mass as illustrated in figure 31.1. This generates a ‘top-hat’ positive density perturbation. One can lay down more than one such perturbation, provided the walls of the perturbations do not overlap, and this simple type of model for inhomogeneity is sometimes referred to as the ‘Swiss-Cheese model’. While highly idealized, this simple model illustrates most of the features of the perturbations of arbitrary shape.

The trajectories $R(t)$ of comoving observers on the surface of a sphere containing mass M obey the energy equation

$$\dot{R}^2 = 2GM/R + 2E \quad (31.1)$$

with $E = \text{constant}$. The solutions of this equation form a two parameter family; we can perturb the energy E , and we can also change the time of the big-bang, or we can make some combined perturbation. In either case, since the mass is fixed, we have $\rho(t)R(t)^3 = \rho'(t)R'(t)^3$, where primed and un-primed quantities refer to the interior and exterior respectively. For a small perturbation $\rho' = \rho(1 + \delta\rho/\rho)$, with $\delta\rho/\rho \ll 1$, to first order in the amplitude we have

$$\frac{\delta\rho}{\rho} = -3\frac{\delta R}{R} \quad (31.2)$$

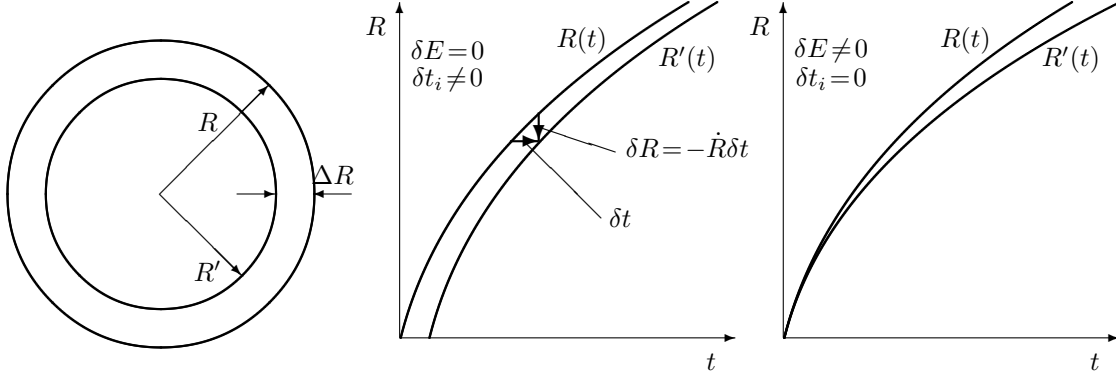


Figure 31.1: One can generate a perturbation of a dust-filled cosmology by excising a sphere of matter and replacing it with a smaller sphere of radius R' . The middle panel illustrates a decaying perturbation produced by a ‘delayed bang’. The right panel shows the more interesting growing perturbation that can be generated by perturbing the energy of the sphere. The space-time in the gap is Schwarzschild.

where $\delta\rho/\rho$ indicates the fractional perturbation to the density at a given time, and similarly for $\delta R/R$. This gives the density perturbation from the perturbation δR to the trajectory. Now we can also determine δR from the perturbation δt_R in the time $t(R)$ to arrive at a given radius R : $\delta t \equiv t'(R) - t(R)$ since $\delta R = -\dot{R}\delta t_R$ (see figure 31.1) and therefore

$$\frac{\delta\rho}{\rho} = 3\frac{\dot{R}}{R}\delta t_R = 3H\delta t_R. \quad (31.3)$$

If we make a perturbation by delaying the bang time within the sphere, but keeping the energy unaltered, the time delay δt_R is constant $\delta t_R = \delta t_i$, so this produces a density perturbation with $\delta\rho/\rho \propto H$ which says that the density perturbation will decay as $1/t$. For perturbations generated in the early universe, these decaying mode density perturbations will be negligible at late times.

A more interesting way to perturb the universe is to keep the bang time fixed, but to perturb the energy of the explosion. Now the interior will again resemble part of another FRW universe, but with lower energy $E' = E + \delta E$ (the quantity δE here being negative for a positive density perturbation). The energy equation for the perturbed radius is

$$\left(\frac{dR'}{dt}\right)^2 = 2GM/R' + 2E + 2\delta E. \quad (31.4)$$

Taking the square root we have, for the time taken to obtain a radius R ,

$$t' = \int dt = \int_0^R \frac{dR'}{\sqrt{\frac{2GM}{R'} + 2E + 2\delta E}} = t - \delta E \int_0^R \frac{dR'}{\left(\frac{2GM}{R'} + 2E\right)^{3/2}} + \dots \quad (31.5)$$

where, in the second step, we have made a Taylor expansion and where \dots denotes terms of 2nd order and higher in δE .

The perturbation in the time to reach radius R is then

$$\delta t_R = -\delta E \int_0^R \frac{dR'}{\left(\frac{2GM}{R'} + 2E\right)^{3/2}} \quad (31.6)$$

and from (31.3) the density perturbation is therefore

$$\frac{\delta\rho}{\rho} = -3\delta E \frac{\dot{R}}{R} \int_0^R \frac{dR'}{\left(\frac{2GM}{R'} + 2E\right)^{3/2}}. \quad (31.7)$$

We can identify two interesting limiting cases

- At early times in all models (and for all time in the Einstein - de Sitter model) the energy constant E is negligible compared to $2GM/R$, the expansion velocity is $\dot{R} \simeq \sqrt{2GM/R}$, the integral is $\int dR(2GM/R)^{-3/2} = (2/5)(2GM)^{-3/2}R^{5/2}$ so the density perturbation is

$$\frac{\delta\rho}{\rho} = -\frac{6}{5} \frac{\delta E}{2GM/R} \quad \Omega \simeq 1. \quad (31.8)$$

This type of density perturbation therefore grows with time as $\delta\rho/\rho \propto R \propto t^{2/3} \propto (1+z)^{-1}$. This was first shown by Lifshitz.

- At late times in a low density universe the energy constant term dominates; the expansion velocity $\dot{R} \rightarrow \text{constant}$ while the integral is $E^{-3/2} \int dR = E^{-3/2}R$, with the net result that the density perturbation becomes asymptotically constant

$$\frac{\delta\rho}{\rho} = -\frac{3}{2} \frac{\delta E}{E} \quad \Omega \ll 1. \quad (31.9)$$

We say that the density perturbation ‘freezes out’ in a low density universe as it ‘peels away’ from the $\Omega = 1$ solution. For $\Omega \ll 1$ the expansion velocity \dot{a} becomes asymptotically constant, so the expansion rate decays as $H \propto 1/a$, and therefore $\Omega \sim G\rho/H^2 \propto 1/a$ also; this means that $\Omega \propto (1+z)$ in a low density universe, and therefore this freeze-out occurs at a redshift $1+z \sim 1/\Omega_0$, though the transition is, in reality, a gradual one.

Note that we could have easily guessed the growth rate by simply arguing that the perturbation to the binding energy of the region is $\delta\phi \sim G\delta M/R \sim (\delta\rho/\rho)(\rho R^2) \propto (\delta\rho/\rho)/R$, and is constant.

The phenomenon described here is often called ‘gravitational instability’, which is something of a misnomer. True, the density perturbation grows with time, but the perturbation to the binding energy is constant in time; the density contrast grows at just the rate required to maintain this.

It is also of interest to calculate the perturbation to the expansion rate $H = \dot{R}/R$ since, as we shall see, this is also a directly observable quantity. The expansion rate within the perturbation is

$$H' = \sqrt{\frac{2GM}{R^3} + \frac{2(E + \delta E)}{R^2}} \quad (31.10)$$

so, to first order in δE , the difference between perturbed and un-perturbed expansion rates is

$$(\delta H)_R = H' - H = \frac{\delta E/R}{\sqrt{\frac{2GM}{R} + 2E}}. \quad (31.11)$$

However, this, as we have been careful to indicate, is the perturbation to the expansion rate *at a fixed radius* R . This is not what is observed, which is the perturbation to the expansion rate *at a given time*. The latter is

$$(\delta H)_t = (\delta H)_R + \dot{H}(\delta t)_R \quad (31.12)$$

Now $\dot{H} = \ddot{R}/R - \dot{R}^2/R^2$, which, with $\ddot{R} = -GM/R^2$, is $\dot{H} = -(3GM/R^3 + 2E/R^2)$ and we have

$$\left(\frac{\delta H}{H}\right)_t = \frac{\delta E}{\frac{2GM}{R} + 2E} \left[1 - \left(\frac{3GM}{R} + 2E\right) \left(\frac{2GM}{R} + 2E\right)^{1/2} \frac{1}{R} \int_0^R \frac{dR}{\left(\frac{2GM}{R} + 2E\right)^{3/2}} \right] \quad (31.13)$$

Again, this rather ugly expression becomes a lot simpler in the limits $2GM/R \gg E$ (i.e. $\Omega \simeq 1$) and $2GM/R \ll E$ (i.e. $\Omega \ll 1$). In the former case we have

$$\left(\frac{\delta H}{H}\right)_t = \frac{2}{5} \frac{\delta E}{2GM/R} \quad \Omega \simeq 1. \quad (31.14)$$

and in the latter limit the fractional perturbation to the expansion rate vanishes.

Comparing with (31.8) for the amplitude of the density perturbation we see that for growing mode perturbations in the Einstein - de Sitter background the perturbation to the Hubble rate is just minus one third of the fractional density perturbation:

$$\frac{\delta H}{H} = -\frac{1}{3} \frac{\delta \rho}{\rho}. \quad (31.15)$$

The fractional perturbation in the expansion rate therefore also grows as $\delta H/H \propto t^{2/3}$. The velocity difference between a pair of particles which straddle the gap is, in this case,

$$-\delta v = H\delta R - R\delta H = HR \left(\frac{\delta R}{R} - \frac{\delta H}{H} \right) = \frac{2}{3} HR \frac{\delta \rho}{\rho}. \quad (31.16)$$

Since $HR \propto \sqrt{1/R}$, the peculiar velocity grows as $\delta v \propto \sqrt{R} \propto t^{1/3}$. Again, this growth rate can be obtained by arguing that the potential $\delta\phi$ is constant in time, but $\delta\phi \sim \delta(v^2) \sim v\delta v$ hence $\delta v \propto 1/v \propto \sqrt{R}$.

To summarize, we have shown that there are two modes for spherical density perturbations in a zero-pressure cosmology. The decaying modes have $\delta\rho/\rho \propto 1/t$. The more interesting growing modes have fractional perturbations in the density and expansion rate which grow as $t^{2/3}$ for $\Omega \simeq 1$. In low density models the density perturbations become asymptotically constant, and the expansion rate perturbation asymptotically vanishes.

The above analysis is Newtonian. This is adequate, provided we are considering perturbations at such times that $\dot{R} \ll c$. That is, for perturbations which are smaller than the current horizon scale. This way of modeling perturbations is, however, not restricted to this regime. In the case of perturbations to a closed universe, for instance, the embedding diagram for the type of density perturbation is partial closed model within the perturbation matching on to a shell of horn-like Schwarzschild geometry which then matches smoothly onto the exterior, this being a closed model with larger radius of curvature.

31.1.2 General Perturbations

The spherical perturbations considered above are highly idealized. In general the density $\rho(\mathbf{x}, t)$ is some arbitrary function of physical position \mathbf{x} . The expansion velocity $\mathbf{v}_{\text{phys}} = \dot{\mathbf{x}}$ will also, in general, be some related function of \mathbf{x} and t . In what follows it is more convenient to work in *comoving spatial coordinate* $\mathbf{r} = \mathbf{x}/a(t)$, where $a(t)$ is the global scale factor (i.e. that for the unperturbed cosmology). This is the same comoving coordinate we denoted by $\boldsymbol{\omega}$ earlier. We define the density perturbation

$$\delta(\mathbf{r}, t) = \frac{\delta\rho(\mathbf{r}, t)}{\bar{\rho}(t)} = \frac{\rho(a\mathbf{r}, t) - \bar{\rho}(t)}{\bar{\rho}(t)}. \quad (31.17)$$

where $\bar{\rho}$ is the mean density. It is also more convenient to define a *peculiar velocity*

$$\mathbf{v} = \mathbf{v}_{\text{phys}} - H\mathbf{x}, \quad (31.18)$$

which is the departure from uniform expansion, and to define a *comoving peculiar velocity* \mathbf{u} , which is the rate of change of \mathbf{r} with time:

$$\mathbf{u} = \dot{\mathbf{r}} = \frac{d(\mathbf{x}/a)}{dt} = \frac{\dot{\mathbf{x}}}{a} - \frac{\dot{a}\mathbf{x}}{a^2} = \frac{\mathbf{v}_{\text{phys}} - H\mathbf{x}}{a} = \frac{\mathbf{v}}{a}. \quad (31.19)$$

Our experience with spherical perturbations suggests that we should be able to decompose a general initial perturbation into growing and decaying components

$$\delta(\mathbf{r}, t_i) = \delta^+(\mathbf{r}, t_i) + \delta^-(\mathbf{r}, t_i). \quad (31.20)$$

where the split into growing and decaying modes is determined by the initial perturbation δ and its rate of change $\dot{\delta}$. In the linear approximation, each of these modes then evolves completely independently, with growth factors $D^\pm(t) = \delta^\pm(t)/\delta^\pm(t_i)$ and the final density perturbation is

$$\delta(\mathbf{r}, t) = D^+(t)\delta^+(\mathbf{r}, t_i) + D^-(t)\delta^-(\mathbf{r}, t_i). \quad (31.21)$$

This is not quite true, however. It turns out that in two or more dimensions there are modes of a completely different character, which did not appear in our analysis of spherical perturbations since they do not admit any spherically symmetric states. To see why, note that, at a given time, we can make a Fourier transform and write the density perturbation field as a Fourier synthesis

$$\delta(\mathbf{r}, t) = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(t) e^{i\mathbf{k} \cdot \mathbf{r}}. \quad (31.22)$$

That is, we are writing the density field as a sum of plane wave ripples with a pattern which is fixed in comoving coordinates (we are assuming here that the wavelength of the perturbation is much less than the curvature scale, so we can consider our spatial coordinates \mathbf{r} to be flat-space Cartesian coordinates). If we do the same thing to the velocity field

$$\mathbf{u}(\mathbf{r}, t) = \sum_{\mathbf{k}} \mathbf{u}_{\mathbf{k}}(t) e^{i\mathbf{k} \cdot \mathbf{r}} \quad (31.23)$$

then it is clear that, for each wave-vector, there are four degrees of freedom — the values of $\delta_{\mathbf{k}}$ and $\mathbf{u}_{\mathbf{k}}$ — not two. The spherical perturbations are special in that if we make a Fourier decomposition, the velocity coefficients $\mathbf{u}_{\mathbf{k}}$ are parallel to the wave-vector \mathbf{k} . These are like longitudinal or compressional sound waves. This is also the characteristic of potential flows. In addition to these so-called *scalar modes* (not to be confused with scalar fields), there are *vector modes*. These have a quite different character, and have non-zero transverse velocity (i.e. the velocity in the directions orthogonal to the wave vector). These modes, for example, can have non-vanishing angular momentum, and are often referred to as ‘torsional modes’. Henceforth we will simply ignore these vector modes, with the rather lame justification that they are uninteresting since there are no growing vector modes. That is, we will exclusively consider perturbations for which

$$\mathbf{u}(\mathbf{r}, t) = \sum_{\mathbf{k}} \hat{\mathbf{k}} u_{\mathbf{k}}(t) e^{i\mathbf{k} \cdot \mathbf{r}}. \quad (31.24)$$

The first step in deriving the growth rate for density perturbations is to note that the density perturbation will give rise to a small perturbation $\delta\phi$ to the Newtonian gravitational potential *via* Poisson’s equation

$$\nabla_x^2 \delta\phi = 4\pi G \bar{\rho} \delta. \quad (31.25)$$

We use the subscript x here to show that we are taking the gradient with respect to physical coordinate \mathbf{x} , rather than say comoving coordinate \mathbf{r} . Applied to a plane wave density ripple $\delta\phi = \phi_{\mathbf{k}} \exp(i\mathbf{k} \cdot \mathbf{r}) = \phi_{\mathbf{k}} \exp(i(\mathbf{k}/a) \cdot \mathbf{x})$ the Laplacian operator gives $\nabla_x^2 \delta\phi = -k^2 \phi_{\mathbf{k}}/a^2$ and therefore Poisson’s equation becomes the algebraic equation

$$k^2 \phi_{\mathbf{k}} = -4\pi G \bar{\rho} a^2 \delta_{\mathbf{k}}. \quad (31.26)$$

The gradient of the potential perturbation $\delta\phi$ is the so-called *peculiar gravity* and causes particles’ world-lines to deviate from $\mathbf{r} = \text{constant}$. However, as we will work in comoving coordinates, there is a slight subtlety: Imagine a particle moving ballistically with no gravitational forces acting. The particle’s physical velocity \mathbf{v}_{phys} will be constant, but as it moves it will be overtaking particles with progressively higher Hubble velocity, so its *peculiar velocity*, i.e. its velocity relative to the comoving observers it is passing, will decrease in time. If, at some instant, it is moving at velocity \mathbf{v} past observer A then after a time interval δt it will have moved a distance $\delta \mathbf{x} = \mathbf{v} \delta t$, and will be passing observer B who is receding from A with Hubble velocity $H \delta \mathbf{x}$, so B will see the particle passing with velocity $\mathbf{v}' = \mathbf{v} - H \mathbf{v} \delta t$; this implies the equation of motion (in the absence of any peculiar gravity)

$$\dot{\mathbf{v}} = -H \mathbf{v}. \quad (31.27)$$

Thus, in an expanding coordinate system there appears to be a *cosmic drag* acting on the particle. The effect of the density perturbation is to subject the particle to an additional acceleration $-\nabla_x \delta\phi$ and the equation of motion is therefore

$$\dot{\mathbf{v}} = -H\mathbf{v} - \nabla_x \delta\phi. \quad (31.28)$$

In terms of $\mathbf{u} \equiv \dot{\omega} = \mathbf{v}/a$, for which $\dot{\mathbf{u}} = (\dot{\mathbf{v}} - H\mathbf{v})/a$,

$$\dot{\mathbf{u}} = -2H\mathbf{u} - \frac{1}{a^2} \nabla_r \delta\phi \quad (31.29)$$

where $\nabla_r = a\nabla_x$ denotes the derivative with respect to comoving coordinate \mathbf{r} . For a plane-wave density ripple, for which $\nabla_r \delta\phi = i\mathbf{k}\phi_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{r}}$ and the velocity is given by (31.24), this gives

$$\dot{u}_{\mathbf{k}} = -2Hu_{\mathbf{k}} - ik\phi_{\mathbf{k}}/a^2. \quad (31.30)$$

The final step is to invoke the continuity equation to connect the velocity \mathbf{u} and the rate of change of δ . In \mathbf{x}, t coordinates the continuity equation is $\partial\rho/\partial t = -\nabla_x \cdot (\rho\mathbf{v}_{\text{phys}})$. The density $1+\delta$ in comoving coordinates similarly satisfies the continuity equation $\partial(1+\delta)/\partial t = -\nabla_r \cdot ((1+\delta)\mathbf{u})$, the linearized version of which

$$\dot{\delta} = -\nabla_r \cdot \mathbf{u}. \quad (31.31)$$

For a single plane-wave density ripple this is

$$\dot{\delta}_{\mathbf{k}} = -iku_{\mathbf{k}}. \quad (31.32)$$

Taking the time derivative of (31.32), and using (31.30) we obtain

$$\ddot{\delta}_{\mathbf{k}} = -ik\dot{u}_{\mathbf{k}} = ik(2Hu_{\mathbf{k}} + ik\phi_{\mathbf{k}}/a^2) = -H\dot{\delta}_{\mathbf{k}} - (k^2/a^2)\phi_{\mathbf{k}} \quad (31.33)$$

and using (31.26) to eliminate $\phi_{\mathbf{k}}$ we have

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - 4\pi G\bar{\rho}\delta_{\mathbf{k}} = 0. \quad (31.34)$$

This is the equation governing the time evolution of a linearized density ripple of amplitude $\delta_{\mathbf{k}}$. However, since \mathbf{k} does not appear in any of the coefficients, and $\delta_{\mathbf{k}}$ appears only at first order, we can multiply by $e^{i\mathbf{k}\cdot\mathbf{r}}$ and sum over modes to obtain

$$\ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\delta = 0. \quad (31.35)$$

This is a local equation governing the evolution of the density perturbation field $\delta(\mathbf{r}, t)$. It is second order in time, so we need to specify both δ (or the displacement) and $\dot{\delta}$ (or the rate of change of the displacement) as initial conditions.

One can readily show that for $\Omega = 1$, equation (31.35) admits the same growing and decaying solutions obtained from the spherical model. Making the ansatz $\delta(\mathbf{r}, t) = AF(\mathbf{r})t^\alpha$, we have $\dot{\delta} = \alpha AF(\mathbf{r})t^{\alpha-1} = \alpha\delta/t$ and $\ddot{\delta} = \alpha(\alpha-1)AF(\mathbf{r})t^{\alpha-2} = \alpha(\alpha-1)\delta/t^2$ which, in (31.35), and dividing by δ , becomes

$$\alpha(\alpha-1)/t^2 + 2\alpha H/t - 4\pi G\bar{\rho} = 0. \quad (31.36)$$

Now in the Einstein - de Sitter model $a \propto t^{2/3}$ so $H = \dot{a}/a = 2/3t$ while $4\pi G\bar{\rho} = 3H^2/2 = 2/3t^2$ so, multiplying the above equation by t^2 , we have

$$3\alpha^2 + \alpha - 2 = 0 \quad (31.37)$$

with solutions $\alpha = -1, +2/3$. The ‘eigen-modes’ of (31.35) are therefore $\delta^+(\mathbf{r}, t) \propto t^{2/3}$ and $\delta^-(\mathbf{r}, t) \propto t^{-1}$. The general solution of (31.35) is then

$$\delta(\mathbf{r}, t) = \delta^+(\mathbf{r}, t_i) \left(\frac{t}{t_i}\right)^{2/3} + \delta^-(\mathbf{r}, t_i) \left(\frac{t}{t_i}\right)^{-1}. \quad (31.38)$$

Here t_i is the initial time.

Taking the time derivative and multiplying by t gives

$$t\dot{\delta}(\mathbf{r}, t) = \frac{2}{3}\delta^+(\mathbf{r}, t_i) \left(\frac{t}{t_i}\right)^{2/3} - \delta^-(\mathbf{r}, t_i) \left(\frac{t}{t_i}\right)^{-1} \quad (31.39)$$

and solving the above equations for $\delta^+(\mathbf{r}, t_i)$ and $\delta^-(\mathbf{r}, t_i)$ yields

$$\begin{aligned} \delta^+(\mathbf{r}, t_i) &= \frac{3}{5} \left(\delta(\mathbf{r}, t_i) + t\dot{\delta}(\mathbf{r}, t_i) \right) \\ \delta^-(\mathbf{r}, t_i) &= \frac{2}{5} \left(\delta(\mathbf{r}, t_i) - \frac{3}{2}t\dot{\delta}(\mathbf{r}, t_i) \right) \end{aligned} \quad (31.40)$$

Given some initial $\delta(\mathbf{r})$, $\dot{\delta}(\mathbf{r})$, this tells us the amplitudes for the growing and decaying modes.

31.2 Non-zero Pressure and the Jeans Length

So far we have considered $P = 0$; this is a good approximation when the Universe is matter dominated and when the gas is decoupled from the radiation. However, prior to a redshift of about 1000 — known as the redshift of recombination or the redshift of decoupling — the gas was highly ionized, and was tightly coupled to the photons of the microwave background by Thomson scattering. Prior to decoupling it is important to allow for the effect of the pressure of the radiation.

31.2.1 Matter Dominated Era

We will first consider the behavior of perturbations in the relatively narrow window between z_{eq} and z_{dec} . During this period, the density of the radiation-plasma fluid is dominated by the baryons, $\rho \simeq \rho_b \gg \rho_{\text{rad}}$, but the radiation provides a pressure $P = P_{\text{rad}} = \rho_{\text{rad}}c^2/3$.

Assuming the baryons and photons to be tightly coupled, the radiation density is proportional to the $4/3$ power of the matter density, or

$$\rho_{\text{rad}} = \frac{\bar{\rho}_{\text{rad}}}{\bar{\rho}_b^{4/3}} \rho_b^{4/3}. \quad (31.41)$$

where $\bar{\rho}_{\text{rad}}$, $\bar{\rho}_b$ are the mean densities of radiation and baryons. The sound speed is given by

$$c_s^2 = \frac{dP}{d\rho} = \frac{c^2}{3} \frac{d\rho_{\text{rad}}}{d\rho_b} = \frac{4c^2}{9} \frac{\bar{\rho}_{\text{rad}}}{\bar{\rho}_b}. \quad (31.42)$$

The sound speed is therefore on the order $c_s \sim c\sqrt{\rho_{\text{rad}}/\rho_b} \propto a^{-1/2}$. We can define a *sound horizon* to be the distance that sound waves can travel in the age of the Universe, or in one expansion time. The *comoving sound horizon* is $r \sim c_s t/a$. However, in the matter dominated era, $t \propto a^{3/2}$ and hence the comoving sound horizon is constant.

The Fourier decomposition approach (unlike the spherical top-hat) is readily modified to incorporate pressure: For $P \ll \rho$ the gravitational acceleration $-\nabla_x \delta\phi$ must be augmented by the pressure gradient acceleration $-\nabla_x P/\rho$, which, for linear perturbations, is $-c_s^2 \nabla_x \delta$. Including this extra acceleration in (31.33), equation (31.34) becomes

$$\ddot{\delta}_k + 2H\dot{\delta}_k - (4\pi G\rho - c_s^2 k^2/a^2)\delta_k = 0. \quad (31.43)$$

The new extra term here radically changes the behavior of the solutions. For small wavelength (high k) this term will dominate and the solutions will be oscillatory — these are simply adiabatic sound waves (though the pressure here comes from the radiation rather than the kinetic motion of the gas particles as in a conventional sound wave).

One can define the *Jeans wavelength*

$$\lambda_J = 2\pi a/k_J = 2\pi c_s (4\pi G\rho)^{-1/2} \quad (31.44)$$

which separates the growing ($\lambda \gg \lambda_J$) modes from oscillatory ($\lambda \ll \lambda_J$) solutions. To order of magnitude, the Jeans length is just the distance a sound wave can propagate in the age of the universe ($\lambda_J \sim c_s t$) which is just the physical sound horizon.

Perturbations of wavelength $\ll \lambda_J$ will oscillate many times per expansion time (this assumes that diffusive damping is negligible; this will be considered later). However, the equation that these perturbations obey is not a simple harmonic oscillator, as it contains a damping term $2H\dot{\delta}$ and also the frequency of the oscillations $\omega \simeq c_s k/a$ is not constant. To determine the secular evolution of the perturbations — i.e. how the amplitude evolves with time — we proceed as follows: First we make a transformation from δ to an auxiliary field χ defined such that

$$\delta(\mathbf{r}, t) = \chi(\mathbf{r}, t)t^\alpha \quad (31.45)$$

with α constant. The partial time derivatives of δ appearing in (31.43) are then

$$\dot{\delta} = \dot{\chi}t^\alpha + \alpha\chi t^{\alpha-1} \quad (31.46)$$

$$\ddot{\delta} = \ddot{\chi}t^\alpha + 2\alpha\dot{\chi}t^{\alpha-1} + \alpha(\alpha-1)\chi t^{\alpha-2} \quad (31.47)$$

so, on multiplying by t^2 (31.43) becomes

$$\ddot{\chi} + 2(\alpha/t + H)\dot{\chi} + \left(\frac{c_s^2 k^2}{a^2} - \frac{3}{2}H^2 + \frac{2H\alpha}{t} \frac{\alpha(\alpha-1)}{t^2} \right) \chi = 0. \quad (31.48)$$

Now in the matter dominated era $a \propto t^{2/3}$, so $H = \dot{a}/a = 2/3t$, so, if we take $\alpha = -2/3$, the coefficient of $\dot{\chi}$ vanishes and we thereby eliminate the damping term. We then have

$$\ddot{\chi} + \left(\frac{c_s^2 k^2}{a^2} - H^2 \right) \chi = 0. \quad (31.49)$$

This is an un-damped oscillator $\ddot{\chi} + \Omega^2 \chi = 0$ with time varying frequency $\Omega^2 = c_s^2 k^2/a^2 - H^2$. We are here most interested in waves with $\lambda \ll c_s t$ — since longer wavelengths will not have had time to oscillate — but this condition is $k = 2\pi a/\lambda \gg Ha/c_s$. To a good approximation we can neglect the term H^2 in the frequency, so $\Omega \simeq c_s k/a$. The sound speed is $c_s \sim c\sqrt{\rho_{\text{rad}}/\rho_{\text{b}}}$ which is proportional to $a^{-1/2}$, so $\Omega \propto a^{-3/2}$. The frequency decreases as the Universe expands, partly because the wave, having fixed comoving wavelength, is being stretched, and partly because the sound speed is decreasing.

In the limit that the time-scale for variation of the frequency is much greater than the period of oscillation — which is just the condition $k \gg Ha/c_s$ again — we can apply the *principle of adiabatic invariance*, which tells us that the amplitude of the χ -field fluctuations should scale as $\chi \propto 1/\sqrt{\Omega}$ or as $\chi \propto a^{3/4}$. The auxiliary field fluctuations therefore grow in amplitude. However, the actual density fluctuation $\delta = \chi t^\alpha = \chi t^{-2/3} \propto \chi/a$, with the net result that the amplitude of acoustic fluctuations damp adiabatically as

$$\chi \propto a^{-1/4} \propto (1+z)^{1/4}. \quad (31.50)$$

Note that in obtaining this we assumed $a \propto t^{2/3}$, as is appropriate for a matter dominated cosmology. However, as the result was obtained using adiabatic invariance, it is independent of the details of the expansion, and would still apply if, for instance, there were some additional field present and affecting the expansion.

31.2.2 Radiation Dominated Era

The above Newtonian analysis is only valid for $P \ll \rho$, and for perturbations which are smaller than the horizon. When the former condition breaks down, one cannot use the non-relativistic Euler, energy equations, and when the latter is broken one cannot use Newtonian gravity. Here we will consider the evolution of acoustic perturbations on scales much less than the horizon, or equivalently

much less than the sound horizon $\lambda \ll c_s t$, since, in the radiation era, the sound horizon $c_s t = ct/\sqrt{3}$ tracks the light horizon, and we will see how the adiabatic damping of such waves is modified.

Now for waves with $\lambda \ll c_s t$ we can safely neglect the gravity due to the perturbation since the pressure gradient acceleration is overwhelming larger. What we shall do here is to compute the evolution of such waves in a fictitious background cosmology in which gravity is also neglected. We do not believe that this is really valid; it is *logically* possible to have a radiation dominated cosmology with $\Omega \ll 1$, but this appears not to be the case for our Universe. However, since the waves evolve adiabatically, the evolution of the wave amplitude, as a function of the scale factor, is independent of the details of the expansion law. By this ruse we are able to compute the evolution using only special relativity and we are able to sidestep the complications of coupling a relativistic plasma to gravity.

The equations of motion are then simply

$$T^{\mu\nu}{}_{,\nu} = 0. \quad (31.51)$$

When we considered self-interacting scalar elasticity sound waves we showed how these four equations could, in the ideal fluid limit, be cast into a more useful form as evolutionary equations for the energy density and for a 3-velocity \mathbf{v} . Here we will repeat that analysis, but using the conventional notation for cosmological density fields. The stress-energy tensor for an ideal fluid is

$$T^{\mu\nu} = (\rho + P/c^2)U^\mu U^\nu + \eta^{\mu\nu}P \quad (31.52)$$

where the 4-velocity $\vec{U} = (\gamma c, \gamma \mathbf{v})$ is that of observers who perceive a stress-energy tensor $T^{\mu\nu} = \text{diag}(c^2\rho, P, P, P)$; i.e. for observers comoving with the fluid.

If we set $\mu = i$ in (31.51) and make use of (31.51) with $\mu = 0$ we obtain the relativistic Euler equation

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{1}{\gamma^2(c^2\rho + P)} \left[c^2 \nabla P + \mathbf{v} \frac{\partial P}{\partial t} \right]. \quad (31.53)$$

Dotting (31.51) with \vec{U} gives

$$0 = U_\mu T^{\mu\nu}{}_{,\nu} = U_\mu \frac{\partial}{\partial x^\nu} [(\rho + P/c^2)U^\mu U^\nu] + U^\nu \frac{\partial P}{\partial x^\nu} \quad (31.54)$$

which, with $\vec{U} = (\gamma c, \gamma \mathbf{v})$ and using $\partial(\vec{U} \cdot \vec{U})/\partial t = 0$ gives the relativistic energy equation

$$\frac{\partial \rho}{\partial t} + (\mathbf{v} \cdot \nabla) \rho = -\frac{\rho + Pc^2}{\gamma} \left[\frac{\partial \gamma}{\partial t} + \nabla \cdot (\gamma \mathbf{v}) \right]. \quad (31.55)$$

Specializing to the case of a radiation density dominated plasma for which $P = \rho c^2/3$, (31.53) and (31.55) become

$$\frac{d\mathbf{v}}{dt} = -\frac{1}{\gamma^2 4\rho} \left[c^2 \nabla \rho + \mathbf{v} \frac{\partial \rho}{\partial t} \right] \quad (31.56)$$

$$\frac{d\rho}{dt} = -\frac{4}{3}\rho \left[\frac{1}{\gamma} \frac{d\gamma}{dt} + \nabla \cdot \mathbf{v} \right] \quad (31.57)$$

where $d/dt = \partial/\partial t + \mathbf{v} \cdot \nabla$ is the total, or convective, time derivative.

Equations (31.56), (31.57) admit exact solutions corresponding to a homogeneous expanding Universe with un-decelerated linear Hubble expansion $\mathbf{v} = H\mathbf{x} = \mathbf{x}/t$. Such solutions do *not* have $\rho(\mathbf{r}, t) = \bar{\rho}(t)$, as one might perhaps have expected. This is because in a homogeneous expanding Universe, the density is not constant on slices of constant coordinate time t , rather it is constant on surfaces of constant proper time since the big bang. For freely expanding matter in Minkowski space-time, the proper time is $\tau = \sqrt{t^2 - x^2/c^2} = t/\gamma$. Surfaces of constant τ are therefore hyperbolae in \mathbf{x}, t space. If we look for a solution with

$$\rho(\mathbf{x}, t) = \bar{\rho}(\tau) \quad (31.58)$$

we find that the factor in brackets on the right hand side of the Euler equation is

$$c^2 \nabla \rho + \mathbf{v} \frac{\partial \rho}{\partial t} = \frac{d\bar{\rho}}{d\tau} \frac{(\mathbf{v}t - \mathbf{x})}{\tau}, \quad (31.59)$$

but $\mathbf{x} = \mathbf{v}t$, so the right hand side of (31.56) vanishes. Therefore $d\mathbf{v}/dt = 0$, and so fluid elements do indeed move with constant velocities. If $d\mathbf{v}/dt = 0$ then $d\gamma/dt = 0$ also, and therefore the energy equation (31.57) becomes, in this context,

$$\frac{d\rho}{dt} = -\frac{4}{3}\rho \nabla \cdot \mathbf{v} = -4\frac{\rho}{t} \quad (31.60)$$

where we have used $\nabla \cdot \mathbf{v} = (1/t)\nabla \cdot \mathbf{x} = 3/t$, or, since $t = \gamma\tau$, and γ is constant along any fluid element world-line

$$\frac{d\rho}{d\tau} = -4\frac{\rho}{\tau} \quad (31.61)$$

which does indeed allow a solution with $\rho = \bar{\rho}(\tau)$, specifically $\bar{\rho} \propto 1/\tau^4$. Since this model is freely expanding (i.e. the scale-factor a is proportional to τ) this corresponds to the usual $\rho \propto 1/a^4$ behavior for relativistic matter.

Having established the exact homogeneous expanding solution, we now want to look at the evolution of perturbations about this ‘background’. In the above analysis we worked in Minkowski space coordinates (\mathbf{x}, ct) , but this proves cumbersome when we add perturbations. Instead we will consider the density and velocity to be functions not of \mathbf{x}, t but of the dimensionless comoving spatial coordinate $\mathbf{r} = \mathbf{x}/ct$ and the proper time τ . We can readily infer the equations governing the evolution of $\rho(\mathbf{r}, \tau)$ and $\mathbf{v}(\mathbf{r}, \tau)$ from equations (31.56), (31.57). Consider a point in space-time where the velocity of the fluid vanishes. At that point, intervals of coordinate time t and proper time τ are identical, $\gamma = 1$, and so these equations become

$$\frac{d\mathbf{v}}{d\tau} = -\frac{c^2}{4\rho} \nabla \rho \quad (31.62)$$

$$\frac{d\rho}{d\tau} = -\frac{4}{3}\rho \nabla \cdot \mathbf{v} \quad (31.63)$$

(note that vanishing of \mathbf{v} does not imply that $\nabla \cdot \mathbf{v} = 0$). But we can always make a Lorentz transformation to make the velocity at any point vanish, and therefore equations (31.62), (31.63) apply everywhere, with understanding that ∇ denotes the gradient with respect to physical displacement in the rest-frame of the fluid — i.e. it is the spatial gradient on surfaces of constant proper time τ .

It is easy to check that the zeroth order solutions

$$\rho = \rho_0 \propto 1/t^4 \quad (31.64)$$

$$\mathbf{v} = \mathbf{v}_0 = \mathbf{x}/t = c\mathbf{r} \quad (31.65)$$

satisfy these equations. Let us now look for solutions $\rho = \rho_0 + \rho_1$ and $\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_1$. Substituting these in (31.62), (31.63) and discarding all but the first order terms yields

$$\dot{\mathbf{v}}_1 = -\frac{c}{4\rho_0\tau} \nabla_r \rho_1 \quad (31.66)$$

$$\dot{\rho}_1 = -\frac{4}{3} \frac{\rho_0}{c\tau} \nabla_r \cdot \mathbf{v}_1 - \frac{4}{3} \frac{\rho_1}{c\tau} \nabla_r \cdot \mathbf{v}_0 \quad (31.67)$$

where $\nabla_r = a\nabla$ denotes spatial derivative with respect to comoving coordinate \mathbf{r} and dot denotes derivative with respect to τ . Now $\nabla \cdot \mathbf{v}_0 = 3c$ so (31.67) can be written as

$$\dot{\rho}_1 + 4\frac{\rho_1}{\tau} = -\frac{4}{3} \frac{\rho_0}{c\tau} \nabla_r \cdot \mathbf{v}_1. \quad (31.68)$$

Taking the derivative of this with respect to proper time gives

$$\ddot{\rho}_1 = \frac{d(\rho_0/\tau)/d\tau}{\rho_0/\tau} \left(-\frac{4}{3} \frac{\rho_0}{c\tau} \nabla_r \cdot \mathbf{v}_1 \right) - \frac{4}{3} \frac{\rho_0}{c\tau} \nabla_r \cdot \dot{\mathbf{v}}_1 - 4\frac{\dot{\rho}_1}{\tau} + 4\frac{\rho_1}{\tau^2}. \quad (31.69)$$

Now $\rho_0 \propto 1/\tau^4$ so $d(\rho_0/\tau)/d\tau = -5(\rho_0/\tau)$, and using equation (31.68) to eliminate $\nabla_r \cdot \mathbf{v}_1$ and (31.66) for $\dot{\mathbf{v}}_1$ yields a second order equation for ρ_1 :

$$\ddot{\rho}_1 + 9\dot{\rho}_1/\tau + 16\rho_1/\tau^2 - \frac{1}{3\tau^2}\nabla_r^2\rho_1 = 0. \quad (31.70)$$

However, for the waves of interest here, for which $\lambda \ll c\tau$, the last term is much larger in magnitude than the penultimate term and we therefore have, in this limit,

$$\ddot{\rho}_1 + 9\dot{\rho}_1/\tau - \frac{1}{3\tau^2}\nabla_r^2\rho_1 = 0. \quad (31.71)$$

We can now treat this exactly as we did for waves in the matter dominated era: letting $\rho_1 = \chi t^\alpha$, but now with $\alpha = -9/2$, eliminates the damping term and results in the oscillator equation

$$\ddot{\chi} - \frac{1}{3\tau^2}\nabla_r^2\chi = 0 \quad (31.72)$$

with frequency $\Omega = 1/\sqrt{3}\tau$, so adiabatic invariance tells us that the χ -field fluctuations evolve as $\chi \propto 1/\sqrt{\Omega} \propto \tau^{1/2}$ and therefore the amplitude of the density perturbation evolves as $\rho_1 \propto 1/\tau^4$. Thus, in this model, the density perturbation amplitude evolves exactly as the background density and the fractional density perturbation amplitude is constant:

$$\delta = \frac{\rho_1}{\rho_0} \propto \tau^0. \quad (31.73)$$

31.2.3 Super-Horizon Scale Perturbations

The rigorous treatment of perturbations with $\lambda \gtrsim ct$ is quite involved, and there are some subtleties involved in defining the perturbation amplitude. In the Newtonian analysis we define the perturbation to be the variation in the density at a given time. For super-horizon scale waves one needs to carefully define the hyper-surface on which one defines the density contrast. One could, for example, take these hyper-surfaces to be surfaces of constant density, in which case the density perturbation would vanish identically. This does not mean that super-horizon scale perturbations are ill-defined, since in this case the expansion law would not be uniform. The results of the more detailed analysis, which we shall not cover here, are in accord with the picture which emerges from the spherical model that an over-dense region can be thought of as part of a different universe (i.e. one with a smaller radius of curvature) evolving independently. These perturbations have constant spatial curvature or gravitational potential, provided $\lambda \gg c_s t$. In the matter dominated case we had $\delta\phi = \text{constant} \sim \delta M/R \sim (R^3\delta\rho)/R \sim (\delta\rho/\rho)\rho R^2$. Constancy of $\delta\phi$ and $\rho \propto 1/R^3$ then implies $\delta\rho/\rho \propto R \propto a$. In the radiation dominated case, the same hand-waving argument, but now with $\rho \propto 1/R^4$, would suggest that $\delta\rho/\rho \propto a^2$, which is in agreement with the more sophisticated analysis that can be found in Peebles' book for instance.

31.2.4 Isocurvature vs Isentropic Perturbations

The type of perturbation we have described above is an *adiabatic perturbation*. This is the perturbation one makes if one takes the matter and radiation in some region and compresses it; the ratio of photons to baryons is fixed and consequently $\delta_{\text{rad}} = (4/3)\delta_{\text{b}}$ and there is a non-zero perturbation in the total matter density

$$\delta_{\text{tot}} = \left(\frac{\delta\rho}{\rho}\right)_{\text{tot}} = \frac{\bar{\rho}_{\text{rad}}\delta_{\text{rad}} + \bar{\rho}_{\text{b}}\delta_{\text{b}}}{\bar{\rho}_{\text{rad}} + \bar{\rho}_{\text{b}}}. \quad (31.74)$$

as illustrated in the left hand panel of figure 31.2.

An alternative that used to be considered is a *isothermal perturbation* where the radiation is unperturbed initially. Nowadays we think of multi-component fluids comprising e.g. pressure free cold dark matter, plus a tightly coupled plasma composed of baryons and radiation. A similar ambiguity then arises as how to set up the initial perturbation. One very natural type of perturbation is to

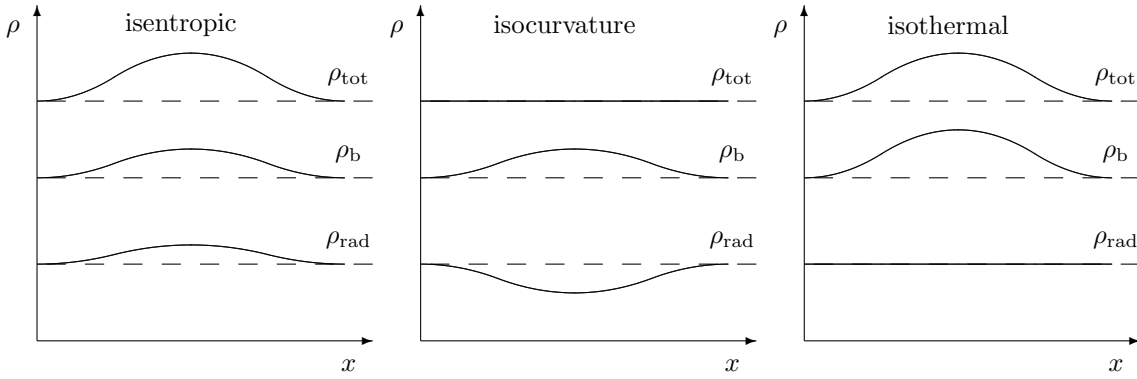


Figure 31.2: The left hand panel shows an ‘adiabatic’ or ‘isentropic’ perturbation of the kind we have been discussing above. In such a perturbation we crush the matter and radiation together. Such perturbations have a net density inhomogeneity and consequently have non-zero curvature or potential perturbations. A very natural alternative is to generate a perturbation in which the initial perturbation in the baryon density is cancelled by a corresponding under-density in the radiation. Such perturbations have, initially, no net density perturbation and therefore no associated curvature perturbation, and are called ‘isocurvature’. For super-horizon scale perturbations the curvature is frozen in, but there is a non-zero pressure gradient, and once the perturbations enter the horizon this becomes effective and will act to annul the pressure gradient. In the example shown in the center panel, there is an inward directed pressure gradient which will act to erase the under-density in radiation, but in doing so will enhance the over-density in the baryons. The radiation density will over-shoot and one will have an oscillation about a state in which the radiation is uniform. The equilibrium state about which these oscillations will occur is shown in the right panel and is known as a ‘isothermal’ perturbation, since the radiation density, and therefore also the temperature, are constant.

say that initially there were no curvature fluctuations, and the perturbation is produced by varying the relative fractions of baryons to photons with the excess in any one fluid being compensated by a deficit in the other. Such a perturbation is termed an *isocurvature perturbation* and is the descendant of the old style isothermal perturbation. The alternative is the so-called *isentropic perturbation* in which the fluids are compressed together and which is the descendant of the old-style adiabatic perturbation. The terminology here arises because the number of photons per baryon is equivalently the entropy per baryon.

31.2.5 Diffusive Damping and Free-Streaming

The Jeans analysis assumed a tightly coupled baryon-photon plasma. There are two situations where this is inappropriate.

At early times the plasma is very optically thick, but as the universe expands the mean-free path for photons increases and the photons tend to leak out of the sound waves and consequently they damp out. This was first analyzed by Joe Silk and is called *Silk damping*. Around $z = 1000$, the ions and electrons (re)combine and become neutral and the photons are no longer locked to the baryons. The consequence of this is that sufficiently small scale perturbations (interestingly close to the mass of galaxies as it happens) are damped out. Nowadays the emphasis is on dark matter models where the damping of acoustic fluctuations is not so critical.

Another interesting (and possibly very relevant) situation is that of *hot-dark-matter* such as a neutrino with a mass on the order of 10 eV. Such particles have large thermal velocities, but are not locked to the other components by scattering. One can still define a *neutrino Jeans length* which is the distance a neutrino can travel in one expansion time. This starts off small (in comoving terms), grows and peaks around the time the neutrinos go non-relativistic. Once the neutrinos become non-relativistic their velocities decay adiabatically as $v \propto 1/a$ according to equation (31.27) and

the comoving Jeans length decreases. As with acoustic waves, perturbations bigger than the Jeans length grow, but smaller perturbations damp out very rapidly as the neutrinos stream out of the perturbation.

31.3 Scenarios

We will now put these results to work and explore a number of scenarios for structure formation.

31.3.1 The Adiabatic-Baryonic Model

The first scenario to be worked out in any detail was the *adiabatic baryonic* model (Peebles and Yu). In this calculation one assumes some initial adiabatic density fluctuations imposed at some very early epoch, and one assumes that the universe is dominated by baryons and radiation. One Fourier decomposes these into plane wave modes and then calculates the temporal evolution of each mode separately. At sufficiently early times all the modes have $\lambda \sim a/k \gg t$, pressure gradients are negligible and the perturbations grow in step with $\delta \propto a^2$. Eventually, short wavelength perturbations enter the horizon. For galaxy scale perturbations ($\lambda_0 \sim 1$ Mpc) this happens when the universe is radiation dominated. These perturbations then oscillate as acoustic waves with constant amplitude. As time goes on progressively longer wavelengths enter the horizon and start oscillating. Eventually we reach the epoch z_{eq} when $\rho_{\text{rad}} = \rho_{\text{matter}}$ and then something rather interesting happens. Before z_{eq} , $P \simeq \rho/3$, $c_s \simeq 1/\sqrt{3}$ and the ‘sound horizon’ $\simeq c_s t/a$ grows like $a \propto t^{1/2}$. After z_{eq} , $P/\rho \propto 1/a$, the sound speed falls like $1/\sqrt{a}$, $t \propto a^{3/2}$ and consequently the comoving sound horizon (or comoving Jeans length) is approximately constant. At $z \simeq 1000 \simeq z_{\text{eq}}/10$, the plasma recombines and the pressure support from the radiation is lost, the Jeans length falls, and all perturbations grow according to the usual $\delta \propto a$ law. The various phases of the life of an adiabatic perturbation are shown schematically in figure 31.3.

The result of this calculation can be expressed in terms of a *transfer function* $T(k)$ expressing the final amplitude of a perturbation of (comoving) wavenumber k relative to its initial amplitude. For a power-law initial power spectrum $P(k) \propto k^n$ for instance, the final power spectrum is $P(k) \propto k^n T^2(k)$. This linear output spectrum can then be given as initial conditions to a numerical simulator who can evolve the density fluctuations into the non-linear regime and compare the results with e.g. galaxy clustering data. The transfer function for the adiabatic baryonic model has some rather interesting features. For wavelengths longer than the maximum Jeans length (roughly the horizon size at z_{eq}) $T(k)$ is constant, but for shorter wavelengths there are oscillations. These arise because one assumes that the perturbations are purely in the growing mode at t_i , so the phase of the sound wave $\varphi = \int \omega(t) dt$ at z_{dec} is smooth and monotonically increasing function of k . If one models the decoupling process as instantaneous, it is easy to show that the amplitude of the growing mode perturbation is then an oscillatory function of k . This results in nodes and anti-nodes (see figure 31.4) and therefore quasi-periodic bumps in the power spectrum. These oscillations extend to about a factor 10 in wavelength, below which they are damped out by photon diffusion. There is a large bump in $T(k)$ at around the maximum Jeans length. This comes about because of the plateau in the Jeans length; a mode with wavelength just greater than λ_{Jmax} will grow uninterrupted while one of slightly shorter wavelength will oscillate with actually slowly decreasing amplitude between z_{eq} and z_{rec} .

This calculation really marked a turning point in structure formation. While the creation of the initial fluctuations was still a subject of speculation, it seemed reasonable to assume a power law initial state, and one then obtained a quantitative prediction for the post- z_{eq} power spectrum. This model has only two free parameters; the initial amplitude of the power spectrum and the initial spectral index n . Moreover, this spectrum had two prominent features; the bump at w_{Heq} and the damping cut-off. The present horizon size is $\simeq 1/H_0 \simeq 3000$ Mpc, and since $w_{\text{H}} \propto t/a \propto a^{1/2} \propto z^{-1/2}$, $a_0 w_{\text{Heq}} \simeq 30$ Mpc which is the scale of superclusters, the largest prominent structures we observe in the universe; a remarkable fact indeed! The damping scale sets a minimum scale for the first structures to form, and it is obviously attractive to identify this scale with galaxies in this theory.

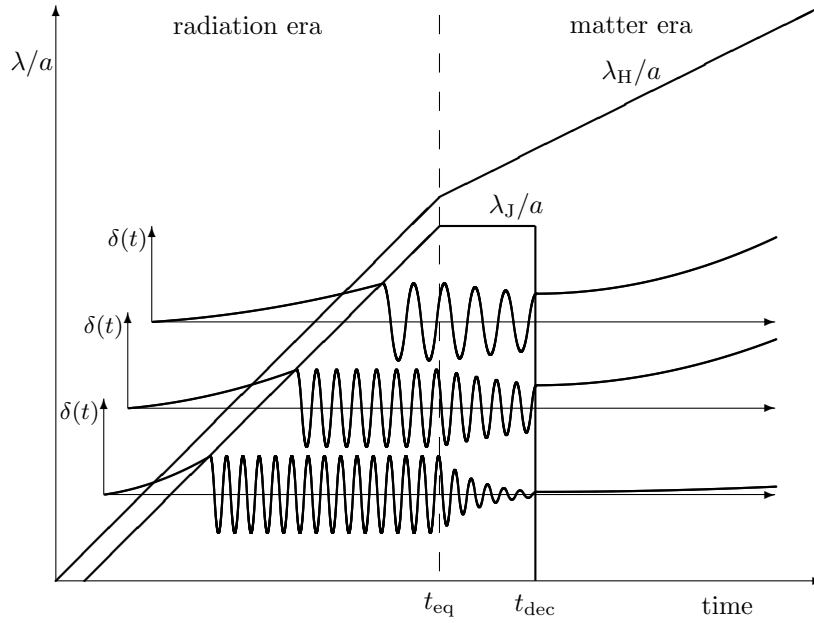


Figure 31.3: Evolution of initially adiabatic (or isentropic) perturbations is shown schematically for perturbations of three different wavelengths. The perturbation passes through three phases. First, when outside the horizon, the perturbation amplitude grows as $\delta \propto a^2$ in the radiation era and as $\delta \propto a$ in the matter dominated era. Perturbations which enter the horizon before t_{eq} oscillate at constant amplitude until t_{eq} . For $t_{\text{eq}} < t < t_{\text{dec}}$ the amplitude decays adiabatically as $\delta \propto 1/a^{1/4}$. Short wavelength perturbations are, in addition, subject to diffusive damping, and are strongly attenuated. Perturbations which persist to t_{dec} then couple to growing and decaying perturbations in the now pressure-free neutral gas.

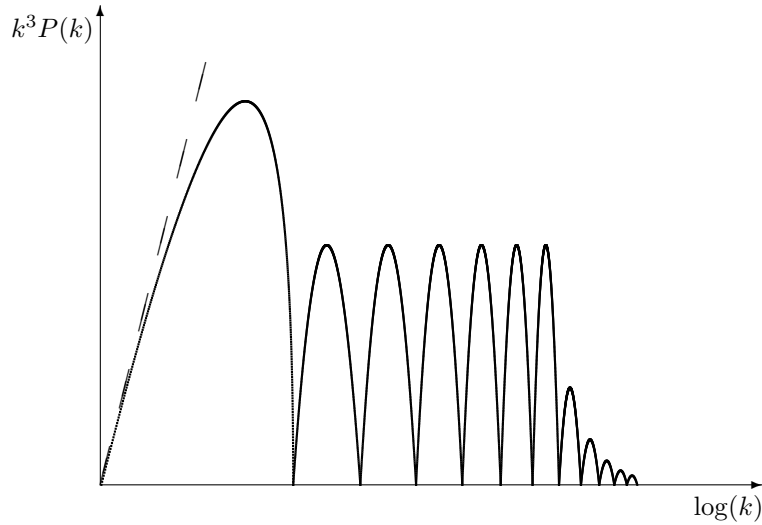


Figure 31.4: Power spectrum in the adiabatic-baryonic model (schematic). The dashed line indicates the initial power spectrum. The main peak is at a scale just larger than the maximum Jeans length, where the perturbations underwent continuous growth. Shorter waves entered the horizon before z_{eq} and subsequently oscillated, so their amplitude is suppressed. The nodes in the output spectrum are those wavelengths which have zero amplitude in the growing mode at the time of decoupling. The cut-off in the power spectrum at high k is due to diffusive damping.

This theory therefore has much to commend it. Its main weakness is that it does not include dark matter. Dark matter seems to dominate over baryonic matter in clusters of galaxies — least-ways it is hard to see how the dark matter there could be in a baryonic form — and big-bang nucleosynthesis predicts an uncomfortably low baryonic density parameter: $\Omega_b \simeq 0.02/h^2 \simeq 0.04$. Such a low density means that perturbations freeze out quite early (when Ω starts to peel away from unity) and this makes it hard to account for both the smallness of the microwave background fluctuations and the presence of galaxies etc. at $z \gtrsim 3$.

As well as the empirical evidence for substantial quantities of dark matter, which implied a density parameter $\Omega \simeq 0.2$, and possibly more if galaxies are biased with respect to dark matter, there was the natural repugnance on the part of theorists to the idea that Ω be quite close to unity, but not exactly so. These facts and prejudices led to the development of models containing large amounts of dark-matter.

31.3.2 The Hot-Dark-Matter Model

The first non-baryonic model to be seriously studied was the *hot dark matter model* (HDM), with the candidate particle being a massive neutrino. Neutrinos are produced in thermal abundance at high temperatures, but the weak interactions freeze-out around an MeV or so and the neutrino are thereafter effectively uncoupled from the rest of matter save through their gravitational influence. The expected number density of a light neutrino species is then roughly the same as the number density of photons in the microwave ground; this is only a rough equality because the number of photons was boosted somewhat when the electrons and positrons recombined, and from consideration of degrees of freedom. Thus if the neutrinos have a mass such that they would go non-relativistic at around z_{eq} they would have a density comparable to the critical density today. This requires $m \simeq 30$ eV, and this theory got a substantial boost when there was experimental evidence to suggest a mass of this order (Lyubimov).

In the HDM model, fluctuations on small scales are damped out by free-streaming. The characteristic smoothing length in this theory is just the comoving distance traveled by a neutrino. At early times, $z > z_{\text{eq}}$, the neutrinos are relativistic and the comoving distance traveled is just the horizon size w_H , after they go non-relativistic the velocity of the neutrinos redshifts $v \propto 1/a$, so the comoving distance traveled per expansion time is $\propto vt/a \propto t/a^2 \propto a^{-1/2}$ which decreases with time. Thus, the total distance traveled is on the order of the horizon size at z_{eq} , which, as we have seen, is roughly the scale of superclusters today.

As there is essentially no power remaining on small scales, the first structures to go non-linear are superclusters and galaxies must then form by fragmentation; this is called a *top down scenario*, as opposed to a *bottom up scenario* in which structures form first on small scales and then cluster hierarchically into progressively larger entities.

The formation of structure in HDM-like models was analyzed extensively by Zel'dovich and co-workers and can be understood analytically in terms of his beautiful approximation in which structures form by ‘pancaking’. While the gross features of large-scale structure predicted in the HDM model agree nicely with the impression of filamentary or sheet-like structure seen in galaxy surveys, it seems hard to reconcile the early appearance of quasars and radio galaxies and the absence of neutral gas at high redshift — the Gunn-Peterson test — with the fact that the large scale structure appears to be forming today.

31.3.3 The Cold Dark Matter Model

The *cold dark matter model* (CDM) also assumes that the bulk of the matter is non-baryonic, but that this material is not in thermal equilibrium at early times. One possible candidate for CDM is the axion (see Kolb and Turner) in which the density resides in coherent oscillations of a scalar field. For the present purposes all we need to know about the axion (or whatever) is that it behaves just like zero-pressure dust.

If we track the evolution of say a galactic scale perturbation then at early times $\lambda \gg t$, pressure gradients are negligible, and the perturbation grows in the usual energy/curvature conserving

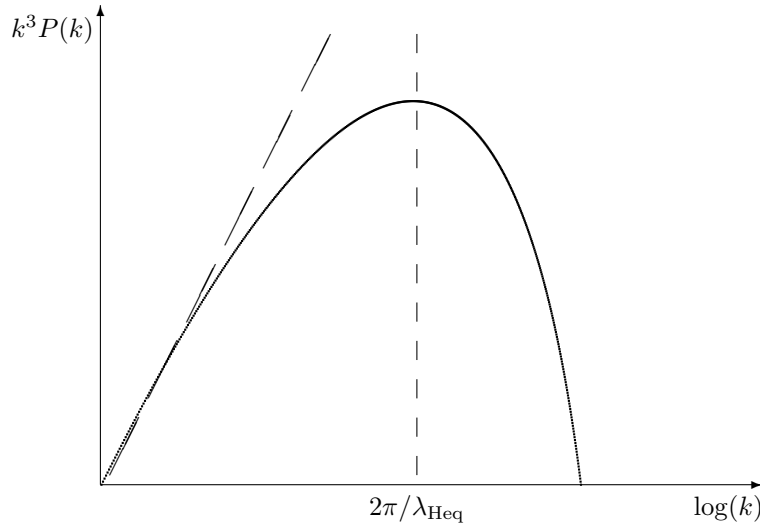


Figure 31.5: Power spectrum in the hot dark matter (HDM) model (schematic). The long-dashed line indicates the initial power spectrum. The vertical dashed line indicates the horizon size at the time the neutrinos become non-relativistic at $z \simeq z_{\text{eq}}$. In the HDM model the first structures to form are super-cluster scale, and smaller scale-structures must form by fragmentation.

manner. The perturbation enters the horizon in the radiation dominated era; the baryon-photon plasma then oscillates acoustically and the growth of the sub-dominant CDM component stagnates. The fluctuations are not erased however, as they are in the HDM model, so at z_{eq} the fluctuations can start to grow. On scales exceeding the horizon size at z_{eq} there is uninterrupted growth. The end result is a suppression of small scale fluctuations by a factor $\sim (\lambda/\lambda_{\text{Heq}})^{-2}$ in amplitude. As we shall see, the preferred initial spectrum has $\delta \propto 1/\lambda^2$ initially, so on small scales the amplitude of the fluctuations becomes asymptotically constant (see figure 31.6). The details have been worked out numerically (Bond and Efstathiou, 1984; Vittorio and Silk, 1984) and it turns out that the progression from the large-scale to small scale asymptotes is very gradual, and consequently one has in effect a hierarchical or ‘bottom-up’ scenario.

The CDM model has been explored in much greater detail than any of the other scenarios, and in many ways makes predictions which agree very well with observations; certainly the qualitative predictions fit very nicely with what is seen. In recent years, however, the simplest, and most attractive, version of the theory — in which the total density parameter is unity — has come under attack from observations of large-scale clustering; there is more power on large scales than the theory predicts. This will be discussed in greater depth below.

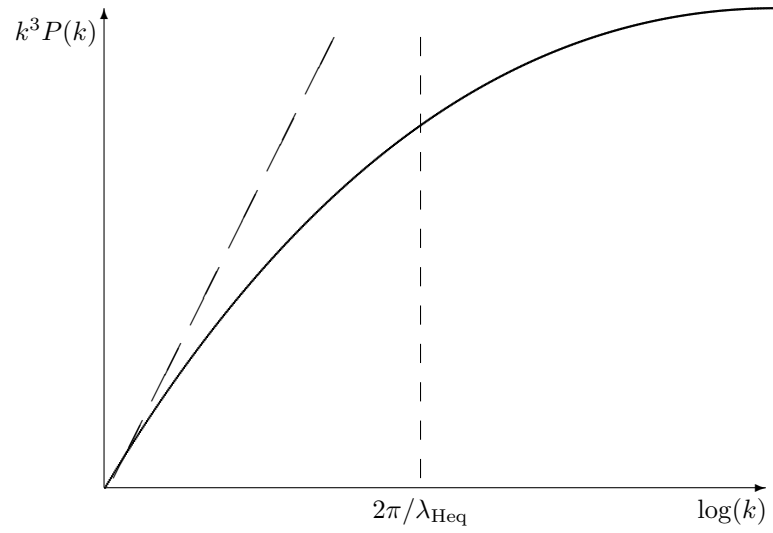


Figure 31.6: Power spectrum in the cold dark matter (CDM) model (schematic). The long-dashed line indicates the initial power spectrum. The vertical dashed line indicates the horizon size at the time the neutrinos become non-relativistic at $z \simeq z_{\text{eq}}$.

Chapter 32

Origin of Cosmological Structure

Having considered the evolution of density perturbations from some given initial state we now explore how the initial seeds for structure may have arisen. We first consider the ‘spontaneous’ generation of fluctuations from the effect of non-gravitational forces in the hot big bang model, and show that it is very difficult to generate large-scale structure in this way. We then consider the generation of density fluctuations from quantum fluctuations in the scalar field during inflation and finally we consider topological defects.

Before embarking on these calculations, it is worth describing what seems to be required observationally. In fact, long before inflation and when cosmological structure formation was still a relatively immature subject, Harrison and Zel’dovich pointed out that if the initial spectrum of fluctuations had a power-law spectrum $P(k) \propto k^n$ extending over a very wide range of scales then it should have index $n = 1$. The argument is that for a power-law, the fluctuations in the potential, and therefore in the curvature, also have a power law spectrum. For most spectral indices, the curvature fluctuations will either diverge at small scales or at large scales. This would result in small black-holes if n is too large, or would lead to the universe being highly inhomogeneous on large scales if n is too small. The ‘happy medium’ (in which the curvature fluctuations diverge at *both* small and large scales, but only logarithmically fast) is that for which the root mean square density fluctuations scale as $\delta\rho/\rho \propto 1/\lambda^2$, so the potential fluctuations $\delta\phi \sim (H\lambda)^2\delta\rho/\rho$ are independent of λ . For a power-law power spectrum, the variance per octave of wave-number is $\langle(\delta\rho/\rho)^2\rangle_k \sim k^3 P(k) \propto k^{3+n} \propto \lambda^{-(3+n)}$. Thus, for $n = 1$, the potential fluctuations are scale invariant. This is known as the *Harrison-Zel’dovich spectrum*. While somewhat philosophically motivated, this kind of spectral index has much to commend it. Gott and Rees, had argued that the structure we see on scales of galaxies, clusters and super-clusters seemed to require an spectral index for the perturbations emerging after z_{eq} of $n \sim -1$. This is not the Harrison-Zel’dovich index, but allowing for the suppression of the growth of small scale perturbations during the physical processes described above during the era around z_{eq} , these are consistent. The real clincher for the $n = 1$ spectrum came with the detection by COBE of roughly scale invariant ripples in the large-angle anisotropy of the CMB. Normalizing the spectrum to cluster or super-cluster scale structures, these fit very nicely to an extrapolation to larger scales using the Harrison-Zel’dovich spectrum.

32.1 Spontaneous Generation of Fluctuations

Consider an initially homogeneous universe and let the pressure become inhomogeneous (this might happen during a phase transition in the early universe, or at much later times when stars form and explode). This non-gravitational force will generate density perturbations. For a spontaneous cosmological phase transition the pressure fluctuations should be uncorrelated on scales larger than the horizon scale at that time. Similarly, a natural model for the pressure perturbation from randomly exploding stars has a flat power spectrum on large scales. What is the amplitude of mass fluctuations on large scales generated by such a process? The answer is very little. Naively, one might imagine that there might be root- N perturbations, with N the number of independent fluctuation regions,

giving $\delta\rho/\rho \propto M^{-1/2}$. Alternatively, one might imagine there would be ‘surface fluctuations’ giving $\delta\rho/\rho \propto M^{-5/6}$. Now it is true that if we measure the density within a sharp-edged top-hat sphere, then there will be fluctuations in the mass of this later order, but the fluctuations in growing modes will be much smaller than this; the amplitude of the growing mode is in fact $\delta\rho/\rho \propto M^{-7/6}$.

Let’s first obtain this result from a Newtonian analysis. What we shall do is compute the perturbation to the large-scale gravitational potential $\delta\phi$ — since this is associated with the growing mode density perturbations — from which we can obtain $\delta\rho/\rho$. Consider first a homogeneous expanding dust-filled cosmology containing an agent who can re-arrange the surrounding matter, but can only influence material at distances $r < R$ (see figure 32.1). What is the perturbation to the Newtonian potential at large scales? The potential is

$$\delta\phi(\mathbf{r}) = -G \int d^3r' \frac{\delta\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} \quad (32.1)$$

where the integrand vanishes for $r' \gtrsim R$. At large distances $r \gg R$, we can expand the factor $1/|\mathbf{r}' - \mathbf{r}|$ as

$$\frac{1}{|\mathbf{r}' - \mathbf{r}|} = (\mathbf{r} \cdot \mathbf{r} - (2\mathbf{r}' \cdot \mathbf{r} - \mathbf{r}' \cdot \mathbf{r}'))^{-1/2} = \frac{1}{r} \left(1 - \frac{1}{2} \frac{2\mathbf{r}' \cdot \mathbf{r} - \mathbf{r}' \cdot \mathbf{r}'}{r^2} \right)^{-1/2} \quad (32.2)$$

Making a Taylor expansion gives

$$\frac{1}{|\mathbf{r}' - \mathbf{r}|} = \frac{1}{r} + \frac{\mathbf{r} \cdot \mathbf{r}'}{r^2} + \frac{1}{2} \left(\frac{\mathbf{r}' \cdot \mathbf{r}'}{r^2} - 3 \frac{(\mathbf{r}' \cdot \mathbf{r})^2}{r^4} \right) + \dots \quad (32.3)$$

Using this in (32.1) gives an expansion in powers of $1/r$. The coefficient of the leading order term (for which $\delta\phi \sim 1/r$) is $\int d^3r' \delta\rho(\mathbf{r}')$. This is the *monopole moment* of the mass distribution, but this vanishes by virtue of conservation of mass. The next term has $\delta\phi \propto 1/r^2$, and has coefficient proportional to $\hat{\mathbf{r}} \cdot \int d^3r' \delta\rho(\mathbf{r}')\mathbf{r}'$, which is the *dipole moment*. This vanishes by virtue of momentum conservation. The next term has $\delta\phi \propto 1/r^3$ with coefficient proportional to the *quadrupole moment*. This does not, in general vanish; the agent, can, for example, rearrange the matter into a ‘dumb-bell’ shaped configuration without exchanging any mass or momentum with the exterior (see figure 32.1). If the mass contained within the perturbation region is ΔM , the large-scale gravitational potential is $\delta\phi(r) \sim G\Delta MR^2/r^3$ where R is the scale of the fluctuation region and this is smaller than the un-shielded monopole term $G\Delta M/r$ by two powers of R/r .

Now consider a multitude of such agents, with separation $\sim R$, each of whom re-arranges the surrounding matter in accordance with mass and momentum conservation, but otherwise in a random manner, such that different fluctuation regions are uncorrelated with each other (see figure 32.2). The mean square large scale potential — averaged over a region containing mass M , or size $r \sim (M/\rho)^{1/3}$ — is then the sum of $N \sim (r/R)^3 \sim M/\Delta M$ quadrupole sources adding in quadrature, so the root mean square potential perturbation is

$$\delta\phi_M = \langle (\delta\phi)^2 \rangle_M^{1/2} \sim N^{1/2} \frac{G\Delta MR^2}{r^3} \propto M^{-1/2}. \quad (32.4)$$

The fluctuations in the potential are therefore a *white-noise* process. Now, for the growing mode, the potential and density fluctuations are related by $\delta\phi \sim (Hr/c)^2 \delta\rho/\rho \propto M^{2/3} \delta\rho/\rho$, so the root mean squared growing mode density perturbations induced by this kind of small-scale ‘curdling’ has rms $\delta\rho/\rho \propto M^{-7/6}$. The mass distribution on large-scales at late times is much smoother than ‘root- N ’ mass fluctuations, and smoother even than the ‘surface fluctuations’.

The argument given above is Newtonian and assumes conservation of mass. Do these conclusions still hold with fluctuations of the relativistic plasma? For example, consider a universe in which the process of baryogenesis is spatially inhomogeneous. If the photon-to-baryon ration — the specific entropy that is — is an incoherent random function of position, this will generate an initially isocurvature perturbation such that the number density of baryons is a white-noise process, but with the initial density fluctuation in the baryons being compensated by the radiation density.

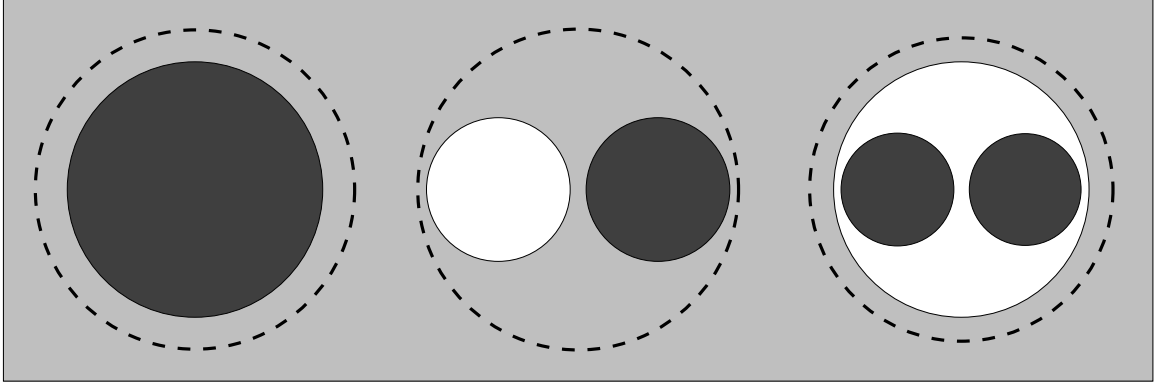


Figure 32.1: Illustration of the type of perturbation that can be generated by a physical process that operates locally (within region delimited by the dashed circle). On the left is a monopole perturbation. This has a net excess mass and would generate a potential perturbation at large scales $\delta\phi \propto 1/r$. Such a perturbation is not allowed, since it requires importing mass from large distances; if one is constrained only to re-arrange the mass within the dashed circle, then for a symmetric mass configuration the net mass excess must vanish. In the center is shown a dipole perturbation with an over-dense region on the right and an under-dense region on the left. Such a perturbation would generate a large-scale gravitational potential $\delta\phi \propto 1/r^2$. The net mass excess inside the dashed circle is now zero, but such perturbations are still now allowed as, in order to generate such a perturbation, one would need to impart a net momentum to the matter. On the right is a quadrupole perturbation. Such a perturbation can be generated by a local physical process while still conserving mass and momentum. A quadrupole source generates a large-scale potential perturbation $\delta\phi \propto 1/r^3$; this falls off much faster than for an ‘un-shielded’ monopole perturbation.

Now as the universe expands the radiation will redshift away and will eventually become negligible. At late times then there will be fluctuations in the net proper mass contained within any comoving region with rms amplitude scaling inversely as the square root of the number of fluctuation regions, or $\delta\rho/\rho \propto M^{-1/2}$. Does this not conflict with the $M^{-7/6}$ rule? Not necessarily, since we do not know what fraction of these perturbations is in the growing mode. To resolve this, recall the behavior of spherical perturbations. To generate a decaying mode we delay the ‘bang-time’ keeping the energy constant, and the proper mass contained in the perturbation is fixed. In the growing mode we perturb the binding energy ϕ . The *gravitational* mass of the perturbation must equal the unperturbed mass, but as the binding energy is negative, we must actually have a slight enhancement of the net proper mass $\delta M \sim -M\delta\phi$ within the perturbation. Thus, the fluctuations in proper mass within comoving regions (which scale as $M^{-1/2}$ in this incoherent isocurvature model) measure $\delta\phi \sim (H^2\lambda^2)\delta_{\text{growing}}$ and we recover the $\delta\rho/\rho \propto M^{-7/6}$ behavior for the growing modes.

The large-scale growing perturbations produced by small-scale rearrangement of mass are therefore very small and this effectively excludes the possibility that the large-scale structure results from curdling of the universe during a phase transition at early times because the horizon size is small then. It would also seem quite difficult to produce the largest scale structures seen from hydrodynamical effect of supernovae explosions (though the fact that a simple estimate of the net energy released based on the abundance of the results of nuclear burning in stars does not fall very far short of what is desired is tantalizing). In any such scenario, accounting for the large-angle fluctuations in the CMB is very difficult indeed, since the prediction is for temperature fluctuations falling off as $\delta T/T \sim \delta\phi \propto \theta^{-3/2}$.

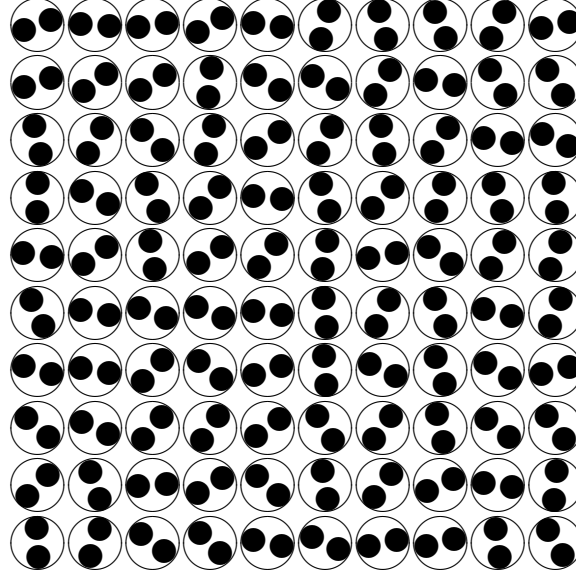


Figure 32.2: Schematic illustration of the type of density inhomogeneity than can be produced by local re-arrangement of the matter. This is the type of perturbations generated by e.g. phase transitions in the hot-big bang model, where the effects of pressure gradients are limited by the horizon size. Each fluctuation region generates a large-scale potential fluctuation $\delta\phi \sim G\Delta MR^2/r^3$, where ΔM and R are respectively the mass within and the size of a fluctuation region. The orientation, and in a realistic model also the amplitude, of the potential fluctuation is a random variable. Therefore the large-scale potential fluctuations are \sqrt{N} times larger than the effect for a single region. With $N \propto M$, the root mean squared potential fluctuation is $\delta\phi \propto \sqrt{M}/r^3 \propto 1/\sqrt{M}$. The potential generated by such a process is therefore a ‘white-noise’ process, with a flat power spectrum $P_\phi(k) = \text{constant}$ (i.e spectral index $n = 0$). The mass fluctuations $\delta\rho$ in the growing mode are related to the potential fluctuations by $\nabla^2\delta\phi = 4\pi G\delta\rho$, so $\delta\rho_{\mathbf{k}} \sim G^{-1}k^2\delta\phi_{\mathbf{k}}$ and the power-spectrum of the mass-fluctuations is therefore $P_\rho(k) \sim \langle |\delta\rho_{\mathbf{k}}|^2 \rangle \propto k^4$. The spectral index is $n = 4$, and the root mean squared mass fluctuations at late times are $\delta\rho/\rho \propto M^{-7/6}$.

32.2 Fluctuations from Inflation

A more promising way to generate density fluctuations is from quantum fluctuations in the scalar field driving inflation. In chapter 29 explored a model of ‘chaotic inflation’ in which the inflaton field has potential $V(\phi) = \lambda\phi^4/\hbar c$. We found that provided $\phi \gg \sqrt{c^4/G}$ — or $\phi \gg m_{\text{pl}}$ in natural units — the stress-energy tensor has $P = -\rho c^2$, as required to drive inflation. While it was not stated there, we also assumed implicitly that the field was weakly self-interacting $\lambda \ll 1$, and that the field value was $\phi \ll \lambda^{-1/4}\sqrt{c^4/G}$, so that the density $\rho \simeq \lambda\phi^4/c^2$ is much less than the Planck-density. With these assumptions, the important scales in the inflaton-radiation-matter dominated cosmology are as depicted in figure 29.1. It is a remarkable feature of the inflationary cosmology that, in addition to solving the flatness, horizon and possibly other problems, it naturally predicts that at late times, there will be density fluctuations re-entering the horizon with amplitude

$$\frac{\delta\rho}{\rho} \sim \sqrt{\frac{\hbar}{c}} \frac{H^2}{\dot{\phi}} \quad (32.5)$$

where these quantities are evaluated as the perturbations leave the horizon during the inflationary era. Since, H and $\dot{\phi}$ are slowly varying during inflation, this naturally predicts seeds for structure formation close to the preferred Harrison-Zel’dovich form.

As usual, since we are dealing with small amplitude fluctuations, the natural approach is to decompose the field into spatial Fourier modes, and compute the evolution of these separately. As

shown in figure 29.1, such a mode, being fixed in comoving wavelength, first appears, or rather becomes describable without a quantum theory of gravity, when the physical wavelength is on the order of the Planck length. As already discussed, for inflation to take place we require that the field fluctuation at that time be in the vacuum state to very high accuracy. This then sets the initial conditions; the initial occupation number for inflatons of this scale must vanish. A detailed calculation of the evolution is extremely technical, involving such tricky issues as the nature of the vacuum in curved space-time, as well as requiring a full general-relativistic treatment for the modes while they are outside the horizon. Here we shall only give a rather hand-waving sketch of the important processes and thereby physically justify the form of the key result (32.5). We will show that the requirement that the final density fluctuation amplitude agree with that required observationally puts very strong constraints on the strength of the interaction term (or mass term) in the inflaton potential. We will also discuss how inflation predicts, in addition to density fluctuations, fluctuations in all fields, and, in particular, predicts a stochastic background of gravity waves. This provides a potential test of the theory.

First, we need to establish the nature of the fluctuations about the large-scale average inflaton field during inflation. We will denote the ‘background’ field by ϕ_0 , and the fluctuations, which, as we shall see, are relatively small, by ϕ_1 . The general equation of motion for the inflaton field is

$$\ddot{\phi} + 3H\dot{\phi} + \frac{c^2}{a^2}\nabla^2\phi + 4\lambda\frac{c}{\hbar}\phi^3 = 0, \quad (32.6)$$

where ∇ denotes the derivative with respect to comoving coordinates. If we decompose the field as $\phi = \phi_0 + \phi_1$, where the ‘background’ field ϕ_0 is assumed to have $\nabla\phi_0 = 0$, and make a Taylor expansion of the interaction term assuming that the fluctuations about the background are relatively small (i.e. $\phi_1 \ll \phi_0$) then the equation of motion for the perturbation, which will not in general have small spatial gradient, is

$$\ddot{\phi}_1 + 3H\dot{\phi}_1 + \frac{c^2}{a^2}\nabla^2\phi_1 + 12\lambda\frac{\phi_0^2 c}{\hbar}\phi_1 = 0. \quad (32.7)$$

Compare this with the equation of motion for a free massive scalar field

$$\ddot{\phi} + 3H\dot{\phi} + \frac{c^2}{a^2}\nabla^2\phi + \frac{m^2 c^4}{\hbar^2}\phi = 0. \quad (32.8)$$

Evidently, the fluctuations about the background field behave like a free field with mass

$$m = \sqrt{12\lambda\hbar\phi_0^2/c^3}. \quad (32.9)$$

The Compton wavelength for the inflaton fluctuations is

$$\lambda_c = \hbar/mc \quad (32.10)$$

which we can compare to the horizon scale c/H . With $H \sim \sqrt{G\rho} = \sqrt{G\epsilon/c^2} \sim \sqrt{G\lambda\phi_0^4/\hbar c^3}$, the ratio of these is

$$\frac{\lambda_c}{c/H} \sim \sqrt{\frac{G\phi_0^2}{c^4}}. \quad (32.11)$$

Therefore, if the field is large enough to allow inflation ($\phi_0 \gg \sqrt{c^4/G}$) then $\lambda_c \gg c/H$; the Compton wavelength is much larger than the horizon. Thus, to a very good approximation, the classical equation governing fluctuations ϕ_1 is that of a free, massless field. This result is not specific to the $V \propto \phi^4$ form for the inflaton potential; the same is true for a $V \propto \phi^2$ theory or for other polynomial potentials.

The initial conditions are that the occupation number for these fluctuations must vanish: $n_{\mathbf{k}} = 0$. There are still zero-point fluctuations of the field with energy $E_{\mathbf{k}} = \hbar\omega_{\mathbf{k}}/2$, but these cannot gravitate. We do not need to understand *why* this is so; we can simply take it as an empirical fact that these zero point fluctuations, which are present in all fields today, and which would predict an energy density on the order of the Planck density of $\rho \sim 10^{95} \text{ gm cm}^{-3}$, are somehow not to

be included in the stress-energy tensor. In order to couple to real density fluctuations, these zero-point quantum fluctuations must somehow develop non-zero occupation numbers. This is easily seen to be inevitable. When the wavelength is much less than the horizon-scale, the effect of the universal expansion is negligible; the field fluctuations evolve just as they would in flat space with adiabatically conserved occupation number $n_{\mathbf{k}} = 0$. However, as the wavelength approaches the horizon scale, the frequency of oscillations $\omega \sim c/\lambda$ approaches the Hubble expansion rate and adiabaticity breaks down. Hopefully it is not too much of a stretch to accept that the result must be occupation numbers $n_{\mathbf{k}}$ of order unity as modes leave the horizon. We can then estimate the amplitude of these now real inflaton fluctuations. The energy per mode is $E_{\mathbf{k}} \sim \hbar\omega$, so the energy density is $\epsilon \sim L^{-3} \sum_{\mathbf{k}} \hbar\omega_{\mathbf{k}} \sim \int d^3k \hbar\omega_k \sim \hbar k^3 \omega_* \sim \hbar c^{-3} \omega_*^4$ where $\omega_* \sim H$. The stress-energy

tensor for massless free fields is $\epsilon = \langle \dot{\phi}_1^2 \rangle / 2c^2 + \langle (\nabla \phi_1)^2 \rangle / 2$, but these terms are equal, so we have $\epsilon \sim \langle \dot{\phi}_1^2 \rangle / c^2 \sim \omega^2 \langle \phi_1^2 \rangle / c^2$. Equating these two expressions for the energy density, and using $\omega \sim H$ for these trans-horizon scale modes, we find that the field variance is

$$\langle \phi_1^2 \rangle \sim \frac{\hbar H^2}{c}. \quad (32.12)$$

Before proceeding, it is interesting to note that the prediction for the field fluctuations derived here are the same as for a field in thermal equilibrium at a temperature $kT \sim \hbar H$. For such a field, the occupation number is exponentially small for wavelengths $k = \omega_{\mathbf{k}}/c \gg kT/\hbar c$, and most of the field variance comes from modes with $k \sim kT/c$. An alternative route to (32.12) is to show that the existence of the event horizon during inflation results in *Hawking radiation* of temperature $kT \sim H/\hbar kc$.

The clear prediction — and this is made more quantitative in the full treatment — is for inflaton field fluctuations at horizon exit of amplitude $\delta\phi \sim \sqrt{\hbar H^2/c}$. To understand how these couple to density and curvature fluctuations at the end of inflation, and subsequently to density fluctuations at horizon re-entry, consider first a region where $\delta\phi$ happens to be small. This region will inflate by a certain number of e -foldings, and will then re-heat to a density determined solely by the nature of the inflaton potential and its couplings to other fields. Now consider a region of the same initial size, but where the field fluctuation happens to be positive. The field in this region starts up ‘higher up the hill’, so this region inflates for slightly longer, and ends up occupying a slightly larger volume when it re-heats (to the same density as the unperturbed region). The extra expansion factor is $\exp(H\delta t)$, where δt is the time taken for the field to roll from $\phi = \phi_0 + \delta\phi$ to $\phi = \phi_0$. This is just $\delta t = \delta\phi/\dot{\phi}$. For small $\delta\phi$ we can expand the exponential as $\exp(H\delta t) \simeq 1 + H\delta t \simeq 1 + H\delta\phi/\dot{\phi}$. This is the excess of volume occupied by the perturbation region as compared to what it would have been had $\delta\phi$ been zero; clearly to replace a given volume in the background model by a slightly larger volume requires that the perturbed region have slightly positive curvature, which, as we have seen, can be related to the Newtonian potential fluctuations. Alternatively, for a large-scale perturbation which enters the horizon after matter domination, there will be an excess of proper mass within the perturbed region $\delta M/M \sim H\delta t$, which is just equal to the Newtonian potential fluctuation, which again is just equal to the density fluctuation at horizon re-entry. Either route yields a prediction for the late-time horizon scale density fluctuation given by (32.5).

The late-time density fluctuation amplitude is therefore set by the values of H and $\dot{\phi}$ at horizon exit. Since the field will be rolling slowly at the terminal velocity $\dot{\phi} = 4\lambda c\phi_0^3/\hbar H$, these will vary slowly with wavelength, so the prediction is for a spectrum with index n close to unity. As further consequence is that since the initial ‘zero-point’ fluctuations are statistically independent, so also will be the complex amplitudes for the density fluctuations; i.e. the prediction is that the density perturbations will take the form of a Gaussian random field.

Observations of the microwave background tell us that the density fluctuation at horizon re-entry is around $\delta_H \sim 10^{-5}$. Matching this requirement places a constrain on the interaction strength parameter λ (or its equivalent for other choices of the inflaton potential form). Using $H^2 \sim G\epsilon/c^2 \sim G\lambda\phi_0^4/\hbar c^3$ and $\dot{\phi} \sim \lambda c\phi_0^3/\hbar H$ we have

$$\delta_H \sim \sqrt{\frac{\hbar}{c}} \frac{H^2}{\dot{\phi}} \sim \left(\frac{G\phi_0^2}{c^4} \right)^{3/2} \lambda^{1/2}. \quad (32.13)$$

Now the first factor must be greater than unity for inflation to take place. In fact, we found that we needed $\phi_0 \gtrsim \sqrt{c^4/\epsilon G}$ where ϵ is the inverse of the number of e -foldings required to solve the horizon problem. This means that the pre-factor on the right hand side of (32.13) is around 100 (though the precise value is dependent on the energy scale of re-heating), and therefore a viable model must have a dimensionless interaction strength

$$\lambda \sim 10^{-4} \delta_H^2 \sim 10^{-12}. \quad (32.14)$$

Finally, while we have focused on the fluctuations in the inflaton field, since it is these which give rise to density fluctuations, the first part of the argument here can be used to predict the horizon-exit value of the amplitude of any fields which are effectively massless during inflation. In particular, the theory predicts that there should be fluctuations in the graviton field — gravitational waves that is — with amplitude on the order of the expansion rate in units of the Planck frequency. These waves are ‘frozen-in’ which the perturbation is outside the horizon and then start to oscillate on horizon re-entry. A high priority for future measurements of the microwave background anisotropies is to measure the strength of these waves.

32.3 Self-Ordering Fields

An alternative, and also highly attractive, possibility is that the seeds of structure may be due to *self-ordering fields*. The idea here is to have some scalar field or such-like which is initially in a highly disordered thermal state, but which has potential function of the kind invoked in spontaneous symmetry breaking. As the universe expands, the field temperature decreases and eventually it becomes energetically favorable for the field to fall into the minimum of the potential function. Such fields try to ‘comb themselves smooth’, but are frustrated in this due to the formation of *topological defects*. The most common example of this phenomenon at low energies is a ferro-magnetic material which, if cooled from high temperature, will undergo a phase transition and will develop domains which are bounded by walls. In cosmology, as at low energies, the character and evolution of such systems depends critically on the dimensionality of the field involved. Here we shall consider first 1-dimensional fields, which give rise to *domain walls*, but, unfortunately do not seem to be consistent the observed state of the universe. We then consider 2-dimensional fields, which, as we shall see, give rise to *cosmic strings*, and which are a more promising mechanism for generating cosmological structure.

32.3.1 Domain Walls

Consider a real scalar field with Lagrangian density

$$\mathcal{L}(\phi, \dot{\phi}, \nabla\phi) = \frac{1}{2c^2} \dot{\phi}^2 - \frac{1}{2} (\nabla\phi)^2 - V(\phi) \quad (32.15)$$

with potential function

$$V(\phi) = V_0 - \frac{m^2 c^2}{\hbar^2} \phi^2 + \frac{\lambda}{\hbar c} \phi^4 \quad (32.16)$$

as sketched in figure 32.3. This field has a negative mass parameter, and a self-interaction parameterized by the dimensionless constant λ . The potential has asymmetric minima at field values $\phi = \pm\phi_0$, the solutions of $dV/d\phi = 0$, where

$$\phi_0 = \sqrt{\frac{m^2 c^3}{4\lambda\hbar}}, \quad (32.17)$$

and has an unstable maximum at $\phi = 0$. If we require that $V(\pm\phi_0) = 0$, the constant V_0 is related to the parameters m , λ by

$$V(0) = V_0 = \frac{m^4 c^5}{16\lambda\hbar^2}. \quad (32.18)$$

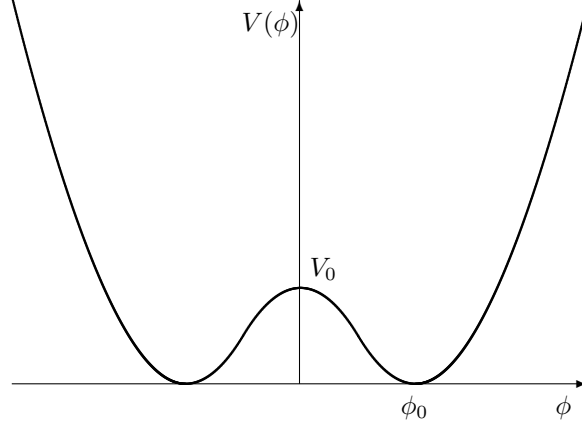


Figure 32.3: The potential function for a real scalar field involved in the generation of domain walls.

Now a field in thermal equilibrium at temperature T has energy per mode $E_{\mathbf{k}} = (n_{\mathbf{k}} + 1/2)\hbar\omega_{\mathbf{k}}$ with $n_{\mathbf{k}} = (e^{\hbar\omega_{\mathbf{k}}/kT} - 1)^{-1}$. If the field is effectively free and massless, the dispersion relation is simply $\omega_{\mathbf{k}} = ck$. Ignoring the zero point energy, and taking the universe to be a periodic box of size L , the energy density is

$$\epsilon(T) = L^{-3} \sum_{\mathbf{k}} n_{\mathbf{k}} \hbar \omega_{\mathbf{k}} = \int \frac{d^3k}{(2\pi)^3} n_{\mathbf{k}} \hbar \omega_{\mathbf{k}} \sim \frac{\hbar}{c^3} \omega_T^4 \sim \frac{(kT)^4}{(\hbar c)^3} \quad (32.19)$$

where $\omega_T = kT/\hbar$. This is just the Stefan-Boltzmann law. The energy density is also related to the mean square field fluctuations by the stress-energy tensor: $\epsilon \sim \dot{\phi}^2/c^2 \sim \omega^2 \phi^2/c^2$, so equating these two expressions for ϵ gives the root mean square field fluctuation at temperature T :

$$\langle \phi^2 \rangle_T^{1/2} \sim \sqrt{\frac{\hbar}{c}} \omega_T \sim \frac{kT}{\sqrt{\hbar c}}. \quad (32.20)$$

This says that the typical field value for a field in thermal equilibrium is proportional to the temperature (in natural units the root mean squared value of the field — which has dimensions of energy — is just equal to kT). If we use this to compute the potential energy density ϵ_{int} due to the interaction term we find

$$\epsilon_{\text{int}} \sim \frac{\lambda}{\hbar c} \langle \phi^2 \rangle^2 \sim \frac{\lambda (kT)^4}{(\hbar c)^3}. \quad (32.21)$$

Thus, provided the dimensionless interaction strength λ is much less than unity, as we shall assume, for a thermal state the interaction energy is a small perturbation to the total energy (32.19).

At high temperatures such that $kT \gg \sqrt{\hbar c} \phi_0$ the typical field values are much greater than ϕ_0 and the hill at the center of the potential is then relatively unimportant for the motion of the field. However, as the universe expands the temperature and the field amplitude decrease until the thermal field fluctuations become of order ϕ_0 and below this temperature the field will be trapped in one or other of the local minima. This *phase transition* occurs at a critical temperature

$$kT_c \sim \sqrt{\hbar c} \phi_0 \sim \frac{mc^2}{\sqrt{\lambda}}. \quad (32.22)$$

Now since the field configuration is initially highly spatially incoherent, different regions of space will want to settle into different minima. What happens (see simulation??) is that the field will become locally smooth within domains with value $\phi = \pm\phi_0$, since this minimizes the $(\nabla\phi)^2$ contribution to the energy density, with domains separated by domain walls where there is a strong localized gradient of the field. One can estimate the thickness of these walls on energetic grounds to be

$$\Delta x \sim \frac{\phi_0}{\sqrt{V_0}} \quad (32.23)$$

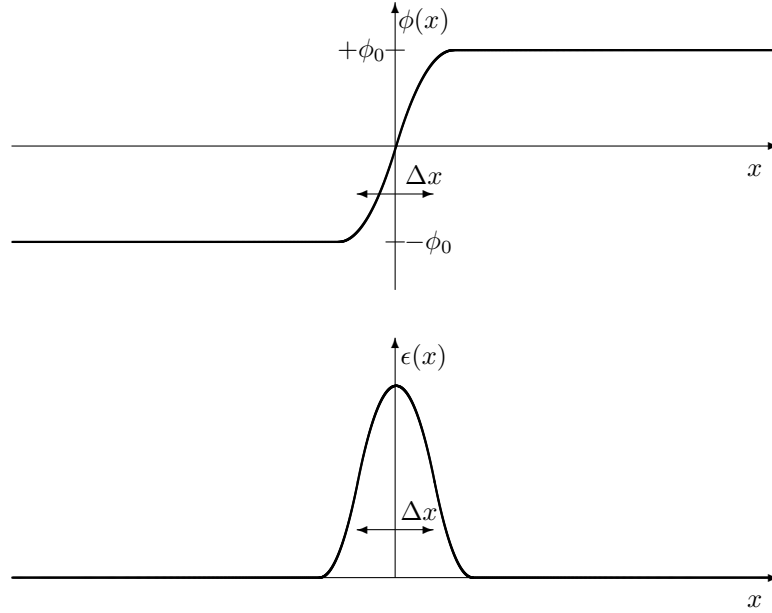


Figure 32.4: The upper panel shows the variation of the field passing through a domain wall of thickness Δx . We can estimate the width, and surface mass density of a domain wall, as follows. If the wall has width Δx then the typical field gradient within the wall is $\nabla\phi \sim \phi_0/\Delta x$. The energy density is then $\epsilon \sim (\phi_0/\Delta x)^2 + V_0$, so the density per unit area Σ is given by $c^2\Sigma \sim \epsilon\Delta x \sim \phi_0^2/\Delta x + V_0\Delta x$. If the wall is too thin, the gradient energy term becomes large while if the wall is too thick the potential term is increased. The total energy is minimized for $\Delta x \sim \phi_0/\sqrt{V_0}$. This is the width of a stable domain wall, for which the mass surface density is $\Sigma \sim \phi_0\sqrt{V_0}/c^2$.

(see figure 32.4 and its accompanying caption), and the mass-energy surface density in a stable wall is

$$\Sigma \sim \frac{\phi_0\sqrt{V_0}}{c^2}. \quad (32.24)$$

The single static planar wall is highly idealized. The initial walls configuration will be highly disordered. Again, energetic considerations tell us that the system will evolve to minimize the total energy in the walls. A simply connected region bounded by a wall will tend to shrink. In doing so, it will convert the potential energy into kinetic energy, so we expect walls to be moving at speeds on the order of the speed of light. Such a region will shrink to zero size on a time scale of order its size divided by c . The energy released will propagate away as waves, but these will damp adiabatically, so between the walls the field will remain relatively smooth. The expectation then is that any regions smaller than the horizon scale $\sim ct$ will disappear, but the field at separations bigger than the horizon scale will remain uncorrelated; the field dynamics will result in domains, at any time, on the order of the horizon size. In fact we expect a *scaling solution* where the field looks the same at any time save for scaling of the mean wall separation with the horizon scale.

This behavior is illustrated, for a field in 2-dimensions in figure 32.5. The equations for a scalar field in 2-dimensions, with a W-shaped potential and with a weak damping term were evolved numerically using a simple centered algorithm. The initial field was a Gaussian random field with a flat spectrum, aside from a smoothing with a small kernel to make the field smooth at the spatial sampling scale. The initial field amplitude was somewhat higher than ϕ_0 , but the damping term cools the field, which starts to separate into domains where $\phi \simeq \pi\phi_0$. As the system evolves, enclosed regions can shrink to zero size and then disappear with a release of energy in a circular out-going wave. The scale of the walls gradually increases with time.

If we say there is on the order of one wall, of area $\sim (ct)^2$ per horizon volume $(ct)^3$, the mean

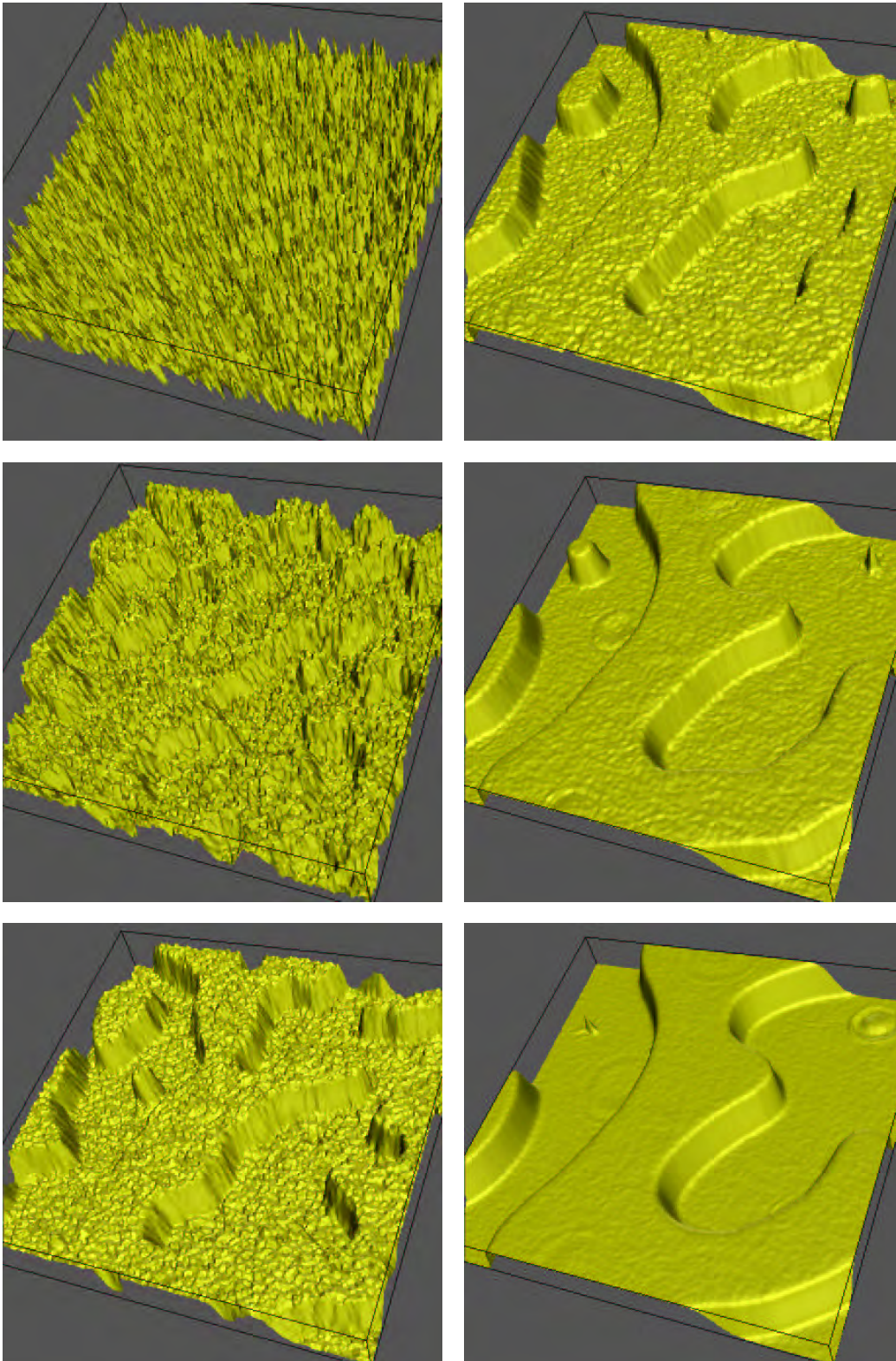


Figure 32.5: A set of frames from a computation of spontaneous symmetry breaking. The full animation can be viewed at <http://www.ifa.hawaii.edu/~kaiser/wavemovies>

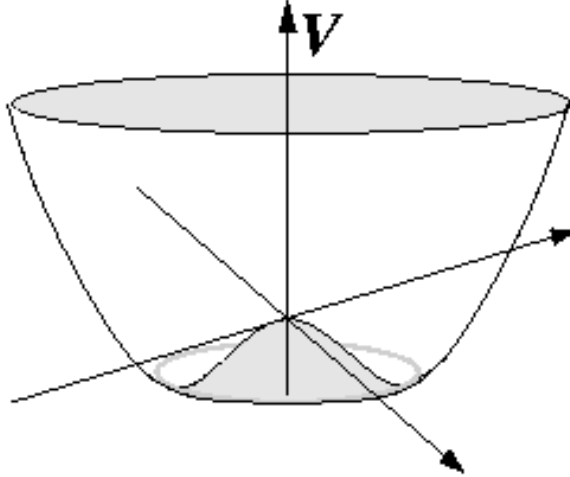


Figure 32.6: The potential function for a 2 component scalar field involved in the generation cosmic strings.

mass-energy density in walls is

$$\rho_{\text{walls}} \sim \frac{\Sigma}{ct}. \quad (32.25)$$

This is a serious problem, since the density of the matter, or radiation, in the universe is $\rho = 3H^2/8\pi G$, which scales as $1/t^2$. Thus the walls will rapidly come to dominate the universe; and one would have very large density inhomogeneity on the horizon scale. This is not what is observed.

32.3.2 Cosmic Strings

Now consider a two-component scalar field ϕ for which the potential $V(\phi)$ is the two dimensional analog of (32.16) as illustrated in figure 32.6. This is often called a ‘Mexican-hat’ or ‘sombrero’ potential. The minimum energy is on the circle $|\phi| = \phi_0$ and the field will try to relax towards this. There will be oscillations about the minimum, but the amplitude of these decreases adiabatically and the field will develop regions where the field lies in the minimum and varies slowly with position. While a completely uniform field is energetically favored, just as for domain walls, the assumed initial incoherence of the field limits the scale of coherence; the formation of a single infinitely large domain being frustrated by the formation of a network of *cosmic strings* — localized regions of energy density where the field sits at $\phi = 0$.

To get an idea of the topology of the initial string network, picture the initial field as filtered white noise with some coherence length set by the filter, and model the initial field evolution as simply rolling ‘downhill’ to the nearest minimum. In most places the field will vary quite smoothly with position, with $\nabla\phi \sim \phi_0/\lambda$, where λ is the coherence length for the initial field. However, at positions where both components of the field ϕ_1 and ϕ_2 were initially very small there will be very large gradients — and therefore very high energy density — localized near the regions where $\phi = 0$ initially. Now in 3-dimensional space the field ϕ_1 will generally vanish on a surface, and similarly for ϕ_2 , so the regions where both components vanish are the intersection of these surfaces; i.e. on lines, or ‘strings’. There is another way of looking at this; if we traverse an arbitrary loop in the real three dimensional space, the field moves along a closed trajectory in 2-dimensional field space. If the field is trapped in the circular trough then it is possible that the field trajectory will pass once around the brim of the sombrero; we would say that this loop has a *winding number* of one (or minus one, depending on the sense of rotation of the field). Now this winding number is a *topological invariant*; we can make a continuous deformation of the loop and the winding number cannot change, provided the field is everywhere confined to the minimum energy circle.

Now the energy for this toy mode is in fact divergent. What is energetically more favorable is

for the field to sit at $\phi = 0$ along the string axis, with the field falling to the potential minimum within some distance — the string thickness Δx . We can estimate what this is as follows. Consider a perfectly axi-symmetric field with unit winding number, and let's assume to start with that the field everywhere (except perhaps exactly on the axis) lies in the minimum. We can choose the spatial coordinate axes such that the field is

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \frac{\phi_0}{r} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (32.26)$$

The field gradient term in the energy density is

$$(\nabla\phi)^2 \equiv \frac{\partial\phi_i}{\partial r_j} \frac{\partial\phi_i}{\partial r_j} = \frac{\phi_0^2}{r^2}. \quad (32.27)$$

If we integrate this from r_{\min} to r_{\max} we find a contribution to the line density

$$\sigma c^2 = \phi_0^2 \int \frac{d^2r}{r^2} = 2\pi\phi_0^2 [\log(r)]_{r_{\min}}^{r_{\max}} \quad (32.28)$$

so the line density diverges logarithmically if we let $r_{\min} \rightarrow 0$. Now consider a crude model in which the field lies in the zero potential for $r \gtrsim \Delta x$ but has $\phi \simeq 0$ for $r \lesssim \Delta x$ with some smooth transition between these. The gradient contribution to the line density will then contain a component $\sigma \sim \phi_0^2 \log(r_{\max}/\Delta x)/c^2$ and there will be a contribution from the potential $\sigma \sim V_0 \Delta x^2/c^2$, so the total line density will be

$$\sigma \simeq \frac{1}{c^2} (\alpha \phi_0^2 \log(r_{\max}/\Delta x) + \beta V_0 \Delta x^2) \quad (32.29)$$

where α, β are dimensionless coefficients of order unity. Setting the derivative of this with respect to Δx to zero gives the string width for minimum line density (i.e. energy)

$$\Delta x \sim \frac{\phi_0}{\sqrt{V_0}} \quad (32.30)$$

just as above for domain walls. The linear mass-density is

$$\sigma \sim \frac{V_0 \Delta x^2}{c^2} \sim \frac{\phi_0^2}{c^2}. \quad (32.31)$$

The generation of this network of strings during a phase-transition involving a two-component field is known as the *Kibble mechanism*.

Just as for walls, the initial string network will be quite contorted. Calculating the stress-energy tensor (or more simply applying energetics arguments) again tells us that the string network will not be static but will develop transverse velocities $\sim c$. The character of the evolution of the string network is qualitatively different, however. Strings can reconnect when they intersect and so loops can be chopped off the network. Such a loop may further intersect itself, but there are stable loop configurations which sit there and oscillate. Such loops have large quadrupole moments and are moving relativistically, so they are quite efficient at radiating gravitational radiation. One can show that such loops will decay after $\sim c^2/G\sigma$ oscillations.

It is reasonable to expect that such a network will evolve towards a *scaling solution* with roughly one long string per horizon volume (that being the distance a string section will typically move). If we estimate the mean density in such strings we find

$$\rho_{\text{string}} \sim \frac{\sigma}{(ct)^2}. \quad (32.32)$$

This is quite different from the case of walls where the density falls as $1/t$; here the string density evolves in the same manner as the mean density of matter or radiation, whichever happens to dominate. Thus we expect to have a constant fraction of the total energy density in string at any time. That the system should tend towards the scaling solution seems very reasonable — if

there were too much string in some region then the interconnection would be more vigorous than on average and *vice-versa* — and early simulations of the evolution of the string network were performed and seemed to confirm this. This led to a simple picture of a continuously evolving network of long strings with a debris of oscillating loops lying around (whose mass spectrum could be crudely estimated from the dynamics of loop production) and it was supposed that the loops would act as point-like ‘seeds’ for structure formation. In this picture the density fluctuations would be highly non-Gaussian, in contrast to the fluctuations arising from inflation for instance. However, subsequent higher resolution simulations showed that this picture was somewhat flawed. The simple intuitive expectation (and low resolution simulations) did not incorporate an important feature; each time strings chop, discontinuities form and propagate along the string as traveling waves. As time proceeds the network develops more and more fine scale structure. It is still suspected that a scaling solution will result, but performing the needed simulations is quite a challenge. Analysis of the higher resolution simulations suggest that the simple one loop-one object picture for structure formation was overly simplified and that the myriad of rapidly moving loops produces something more akin to Gaussian fluctuations.

Perhaps the nicest feature of the string model is that the model has only one free parameter — the line-density of the strings σ . This sets the amplitude of density fluctuations at horizon crossing. We can estimate this as follows: The total density is $\rho_{\text{tot}} \sim H^2/G$, so the ratio of string to total density is

$$\frac{\rho_{\text{string}}}{\rho_{\text{tot}}} \sim \frac{\sigma G}{c^2 H^2 t^2} \sim \frac{G\sigma}{c^2} \sim \frac{G\phi_0^2}{c^4} \sim \frac{G}{c^4} \frac{(kT_c)^2}{\hbar c} \simeq \left(\frac{kT_c}{E_{\text{pl}}} \right)^2 \quad (32.33)$$

where we have used $\sigma \sim \phi_0^2/c^2$, $E_{\text{pl}} \sim \sqrt{\hbar c^5/G}$ and $kT_c \sim \sqrt{\hbar c}\phi_0$. Now the energy density fluctuations in the strings are of order unity at the horizon scale — there being on the order of one string per horizon — and therefore the total density perturbation at horizon crossing is

$$\frac{\delta\rho}{\rho} \sim \frac{\rho_{\text{string}}}{\rho_{\text{tot}}} \sim \left(\frac{kT_c}{E_{\text{pl}}} \right)^2. \quad (32.34)$$

The gravity associated with the string network drives motions of the rest of the matter and thus excites growing density perturbations which could plausibly account for the structure we see. This is very encouraging. First, the theory naturally generates perturbations with scale invariant amplitude at horizon crossing; the Harrison-Zel’dovich spectrum. Second, for strings formed at around the GUT scale of $kT_c \sim 10^{16} \text{ GeV} \sim 10^{-3} E_{\text{pl}}$, this predicts $\delta \sim 10^{-6}$, which is not far from that observed. Unfortunately, while the formation of strings at the GUT time is not mandatory, the formation of monopoles is, and these monopoles are a disaster. They can be gotten rid of by inflation — and one major motivation for inflation was the monopole problem — but then one would inflate away the strings as well.

An interesting feature of the negative tension is that the stress-energy tensor for an infinite static string is trace-free and consequently the string produces no tidal field. Outside of such a string spacetime is flat, but it is topologically different from ordinary Minkowski space in that there is a small deficit in azimuthal angle $\delta\varphi = 4\pi G\sigma/c^2$. A particularly distinctive features of the cosmic string model arises *via* gravitational lensing; lensing by long strings can produce a unique signature both in images of distant galaxies and in the microwave background.

There are other defects which can form. We have already discussed formation of walls from 1-dimensional fields, and monopoles from three dimensional fields, both of which are dangerous. A four-component field is more benign and gives rise to *texture*. A texture is not a topologically stable defect. Textures are most easily pictured in 1-D — where they result from having a 2-component field — and such a texture can shrink until $(\nabla\phi)^2 \sim V(0)$ at which point it will unwind.

Chapter 33

Probes of Large-Scale Structure

33.1 Introduction

The inflationary/CDM model makes quite definite predictions for the power spectrum of the density perturbations emerging from the early universe, and also that the fluctuations should take the form of a Gaussian random field. On small-scales, structures have gone non-linear, resulting in galaxies and groups and clusters of galaxies. In many ways the cleanest tests of cosmological structure formation theory comes from measurements of large-scale structure, where the fluctuations are still in the linear regime ($\delta\rho/\rho \ll 1$), and thus directly reveal the initial conditions. There are four classic probes of large scale structure: *galaxy clustering* §33.2; deviations from pure Hubble expansion or *bulk-flows* §33.3; anisotropy of the microwave background §33.4; and the distortion of the shapes of faint galaxies from *weak lensing* §33.5. Rather than try to present a snap-shot of the current results in each of these fields — which would rapidly become out-dated — we will focus on the basic physical principles and set out what are the key strengths and weaknesses of the different probes.

33.2 Galaxy Clustering

In some ways, galaxy clustering is the most straightforward probe of large-scale structure. The galaxies are like a dust of test-particles which flow with the matter in general and their spatial distribution thereby traces the underlying total density. By measuring the statistical properties of the galaxy distribution — auto-correlation function, power spectrum and higher order statistics — we can test the predictions of early universe theory. There are two basic techniques that are used here. One is to use *angular surveys*, where one measures the angular positions of galaxies and extracts for example $\omega(\theta)$, the two-point angular galaxy correlation function. The other is to use *redshift surveys*, where one has the three-dimensional distribution of galaxies, with distance being inferred from the redshift. The advantages of angular surveys is that they are cheap and deep, particularly with the advent of large-format CCD detectors. The disadvantage is that the structure tends to get washed out in projection. Also, because we see a large number of structures along any line of sight, the projected fluctuations become Gaussian by virtue of the central limit theorem, making it hard to test for primordial non-Gaussianity. Redshift surveys avoid the latter problems, but are much more expensive to obtain; one can determine the positions of on the order of 10^5 galaxies in a 30 minute observation with a large telescope, reaching magnitude $m_R \sim 26 - 27$. The deepest redshift surveys currently feasible are restricted to $m_R \lesssim 24$, and take many hours per field, with much smaller field of view. A complication in analysing such surveys is that the redshift does not strictly measure the distance; the peculiar velocities associated with the growing structure result in *redshift space distortion*, though, as we shall see, we can turn this to our advantage and use it as a way to determine the density parameter.

33.2.1 Redshift Surveys

The technique here is to take images of a field to determine the locations of the galaxies in some chosen range of magnitudes, and then to obtain optical spectra, this typically being done using either a multi-slit spectrograph or using optical fibres to conduct the light from the focal plane to a single spectrograph. State of the art systems are capable of collecting several hundred spectra simultaneously. The result of this exercise is a catalog containing angular positions and redshifts for all of the galaxies in the chosen region of the sky which meet the selection criterion.

33.2.2 Poisson Sample Model

In order to analyze this kind of catalog and relate the results to theoretical predictions we need to have some kind of model for how galaxies are related to the underlying density field. Now one possibility is a ‘what-you-see-is-what-you-get’ model, where the luminous and dark matter densities are precisely the same. However, this is not really viable; first of all, we know empirically from rotation curve studies and from gravitational lensing that galaxies have extended haloes. Evidently the luminous matter is much clumpier than the dark matter on small scales. This property is also readily understood physically; when galaxy halo sized objects form, the baryons can dissipate their binding energy quite efficiently *via* collisional effects (see later), and so can sink to the bottom of the potential well while the dark matter, being like a collisionless fluid cannot contract because of phase-space density conservation. Second, there is good reason to believe that the luminosity of a galaxy is not simply proportional to the amount of baryons in the region from which it forms. There is almost certainly a strong stochastic element to the luminosity to baryon mass density. Young galaxies are very bright, and then fade, and galaxies may undergo subsequent bursts of star formation. These considerations motivate the model for the relation between galaxies and the underlying, or total mass, density field, known as the *Poisson sampling model*. One way to visualize this model is to consider a sea of particles so numerous that they have an effectively continuous number density $\rho(\mathbf{r})$. Now assign to each particle a uniformly distributed independent random number p in the range $0 \leq p \leq 1$ and paint those particles with $p < \epsilon$ red, where ϵ is some small constant. The red particles define a Poisson sample of the continuous field $\rho(\mathbf{r})$.

Mathematically one can formalize this by imagining space to be divided into tiny cubical cells with labels i having volume δV_i and occupation number n_i . The mean occupation number is

$$\bar{n}_i = \epsilon \rho(\mathbf{r}_i) \delta V_i. \quad (33.1)$$

The probability that a cell is occupied is

$$P(n_i = 1) = \bar{n}_i \quad (33.2)$$

and, in the limit that the cells become infinitesimal, the probability that $n_i > 1$ becomes negligible, so the probability that the cell is empty is

$$P(n_i = 0) = 1 - \bar{n}_i. \quad (33.3)$$

So far galaxies have been modelled as identical, featureless points. The Poisson sample model can be extended to incorporate a distribution of intrinsic luminosities or other characteristics. If the *luminosity function* — the distribution function for luminosities — is $\phi(L)$ then the extension of the Poisson sample model is that the probability that a cell in the 4-dimensional position-luminosity space is occupied is

$$P(n_i = 1) = \epsilon \rho(\mathbf{r}) \phi(L) d^3V dL. \quad (33.4)$$

This model says that one draws particles at random from the quasi-continuous sea, and then assigns each of them a luminosity drawn at random from the luminosity function.

The Poisson sample model underlies most observational work in galaxy clustering — so much so that the adoption of this model is rarely explicitly stated in the literature. However, we should remember that it is only a model, and one that cannot be strictly correct. It is the opposite extreme from the ‘light traces mass exactly’ WYSIWIG model described above. Reality probably lies somewhere in between.

33.2.3 Correlation Functions

In this model, the probability that two particular cells (labelled 1 and 2), separated by \mathbf{r}_{12} are both occupied is just the product of the probabilities that each cell is occupied $P(n_1 = 1, n_2 = 1) = \epsilon^2 \delta V^2 \rho(\mathbf{r}_1) \rho(\mathbf{r}_2)$. We are ignoring here, for simplicity, the distribution of luminosities and are treating galaxies as identical featureless points. The mean value of the product of cell occupation numbers taken over all pairs of cells with separation \mathbf{r}_{12} is then

$$\overline{n_1 n_2} = \epsilon^2 \rho(\mathbf{r}_1) \rho(\mathbf{r}_2) \delta V_1 \delta V_2. \quad (33.5)$$

This is, if you like, the average taken over an ensemble of realizations of points generated by sampling a single given density field $\rho(\mathbf{r})$. Now the density field $\rho(\mathbf{r})$ is itself a random process and Nature, by creating us at a particular location in space, has provided us with a specific realization of this field in the region around us that we can observe. Learning about the specific density field configuration around us is interesting, of course, but for testing theories of structure formation, we are more interested in learning about the statistical properties of the ensemble of all density fields. Under the Copernican hypothesis — that we are placed at a random location in a statistically homogeneous random universe — these properties can be encoded in the hierarchy of correlation functions. A nice feature of the Poisson sample model is that it makes it relatively easy to relate the correlation function of the underlying density field $\rho(\mathbf{r})$ to counts of galaxies.

Consider the ensemble average of the product of the occupation numbers for a pair of cells

$$\langle n_1 n_2 \rangle = \epsilon^2 \langle \rho(\mathbf{r}_1) \rho(\mathbf{r}_2) \rangle \delta V_1 \delta V_2 \quad (33.6)$$

and defining the fractional density perturbation $\delta(\mathbf{r})$ such that $\rho(\mathbf{r}) = \bar{\rho}(1 + \delta(\mathbf{r}))$ this becomes

$$\langle n_1 n_2 \rangle = \epsilon^2 \bar{\rho}^2 \delta V^2 (1 + \xi(\mathbf{r}_{12})) \quad (33.7)$$

where $\xi(\mathbf{r}_{12}) = \langle \delta(\mathbf{r}) \delta(\mathbf{r} + \mathbf{r}_{12}) \rangle$ is the auto-correlation function of the density fluctuation field. If there were no underlying structure ($\xi(\mathbf{r}) = 0$) the expectation of the product of cell numbers would therefore be $\langle n_1 n_2 \rangle = \epsilon^2 \bar{\rho}^2 \delta V^2$. From a purely observational standpoint, the two point correlation function $\xi(\mathbf{r})$ measures the fractional excess probability that a pair of cells are both occupied, given that they have a separation \mathbf{r}_{12} . It also measures the fractional excess probability that an observer living on a galaxy has a neighbour in a cell at that relative location as compared to an observer at a randomly chosen location (problem??).

Now we can estimate $\xi(\mathbf{r})$ by counting pairs with an appropriate separation. Obviously we must average over some range of separation in order to get a sensible number of pairs. Given some bounded region with cells labelled i we can write the number of pairs of galaxies with separation $|\mathbf{r}_{ij}|$ in the range $r - \Delta r/2 < |\mathbf{r}_{ij}| < r + \Delta r/2$ as a double sum over cells:

$$N_p(r, \Delta r) = \sum_i \sum_j n_i n_j S(|\mathbf{r}_{ij}|; r, \Delta r) \quad (33.8)$$

where $S(|\mathbf{r}_{ij}|; r, \Delta r)$ is unity (zero) if the the separation lies in (outside) the allowed range. Taking the geometry of the observed region to be given, the expectation value of N_p is

$$\langle N_p(r, \Delta r) \rangle = \sum_i \sum_j \langle n_i n_j \rangle S(|\mathbf{r}_{ij}|; r, \Delta r) \quad (33.9)$$

and using (33.7) this becomes

$$\langle N_p(r, \Delta r) \rangle = \epsilon^2 \bar{\rho}^2 \sum_i \delta V_i \sum_j \delta V_j (1 + \xi(\mathbf{r}_{ij})) = \epsilon^2 \bar{\rho}^2 \int d^3 r_i \int d^3 r_j (1 + \xi(\mathbf{r}_i - \mathbf{r}_j)). \quad (33.10)$$

Let's assume that the range of separations is quite small $\Delta r \ll r$, and that $\xi(\mathbf{r})$ depends only on $r_{ij} = |\mathbf{r}_{ij}|$ and is such that, at separation \mathbf{r} the scale for changes in $\xi(\mathbf{r})$ is on the order of r ;

i.e. $d\xi/dr \sim \xi/r$. Under these assumptions we can take $\xi(\mathbf{r}_{ij})$ to be the constant value for the center of the averaging bin $\xi(r)$, so

$$\langle N_p(r, \Delta r) \rangle = \bar{n}^2 (1 + \xi(r)) \int d^3 r_i \int d^3 r_j \quad (33.11)$$

where $\bar{n} = \epsilon \langle \rho \rangle$ is the mean density of galaxies.

To convert this to a practical estimator for $\xi(\mathbf{r})$ we just need some way to estimate the double spatial integral here. Since the survey volume is usually specified by giving a range of angular coordinates on the sky, perhaps subject to a set of ‘bad regions’ (bright stars, data drop-outs etc), this is most easily computed in a Monte-Carlo fashion by generating a very large set of points with random independent positions and then applying a filter to restrict them to the survey region. Let the mean number density of random objects be \bar{n}_{rand} and that of real objects be $\bar{n}_{\text{real}} = \bar{n}_{\text{rand}}/\alpha$. We denote the number of real pairs by $\text{DD} = N_{\text{p,real}}$ (for ‘data-data’ pairs) and define $\text{RR} = N_{\text{p,rand}}/\alpha^2$ (for ‘random-random’ pairs, but applying a dilution factor $1/\alpha^2$ to allow for their larger mean number density). The expectation value of DD is then

$$\langle \text{DD} \rangle = (1 + \xi(r)) \text{RR}. \quad (33.12)$$

A fair correlation function estimator $\hat{\xi}(r)$ — i.e. one for which $\langle \hat{\xi}(r) \rangle = \xi(r)$ — is

$$\hat{\xi}(r) = \frac{\text{DD}}{\text{RR}} - 1. \quad (33.13)$$

This is a simple, and often used estimator, but it has a number of drawbacks:

- The mean density of galaxies must usually be determined from the actual density of galaxies in the survey. This introduces a so-called *integral constraint* that the volume integral of the two-point function estimator is forced to vanish. Since $\xi(r)$ tends to be positive at small separations, this introduces a negative bias in $\hat{\xi}(r)$ at separations approaching the size of the survey. This is largely unavoidable. One can try to model the effect, but this requires *assuming* some form for $\xi(r)$ which doesn’t seem fair. The best solution is to obtain more data!
- The density fluctuations couple with the sharp boundary of the survey to produce spurious results. Say the density field is rather smooth, with fluctuations on some ‘coherence-scale’ r_c . If we slice through a structure the resulting density field has the appearance of sharp, or high spatial frequency, structure and this will cause an error in measurements of $\xi(r)$ at small scales. This is not a *bias*, since the error is as likely to be positive or negative, but is an unwanted kind of ‘cross-talk’ between large-scale fluctuations and smaller scales which complicates the error analysis. This problem can be ameliorated somewhat with an alternative estimator (see below), or by ‘apodizing’ the survey; i.e. applying weights which taper off towards the edges.
- A proper error analysis is very difficult, since, as we shall see, the estimates of $\xi(r)$ for different bins are not independent.

The results of redshift surveys indicate that the galaxy-galaxy correlation function has a power-law form $\xi(r) \simeq (r/r_0)^{-\gamma}$ for $r \leq 10\text{Mpc}$ and with $\gamma \simeq 1.8$. On larger scales the correlation function departs from a power law.

Regarding the error analysis, it has often been suggested that one estimate the uncertainty in $\hat{\xi}(r)$ by some form of ‘bootstrap’ approach. An example is to take the actual data, but assign random weights to the galaxies (e.g. one might discard half of the galaxies at random). The differences between such an ensemble of such estimates provides some idea of the fluctuations $\hat{\xi}(r)$ about the true value, but is not usually what one wants. There are two elements of randomness in the generation of a sample of galaxies. The first is the realization of a particular density field $\rho(\mathbf{r})$ from the ensemble of all statistically equivalent initial conditions. Equivalently, by ergodicity, this is the randomness arising from our particular randomly chosen location. The second is the randomness inherent in the

assumed Poisson sampling of the density field. Now on large scales the former tends to dominate, while bootstrap methods are only sensitive to the latter, sub-dominant, type of uncertainty. The latter type of error can best be thought of as a ‘measurement error’; just as in photometry, where we have $1/\sqrt{N}$ fluctuations in the flux from the number of photons, here we have $1/\sqrt{N}$ fluctuations in any estimate of $\delta(\mathbf{r})$ arising from the finite number of galaxies. On large-scales, however, the number of galaxies per ‘structure’ is very large. Superclusters — objects living right at the boundary between the linear and non-linear regimes — contain hundreds or even thousands of galaxies. The $1/\sqrt{N}$ fluctuations — often called the *sampling variance* or *cosmic variance* arising from the finite number of structures is much larger than $1/\sqrt{N_{\text{gals}}}$. Since, in a magnitude limited survey, the mean number density of galaxies decreases with distance, this suggests that one should give more weight to more distant galaxies. To understand the estimation of uncertainties in $\hat{\xi}(r)$ and make these ideas more precise we now turn to power-spectrum estimation.

33.2.4 The Power Spectrum

The two-point correlation function $\xi(r)$ is the Fourier transform of the power spectrum $P(k)$ and *vice versa*, so the two statistics are mathematically equivalent. One can estimate $P(k)$ as follows:

First take the Fourier transform of the galaxies, considering them to be the function consisting of a set of δ -functions:

$$\tilde{f}(\mathbf{k}) = \sum_{\text{galaxies, } g} e^{i\mathbf{k}\cdot\mathbf{r}_g} = \sum_{\text{cells, } i} n_i e^{i\mathbf{k}\cdot\mathbf{r}_i}. \quad (33.14)$$

Multiplying this by its complex conjugate and taking the expectation value gives

$$\langle |\tilde{f}(\mathbf{k})|^2 \rangle = \sum_i \sum_j \langle n_i n_j \rangle e^{i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j)}. \quad (33.15)$$

Using (33.7), and assuming, for simplicity that the mean number density of galaxies is spatially constant, we have

$$\langle |\tilde{f}(\mathbf{k})|^2 \rangle = \sum_i \langle n_i^2 \rangle + \sum_i \delta V_i \sum_{j \neq i} \delta V_j (1 + \xi(\mathbf{r}_i - \mathbf{r}_j)) e^{i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j)} \quad (33.16)$$

where we have realized that (33.7) applies only for $i \neq j$. Now since $n_i = 0$ or 1 , $n_i^2 = n_i$ and hence $\langle n_i^2 \rangle = \bar{n} \delta V$, and, converting the sums to integrals, we have

$$\langle |\tilde{f}(\mathbf{k})|^2 \rangle = \bar{n} \int d^3r W(\mathbf{r}) + \bar{n}^2 \int d^3r W(\mathbf{r}) \int d^3r' W(\mathbf{r}') (1 + \xi(\mathbf{r} - \mathbf{r}')) e^{i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r}')}. \quad (33.17)$$

Where $W(\mathbf{r})$ is the ‘window function’ describing the shape of the survey volume. Next, using $W(\mathbf{r}) = (2\pi)^{-3} \int d^3k \tilde{W}(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{r}}$, and $\tilde{W}(\mathbf{k}) = \int d^3r W(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}$, we have

$$\langle |\tilde{f}(\mathbf{k})|^2 \rangle = \bar{n} \tilde{W}(0) + \bar{n}^2 \tilde{W}^2(\mathbf{k}) + \int \frac{d^3k'}{(2\pi)^3} \int \frac{d^3k''}{(2\pi)^3} \tilde{W}(\mathbf{k}') \tilde{W}^*(\mathbf{k}'') \int d^3r e^{i(\mathbf{k}' - \mathbf{k}'')\cdot\mathbf{r}} \int d^3z \xi(\mathbf{z}) e^{i(\mathbf{k} - \mathbf{k}')\cdot\mathbf{z}} \quad (33.18)$$

$$= \bar{n} \tilde{W}(0) + \bar{n}^2 \tilde{W}^2(\mathbf{k}) + \int \frac{d^3k'}{(2\pi)^3} \int \frac{d^3k''}{(2\pi)^3} \tilde{W}(\mathbf{k}') \tilde{W}^*(\mathbf{k}'') (2\pi)^3 \delta(\mathbf{k}' - \mathbf{k}'') \int d^3z \xi(\mathbf{z}) e^{i(\mathbf{k} - \mathbf{k}')\cdot\mathbf{z}} \quad (33.19)$$

$$= \bar{n} \tilde{W}(0) + \bar{n}^2 \tilde{W}^2(\mathbf{k}) + \int \frac{d^3k'}{(2\pi)^3} P(\mathbf{k} - \mathbf{k}') |\tilde{W}(\mathbf{k}')|^2. \quad (33.20)$$

Thus simply by taking the square of the transform of the positions of the galaxies we have obtained a quantity whose expectation value is, aside in essence, a convolution of the power spectrum of the density field $\rho(\mathbf{r})$ with a kernel which is the square of the transform of the window function. We can identify the terms on the right hand side with a \mathbf{k} -independent ‘shot-noise’ contribution, the transform of the sample volume squared, and finally a term arising from the actual fluctuations in

the underlying density field $\rho(\mathbf{r})$. The appearance of the convolving kernel is readily understood; if the survey has size L then we do not expect to be able to ‘resolve’ features in the power spectrum with $\delta k \lesssim 1/L$. If we are interested in the spectrum a spatial frequency $k \gg L^{-1}$ where L is the size of the survey, then, since $\tilde{W}(\mathbf{k})$ has width $\Delta k \sim 1/L$, we can neglect the effect of the convolving kernel and we have, as a fair estimator of the power spectrum

$$\hat{P}(\mathbf{k}) = |\tilde{f}(\mathbf{k})|^2 - \bar{n}\tilde{W}(0) - \bar{n}^2\tilde{W}^2(\mathbf{k}). \quad (33.21)$$

To obtain this we took the transform of the galaxies, squared it, and then subtracted the expectation of the power spectrum for random galaxies. This is very similar to the estimator of the correlation function (33.13) above. Now there is an alternative. We could have subtracted the power-spectrum for a random catalog of objects, suitably scaled, from $\tilde{f}(\mathbf{k})$ before squaring. If we define the scaled transform of a very numerous catalog of random galaxies $\tilde{f}(\mathbf{k})_{\text{rand}} = \alpha^{-1} \sum e^{i\mathbf{k}\cdot\mathbf{r}}$ then we find that

$$\langle |\tilde{f}(\mathbf{k}) - \tilde{f}_{\text{rand}}(\mathbf{k})|^2 \rangle = \bar{n}\tilde{W}(0) + \int \frac{d^3k'}{(2\pi)^3} P(\mathbf{k} - \mathbf{k}') |\tilde{W}(\mathbf{k}')|^2. \quad (33.22)$$

This estimator has somewhat superior performance. The analog for estimating the correlation function is

$$\hat{\xi}(r) = \frac{\text{DD} - 2\text{DR} + \text{RR}}{\text{RR}} \quad (33.23)$$

where DD and RR are as before and where DR is the number of data-random pairs. This estimator is somewhat less affected by sharp edges of the survey volume.

The power-spectrum approach makes it easier to estimate the uncertainty. In inflationary models, the density field is predicted to have Gaussian statistics; i.e. the different Fourier modes are uncorrelated. The same is true for the spectrum of the galaxies. This means that the measured ‘raw’ power (before subtracting off the shot noise from the galaxy discreteness that is) is some positive speckly function with speckle width $\Delta k \sim 1/L$. The expectation value of the measured power is the true power (plus the expected shot noise power), but there are fractional fluctuations about the mean of order unity. The fractional fluctuations in the measured power for a band of power are at the $1/\sqrt{N}$ level, but where N is the number of distinct Fourier modes. For a cubical survey volume, the spacing of such independent modes is just $\Delta k = 2\pi/L$ and the concept can also be made precise for non-cubical survey volumes. Thus, in power spectrum estimation, and assuming Gaussian initial conditions, error estimation is fundamentally a counting exercise.

Comparing the two approaches, it is interesting to note that the correlation function estimator is relatively insensitive to the survey geometry, whereas with for the power spectrum the result is convolved with a window function. This becomes a serious problem for e.g. pencil beam surveys, where the kernel function becomes a thin disk (the transform of a needle). One consequence of this is that when one estimates the spectrum at some low spatial frequency, the result is dominated by high spatial frequency power being scattered down by the extended ‘side-lobes’ of the kernel function. However, for sensible survey geometries the advantage of being able to simply estimate the uncertainty in the power makes it the technique of choice.

Power spectrum estimates from e.g. the SDSS and 2df surveys are now probing well into the linear regime. The theoretical prediction is that the power, which rises with decreasing spatial frequency at small scales as $k^{-1.2}$, will peak at a scale corresponding to the horizon scale at z_{eq} and will then decrease asymptotically as $P(k) \propto k$ — the $n = 1$ Harrison-Zel’dovich spectrum — to very large scales.

33.2.5 Redshift Space Distortion

The power spectrum is a useful way to reveal a rather interesting property of the clustering pattern in redshift space. So far we have pretended that redshift simply measures distance. There are, however, peculiar velocities associated with the growth of structure which distort the distribution of galaxies as seen in redshift space. It is easy to understand the nature of the effect; consider an

overdense spherical region. According to (31.15) the Hubble rate within the region will be reduced, and consequently the volume of the region will appear smaller than it really is. Transverse dimensions are unaffected by the velocity, so the effect is that a overdense sphere will appear squashed along the line of sight.

We can make this more quantitative. Consider a single plane-wave density ripple with wave vector \mathbf{k} lying along the line of sight. Let the density contrast amplitude be δ_r (the subscript r denoting real-space). This density wave is generated by a sinusoidal displacement $\Delta r = \Delta r_0 \cos(kr)$. The density contrast is the derivative of the displacement: $\delta_r = d\Delta r/dr = -k\Delta r_0 \sin(kr)$, so the displacement and density contrast are related by $\Delta r = \delta_r/k$. Now the continuity equation tells us that the peculiar comoving velocity \mathbf{u} is related to the perturbation amplitude by $\nabla \cdot \mathbf{u} = -\dot{\delta}$, or $u = -\dot{\delta}/k$. The physical peculiar velocity is $\mathbf{v} = a\mathbf{u} = a\dot{\delta}/k$, giving rise to a physical displacement in redshift space $\Delta x = v/H$ and therefore a comoving displacement $\Delta r_z = \Delta x/a$, or $\Delta r_z = \dot{\delta}/Hk$. The ratio of the extra displacement in redshift space to the true displacement is

$$\frac{\Delta r_z}{\Delta r_0} = \frac{\dot{\delta}}{H\delta}. \quad (33.24)$$

In an Einstein - de Sitter cosmology, the density perturbation grows as $\delta \propto a$, so $\dot{\delta}/\delta = \dot{a}/a = H$. In this case, the extra displacement is just equal to the real displacement. This means that the amplitude of the density ripple in redshift space will appear twice as large as it really is. In a low density universe we saw that the perturbation to the Hubble rate is, for a given density contrast, lower. A simple, yet very good, approximation to the Ω dependence of the expansion rate perturbation is

$$\frac{\Delta H}{H} = -\frac{1}{3}\Omega^{-0.6}\frac{\Delta\rho}{\rho}. \quad (33.25)$$

It is not hard to generalize the analysis to a plane wave of arbitrary direction \mathbf{k} . The full effect is seen for waves along the line of sight whereas a wave lying transverse to the line of sight is unaffected. The final result for the amplitude $\delta_z(\mathbf{k})$ of the wave in redshift space is

$$\delta_z(\mathbf{k}) = (1 + \Omega^{0.6}\mu^2)\delta_r(\mathbf{k}) \quad (33.26)$$

where $\delta_r(\mathbf{k})$ is the amplitude of the wave in real-space and μ is the cosine of the angle between the wave vector and the line of sight. The power spectrum is the expectation value of the square of the Fourier transform, so the power-spectra in real and redshift-space are related by

$$P_z(\mathbf{k}) = (1 + \Omega^{0.6}\mu^2)^2 P_r(\mathbf{k}). \quad (33.27)$$

The redshift space distortion has been detected quite clearly in the 2df survey and appears to indicate $\Omega \simeq 0.3$, in accord with other estimates. The measurement of the effect is a little more difficult than the simple analysis would suggest. This is because non-linear structures produce an elongation of structures along the line-of-sight from the *finger of god* effect. Care must be taken so that this does not contaminate the linear theory squashing effect. The other main weakness of this probe is that we have assumed that galaxies are unbiased tracers of the mass. If galaxies are positively biased, the galaxy density contrast in real space is greater than that of the mass density (which is what drives the peculiar velocity). The result of a positive bias is to give an artificially low estimate of the density parameter. If we model the bias as a constant multiplier: $\delta_g(\mathbf{r}) = b\delta(\mathbf{r})$ then the redshift space power spectrum is

$$P_z(\mathbf{k}) = (1 + \beta\mu^2)^2 P_r(\mathbf{k}) \quad (33.28)$$

where

$$\beta = \frac{\Omega^{0.6}}{b}. \quad (33.29)$$

33.2.6 Angular Clustering Surveys

An alternative to redshift surveys is to measure the clustering from the fluctuations in the counts of galaxies projected onto the sky as revealed by a photometric survey. One can define a 2-dimensional angular correlation function $w(\theta)$, and angular power spectrum $P(\kappa)$ much as in 3-dimensions. As already mentioned, photometric surveys allow one to probe much larger volumes of space than available using redshift surveys, but the disadvantages are that the structures tend to get washed out in projection, and that non-Gaussianity becomes harder to discern. We will now show how the 2-dimensional statistics are related to their 3-dimensional analogs. In both cases we will assume that we are trying to measure structure on scales much less than the depth of the survey.

Limber's Equation

Limber showed how to relate the angular correlation function $w(\theta)$, to the 3-dimensional density correlation function $\xi(r)$. We define $w(\theta)$ such that the expectation value of the product of the occupation numbers for two small cells on the sky of solid angles $\delta\Omega_1$, $\delta\Omega_2$ and separation θ is

$$\langle n_1 n_2 \rangle = \bar{n}_\theta^2 \delta\Omega_1 \delta\Omega_2 (1 + w(\theta)). \quad (33.30)$$

Here \bar{n}_θ is the mean density of galaxies on the sky.

For simplicity we will assume that the universe can be treated as spatially flat on scales up to the size of the survey and we will work in Cartesian comoving coordinates $\mathbf{r} = (x, y, z)$. We will also assume that the survey solid angle is $\Omega \ll 1$, so we can erect a 2-dimensional Cartesian co-ordinate system (θ_x, θ_y) on the sky also, with $\mathbf{r} = (x, y, z) \simeq (z\theta_x, z\theta_y, z)$ to a very high precision. For an infinitesimal cell, and for a given density contrast field $\delta(\mathbf{r})$ the probability that a cell at position θ is occupied is

$$P(n_1) = \delta\Omega_1 \int dz \bar{n}(z) z^2 (1 + \delta(z\theta_x, z\theta_y, z)). \quad (33.31)$$

Taking the average over the ensemble of density fields, and using $\langle \delta \rangle = 0$, the expectation is

$$\langle n_1 \rangle = \delta\Omega_1 \int dz \bar{n}(z) z^2 \quad (33.32)$$

so the mean density on the sky is

$$\bar{n}_\theta = \frac{\langle n_1 \rangle}{\delta\Omega} = \int dz \bar{n}(z) z^2. \quad (33.33)$$

The probability that two cells, which we take to lie at the origin and at a distance θ along the θ_x axis, both be occupied is

$$P(n_1, n_2) = \langle n_1 n_2 \rangle = \delta\Omega_1 \delta\Omega_2 \int dz \bar{n}(z) z^2 \int dz' \bar{n}(z') z'^2 (1 + \langle \delta(0, 0, z) \delta(\theta z', 0, z') \rangle) \quad (33.34)$$

From the definition of $w(\theta)$ (33.30) we obtain

$$w(\theta) = \frac{\int dz \int dz' \bar{n}(z) z^2 \bar{n}(z') z'^2 \xi\left(\sqrt{\theta^2 z'^2 + (z - z')^2}\right)}{[\int dz \bar{n}(z) z^2]^2}. \quad (33.35)$$

If we are measuring at a small angle $\theta \ll 1$, and the spatial correlation function is a rapidly falling function, as seems to be the case, then the spatial correlation function will become very small for $(z - z') \gtrsim z\theta$. In this limit, making the dz integration over the correlation function is a lot like integrating over a δ -function. We can replace the slowly varying function $\bar{n}(z) z^2$ by its value at $z = z'$ and, changing integration variable we have

$$w(\theta) = \frac{\int dz' \bar{n}(z')^2 z'^4 \int dz \xi(\sqrt{\theta^2 z'^2 + z^2})}{[\int dz \bar{n}(z) z^2]^2}. \quad (33.36)$$

This is known as *Limber's equation*.

Angular Power Spectrum

In the same approximation, one can show that the angular power spectrum

$$P_\theta(\boldsymbol{\kappa}) = \int d^2\theta w(\theta) e^{i\boldsymbol{\kappa}\cdot\boldsymbol{\theta}} \quad (33.37)$$

is related to the 3-dimensional power spectrum by

$$P_\theta(\boldsymbol{\kappa}) = \int dz \bar{n}(z)^2 z^2 P(\boldsymbol{\kappa}/z). \quad (33.38)$$

To understand this one can think of the total projected galaxy counts as the superposition of a set of slabs. If we measure the transform of the counts at some wave-vector $\boldsymbol{\kappa}$ then we will see contributions from all 3-dimensional plane waves which project to the appropriate angular frequency; ie. they must have perpendicular wave-vector $k_\perp = \kappa/z$. If the slabs are each thick compared to the wavelength then only waves with \mathbf{k} very nearly perpendicular to the line of sight will contribute appreciably; they must have $k_\parallel \lesssim 1/\Delta r$, with Δr the slab thickness. This means that the power spectrum at wave-number $\boldsymbol{\kappa}$ can only feel a contribution from modes with $\mathbf{k} \simeq \boldsymbol{\kappa}/z$. Again, if Δz is large compared to the scale of clustering we can take the different slabs to be effectively uncorrelated so we can just add the power from all the slabs. Letting the sum $\sum \Delta r \dots \rightarrow \int dr \dots$ we obtain the result above.

Results

Either (33.36) or (33.37) can be used to predict the two-dimensional statistics from the 3-dimensional 2-point function or power spectrum. The simple analysis here can readily be generalized to non-flat spatial geometry, and one can also include evolution of $\xi(r)$ or $P(k)$. The relation is particularly simple when the correlation function is a power-law $\xi(r) \propto r^{-\gamma}$, for which we readily find $w(\theta) \propto \theta^{1-\gamma}$ or $w(\theta) \propto \theta^{-0.8}$ if $\gamma \simeq 1.8$. In Fourier space, the angular power spectrum has the same spatial index as the 3-dimensional power spectrum. At bright magnitudes the observations agree with this prediction, but there seems to be some flattening of the slope at faint magnitudes (corresponding to galaxy redshifts on the order of unity).

Of great interest is the form of the 2-point functions at large scales — the small-scale behaviour being relatively well understood — but great care must be taken to correctly determine $\bar{n}(z)$ since small-scale fluctuations arising nearby can masquerade as large-scale fluctuations at greater distance.

Statistical Uncertainty

Great care is needed in interpreting estimates of the angular correlation function since the correlations between estimates at different angular scales are highly correlated. To see why, it is again easiest to consider the power spectrum. As always, the angular power spectrum has a speckly form, with coherence scale, or speckle size, $\Delta\kappa \sim 1/\Theta$, with Θ the size of the survey. The expectation value of the power-spectrum is the true power, but there are relative fluctuations of order unity within each coherence patch. The error in $\hat{w}(\theta)$, which we will denote by Δw is driven by the fluctuations $\Delta P(\kappa)$ about the mean value. If we compute the variance we can model the estimated power as a set of discrete independent values with mean value $\langle \hat{P} \rangle = P(\kappa)$ and variance $\langle (\Delta P)^2 \rangle = P^2(\kappa)$. If we compute the error variance in $\hat{w}(\boldsymbol{\theta}) = (2\pi)^{-2} \int d^2\kappa \hat{P}(\kappa) e^{i\boldsymbol{\kappa}\cdot\boldsymbol{\theta}}$ we find

$$\langle (\Delta w)^2 \rangle \sim (\Delta\kappa)^4 \sum_i \sum_j \Delta P_i \Delta P_j \sim (\Delta\kappa)^4 \sum_i \langle (\Delta P)^2 \rangle \sim (\Delta\kappa)^2 \int d^2\kappa P^2(\kappa). \quad (33.39)$$

Now if $P(\kappa) \propto k^{-1.2}$, as seems to be the case, then this integral is ‘infra-red divergent’; it is dominated by few lowest frequency speckles. The error $\Delta w(\theta)$ produced by these dominant speckles has very strong long range correlations.

Performing the same analysis for the 3-dimensional correlation function we find

$$\langle (\Delta\xi)^2 \rangle \sim (\Delta k)^6 \sum_i \sum_j \Delta P_i \Delta P_j \sim (\Delta k)^3 \sum_i \langle (\Delta P)^2 \rangle \sim (\Delta k)^3 \int d^3k P^2(k). \quad (33.40)$$

With $P(k) \propto k^{-1.2}$ this integral is ‘ultra-violent’ divergent; more realistically, what this means is that if we average the correlation function over bins of width Δr then the integral will be cut-off at $k_{\max} \sim 1/\Delta r$. There is still a long-range component to the errors in $\hat{\xi}$, but it is smaller than the bin-to-bin variance. In either case, the best way to understand the statistical uncertainty is *via* power spectrum analysis, and most recent studies use this technique.

33.3 Bulk-Flows

Deviations from pure Hubble expansion, or *bulk-flows*, provide another probe of large scale structure. This probe exploits the linearized continuity equation $\dot{\delta} = -\nabla \cdot \mathbf{u}$ which relates the comoving peculiar velocity to the growth of the density contrast $\delta(\mathbf{r})$. Bulk-flow studies therefore provide a direct probe of the *mass* fluctuations, and are independent of biasing. In principle, these studies can be used to determine the power-spectrum of the mass fluctuations. However, as we shall describe, bulk flow measurements can only be made for nearby galaxies, so the volume that can be probed is quite small and the power spectrum estimates therefore have large sample variance. Instead, what is more usually done is to compare the flows and the galaxy density contrast, and thereby try to determine the fluctuation growth rate.

33.3.1 Measuring Bulk-Flows

The velocity in question is the peculiar velocity; the deviation from pure Hubble flow. One very useful datum comes from the *CMB dipole anisotropy*, which measures the Earth’s peculiar velocity with respect to the frame in which the CMB is isotropic. Since the dipole anisotropy is much larger, by a factor ~ 30 , than the intrinsic anisotropy, this is effectively just our peculiar velocity. This tells us that the earth is moving at about 300km/s. We are moving in roughly the opposite direction around the Milky Way at a speed of about 220km/s, so taking the vector sum we find that the galaxy is moving at about 500km/s relative to the cosmic rest frame.

In principle, one can measure the line-of-sight peculiar velocities for clusters of galaxies from the *kinematic Sunyaev Zel’dovich effect*, but there is little useful information as yet.

What has proved more useful is to measure the relative velocities of other galaxies in our neighbourhood, and thereby extend the measurement of the Milky Way motion to larger scales. Measuring bulk-flows is conceptually straightforward; we simply need to determine the distance r to the galaxy in question, we then take the recession velocity of the galaxy and correct this for the motion of the Milky Way (to obtain the recession velocity v_{CMB} that would be measured by an observer at our location who happens to see no CMD dipole moment). The component of the galaxies peculiar velocity along the line of sight is

$$v_{\text{pec}} = v_{\text{CMB}} - Hr. \quad (33.41)$$

The tricky part is determining accurate distances. If galaxies were ‘standard candles’ this would be straightforward, but galaxies have a wide range of intrinsic luminosities. There are, however, strong correlations between the intrinsic luminosity and distance independent measureable quantities such as the rotation velocity or velocity dispersion. This is not surprising; more massive galaxies have larger rotation velocities and might be expected to be more luminous. Spiral galaxies, for instance, obey a strong correlation between intrinsic luminosity and rotation velocity of the form $L \propto v_c^\alpha$, with $\alpha \simeq 4$ (the best fitting slope depends on the passband in which the flux is measured). This is known as the *Tully-Fisher relation* (see figure 33.1). The rotation velocity provides an estimate of the intrinsic luminosity and comparing with the measured flux then gives the distance to the galaxy. There is a similar relation for elliptical galaxies between the luminosity and the velocity dispersion, known as the *Faber-Jackson relation*. It was subsequently realized that a more accurate

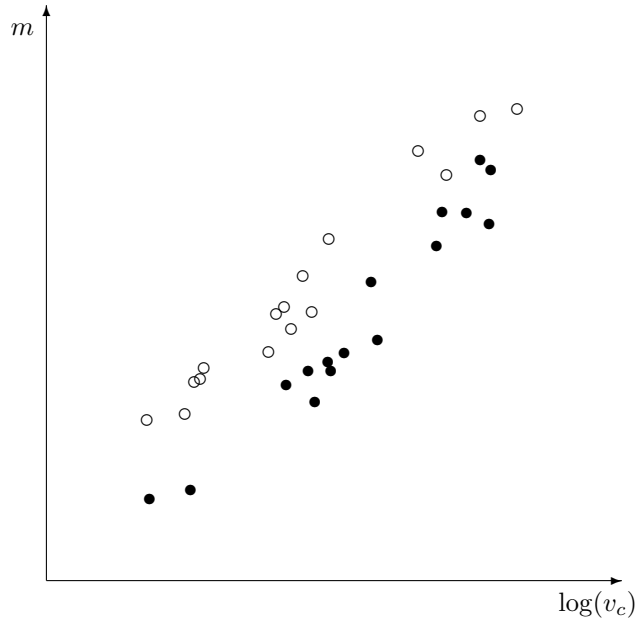


Figure 33.1: Illustration of the Tully-Fisher relation. Open and closed dots denote measurements of the logarithm of circular velocity and apparent magnitude for galaxies in two regions of space, say in two different clusters. There is a tight relation between m and $\log(v_c)$ in both cases, but with an offset. The most straightforward interpretation of this is that there is a universal relation between intrinsic luminosity and v_c but that the clusters are at different distances. This type of measurement allows one to determine relative distances to distant galaxies.

prediction for the intrinsic luminosity can be given from measurements of both the velocity dispersion and the surface brightness; another distance independent observable. This relation is known as the *fundamental plane*. Other methods for determining distances to galaxies include supernovae and *surface brightness fluctuation*.

A common property of all such distance estimates is that they give distances with a constant, or nearly constant, fractional error. The most precise methods give fractional distance errors at about the 10% level. If we assume that our motion of $\sim 500\text{km/s}$ is not atypical — as seems to be the case — this says that, for a single galaxy, the peculiar velocity can only be determined if the recession velocity is $v \ll 5000\text{km/s}$. One can do somewhat better by averaging together distances for a collection of galaxies in some region of space, which are assumed to share a common peculiar velocity, but the range of such methods is still quite limited.

The earliest application of such measurements were to the *local supercluster* (LSC). This is a system whose center lies $\sim 10\text{Mpc}$ from the Milky Way and which appears to have a density contrast in galaxies $\Delta n/n \sim 2$ within our radius. Measuring distances to, and recession velocities of, galaxies in this system indicates that the Hubble expansion is retarded by about 30%. This indicates a fairly low Ω . Only part of the $\sim 500\text{km/s}$ motion of the Milky Way can be attributed to the gravitational pull of the local supercluster. More accurate inspection of the flows within the LSC revealed a tidal shear, indicating the presence of some external mass. Once a deeper sample of elliptical galaxies became available it became apparent that there are several superclusters in our vicinity that produce a rather complicated flow pattern.

These measurements have been combined with redshift surveys to determine Ω (or more generally β). There are two approaches that have been used. The first is to use the galaxy density field to compute the acceleration field and thereby predict the peculiar velocity. This is a non-local method, since the velocity of a galaxy may depend on quite distant mass concentrations. It is then important that the redshift survey have nearly full sky coverage and be sufficiently deep to encompass all of the important mass fluctuations. The second approach is to compute the divergence of the velocity

field and then compare this with the density. This is a local comparison. While it would seem that computing $\nabla \cdot \mathbf{u}$ would require all 3 components of the velocity field, this is not the case. The reason for this is that in linear perturbation theory the flow is a potential field. This means that one can determine the velocity potential simply by integrating the line of sight component of the velocity, and one can thereby reconstruct all 3 components of the velocity field from measurements of only one. Application of these methods, mostly using the redshift surveys derived from the IRAS satellite observations, indicate fairly high density parameter. How this can be reconciled with the consensus for low values is not yet clear.

33.4 Microwave Background Anisotropies

A third probe of linear structures are *cosmic microwave background anisotropies*. This is rather different from galaxy clustering and bulk flows in that the anisotropy is generated at early times. Small angle anisotropy, for instance, probes the state of the universe at the redshift of recombination $z_{\text{dec}} \simeq 1000$.

33.4.1 Recombination and the Cosmic Photosphere

First we need to understand where the photons we are seeing originated. In the standard cosmology the universe was highly ionized prior to z_{dec} and then rather rapidly became neutral. Detailed calculations show that the ‘visibility function’ — this gives the distribution over redshift of last scattering for CMB photons — is correspondingly narrow, with width $\Delta z/z \sim 1/10$. In an Einstein - de Sitter cosmology, the comoving distance ω is related to redshift by

$$\omega = 1 - \frac{1}{\sqrt{1+z}}. \quad (33.42)$$

This means that the horizon (the surface of infinite redshift) is at $\omega = 1$, whereas the surface $z = z_{\text{dec}}$ is at $\omega_{\text{dec}} \simeq 1 - 1/\sqrt{1000} \simeq 0.97$. The *cosmic photosphere* is then a rather narrow fuzzy shell quite close to the horizon. If the universe is spatially flat then the angle subtended by the horizon size at decoupling is $\theta_{H,\text{dec}} \simeq 1/\sqrt{z_{\text{dec}}}$ or about 2 degrees. If the universe is open, this angular scale is reduced.

33.4.2 Large-Angle Anisotropies

Consider first the anisotropy generated by perturbations larger than the horizon size at decoupling. The situation is sketched in figure 33.2. The primary effect driving large-angle anisotropy is the so-called *Sachs-Wolfe effect*. Photons which emerge from an over-dense (under-dense) region suffer what is effectively a gravitational redshift (blueshift). However, for subtle reasons, the fractional photon energy change — which is equal to the fractional change in temperature — is one third of the Newtonian potential perturbation at the point of emission. The Sachs-Wolfe temperature anisotropy is

$$\frac{\Delta T}{T} \sim \delta\phi_{\text{em}}. \quad (33.43)$$

The temperature anisotropy is therefore on the order of the density perturbation amplitude at horizon crossing.

In addition to the temperature anisotropy induced as the photons ‘climb out’ of the potential well where they originate, there is an additional source of temperature anisotropy generated as the photons pass through intervening inhomogeneity (photon C in figure 33.2). This effect was first calculated by Rees and Sciama. This effect, however, is sub-dominant. First, in a flat universe, and in linear theory, the effect vanishes. In an open universe there is a non-vanishing effect, but the amplitude is on the order $\Delta T/T \sim HR\delta\phi$, where R is the size of the perturbation; this is smaller than the Sachs-Wolfe effect by the factor HR which is the size of the perturbation in units of the horizon size. Now the effect of multiple perturbations along the line of sight will add in quadrature,

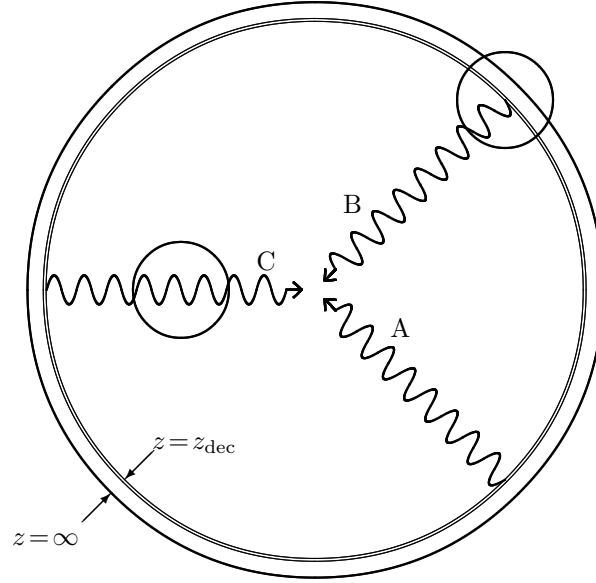


Figure 33.2: Schematic illustration of the generation of large angle CMB anisotropy. Photon A arrives from an unperturbed region of the universe. Photon B emerges from an over-dense region. It suffers the ‘Sachs-Wolfe’ effect — effectively a gravitational redshift — and has slightly lower energy than photon A when it reaches the observer. Photon C passes through an over-dense region on its way from the last scattering surface, and suffers the ‘Rees-Sciama’ effect.

resulting in a net anisotropy $\Delta T/T \sim \sqrt{HR}\delta\phi$, but this is still small compared to the Sachs-Wolfe effect.

The large-scale temperature anisotropy effectively provides a map of the Newtonian potential variation on the last scattering surface. The prediction of inflation (and also cosmic string models) is that the temperature fluctuation field should take the scale-invariant flicker-noise form corresponding to spectral index $n = 1$. This is just what was observed by the COBE satellite.

33.4.3 Small-Angle Anisotropies

Small-scale anisotropies are more complicated. As discussed in chapter 31, prior to de-coupling, and on scales small compared to the sound horizon, the baryon-photon fluid undergoes acoustic oscillations. In such sound waves, there are adiabatic variations of the temperature $\Delta T/T \sim \Delta\rho/\rho$ and there are also peculiar velocities $v \sim c_s \Delta\rho/\rho$. Both of these give rise to temperature anisotropy. Detailed calculation of small-scale anisotropy is quite complicated and must be performed numerically. The net result of such calculations is the angular power spectrum for the fluctuations. This is usually stated in terms of the mean square Legendre polynomial coefficient C_l^2 . However, since these effects appear at small angle, one is free to approximate the sky as effectively flat, and one can then express the temperature variance in terms of the power-spectrum. The result is a spectrum of small-angle anisotropies with bumps and wiggles extending from the horizon scale down to the damping cut-off. The amplitude of the wiggles in the power spectrum are larger the larger is the baryon content.

33.4.4 Polarization of the CMB

In addition to the temperature anisotropy $\Delta T/T$, the CMB is predicted to display polarization. Polarization of the CMB arises by virtue of the anisotropic nature of Thomson scattering; if an electron is illuminated with radiation which is anisotropic then the radiation it scatters will be polarized. For example, if we lie along the z -axis and observe an electron at the spatial origin which is being illuminated by a lamp sited out along the y -axis then the radiation we see will be linearly

polarized in the x -direction (see §10.7). In general, the degree of polarization is proportional to the quadrupole moment of the incident radiation. Measurement of the CMB polarization therefore provide us with a kind of ‘remote-sensing’ of the anisotropy of the radiation field as it was at the time of last scattering.

33.5 Weak Lensing

Another probe of large-scale structure is *weak gravitational lensing*. Light rays propagating to us through the inhomogeneous universe get tugged from side to side by mass concentrations. The deflection of a light ray that passes a point mass M at impact parameter b is

$$\theta_{\text{def}} = \frac{4GM}{c^2 b}. \quad (33.44)$$

This is just twice what Newtonian theory would give for the deflection of a test particle moving at $v = c$. In this picture we are imagining the radiation to be test particles being pulled by a gravitational acceleration. There is another useful way to look at this using wave-optics; the inhomogeneity of the mass distribution causes space-time to become curved. As discussed earlier, the space in an over-dense region is positively curved, as illustrated in figure 28.7. This means that light rays propagating through the over-density have to propagate a slightly greater distance than they would in the absence of a the density perturbation. Consequently the wave-fronts get retarded slightly in passing through the over-density and this results in focusing of rays. There is still another way to picture the situation: The optical properties of a lumpy universe are, in fact, essentially identical to that of a block of glass of inhomogeneous density where the refractive index is

$$n(\mathbf{r}) = (1 - 2\phi(\mathbf{r})/c^2) \quad (33.45)$$

with $\phi(\mathbf{r})$ the Newtonian gravitational potential. In an over-dense region, ϕ is negative, so n is slightly greater than unity. In this picture we think of space as being flat, but that the speed of light is slightly retarded in the over-dense region. All three of the above pictures give identical results.

The gravitational potential of a bound structure is on the order of $\delta\phi \sim \sigma_v^2$, where σ_v is the velocity dispersion or circular velocity. Now the velocity dispersion for a massive cluster of galaxies, for example, is $\sigma_v \simeq 1000\text{km/s}$, or about 0.003 times c . One such object can cause a deflection of on the order of $20''$, and multiple objects along the line of sight would give random deflections adding in quadrature to give still larger deflection. Images of distant objects are therefore shifted from their ‘true’ positions (i.e. the positions they would have if we could somehow switch off the gravity perturbation). Unfortunately, this deflection is not easily measurable, since we do not know the true positions. Instead, the effect exploited in weak lensing is the differential deflection — i.e. the fact that the deflection suffered by the light from one side of a distant galaxy is slightly different than the deflection for the other side. This causes a systematic distortion, or ‘shearing’, of the shapes of the distant galaxies. The effect is similar to the distortion of distant objects seen in a ‘mirage’, though is a much weaker effect.

We can estimate the size of the effect as follows: Consider a mass concentration of size R and mass M at a distance D_{ol} from the observer as shown schematically in figure 33.3. Now consider a thin conical bundle of rays with opening angle θ_0 emerging from the observer and propagating back to some distant ‘source plane’. In the absence of the lens this bundle would intersect the source plane in a circle of radius $l = D_{os}\theta_0$. Now the lens will introduce some deflection $\theta_{\text{def}} \sim GM/c^2 b$, and there will also generally be some change in the deflection across the bundle of $\Delta\theta_{\text{def}} \sim GM\delta b/c^2 b^2$. This relative deflection will cause the initially circular bundle of rays to become elliptical. The length of the ellipse in the radial direction will be $l' = l + D_{ls}\Delta\theta_{\text{def}}$. The fractional stretching of the ellipse, or ‘image shear’, is

$$\gamma = \frac{l'}{l} - 1 = \frac{D_{ls}}{D_{os}} \frac{\Delta\theta_{\text{def}}}{\theta_0}. \quad (33.46)$$

Now we can also write the differential deflection as $\Delta\theta_{\text{def}} = \theta_0 d\theta_{\text{def}}/d\theta$, where $d\theta_{\text{def}}/d\theta$ is the *distortion tensor*; it gives the rate of change of deflection angle with position on the sky. The image

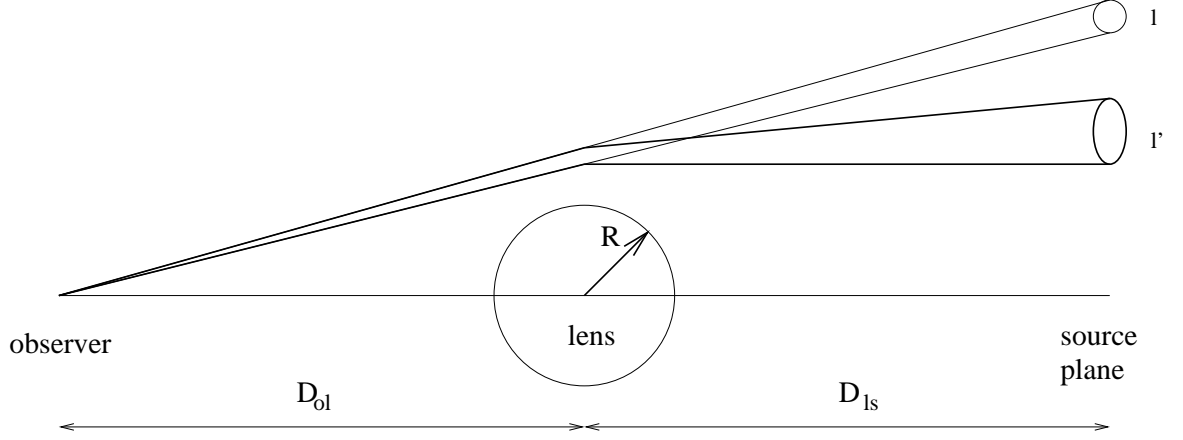


Figure 33.3: Schematic illustration of weak lensing. In the absence of the deflecting lens, a conical bundle of rays emerging from the observer will intercept the source plane in a circle. In the presence of the lens, the bundle will intercept the source plane in an ellipse. An object of this ellipticity on the source plane will therefore appear circular. Conversely, a circular object will appear elliptical, but with an ellipticity of the opposite ‘sign’ (i.e. it will appear stretched along the tangential direction in this example). As shown in the main text, the fractional stretching of the image, also known as the ‘image-shear’, is $\gamma = l'/l - 1 = (D_{ls}/D_{os})\partial\theta_{\text{def}}/\partial\theta$.

shear is then

$$\gamma = \frac{D_{ls}}{D_{os}} \frac{d\theta_{\text{def}}}{d\theta}. \quad (33.47)$$

Now consider an over-dense lens of size R and density contrast $\delta\rho/\rho$. The angular size of the lens is $\theta \sim R/D_{ol}$. The deflection angle is

$$\theta_{\text{def}} \sim \frac{G\delta M}{Rc^2} \sim \frac{G\delta\rho R^2}{c^2}. \quad (33.48)$$

The deflection angle will vary smoothly with impact parameter, so the distortion is

$$\frac{d\theta_{\text{def}}}{d\theta} \sim \frac{\theta_{\text{def}}}{\theta} \sim \frac{G\delta\rho R^2/c^2}{R/D_{ol}} \sim \frac{H^2 D_{ol} R}{c^2} \frac{\delta\rho}{\rho} \quad (33.49)$$

where we have used $H^2 \sim G\rho$. The image shear is therefore

$$\gamma \sim \frac{H^2}{c^2} \frac{D_{ol} D_{ls}}{D_{os}} R \frac{\delta\rho}{\rho}. \quad (33.50)$$

Note that the strength of the effect scales as the product of the density and the size of the object; i.e. it is proportional to the surface density of the lens.

This is the effect due to a single object. This may be relevant for highly non-linear objects such as clusters of galaxies, where the space filling factor is small and the probability that a line of sight intercepts such an object is small. For large-scale structures with $\delta\rho/\rho \lesssim 1$ the filling factor is of order unity, and the net effect will be the superposition of a large number of structures along the line of sight. In fact, the shear is the integral of the tidal field along the line of sight. Each structure will give a contribution to the shear with a random direction and strength. This means that the net shear variance $\langle\gamma^2\rangle$ will be the sum of the individual shear variances, or, equivalently, that the net effect will be larger than that from a single object by roughly \sqrt{N} where $N \sim D_{os}/R$ is the number of structures. The factor $D_{ol}D_{ls}$ means that we get a small effect from structures which are very close to either the source or the observer, so, to get a crude estimate of the effect we can assume

that all of the distances are of the same order of magnitude. This gives the prediction for the root mean square shear:

$$\langle \gamma^2 \rangle^{1/2} \sim \frac{H^2 D_{\text{os}}^{3/2} R^{1/2}}{c^2} \left\langle \left(\frac{\delta \rho}{\rho} \right)^2 \right\rangle^{1/2}. \quad (33.51)$$

This tells us that the strength of the effect increases as the $3/2$ power of the distance to the source; therefore to detect the effect one wants to use sources at cosmological distance (this also gives a large number of background galaxies, with also helps). For sources at $z \sim 1$, the distance is $D \sim c/H$, so the strength of the effect is then

$$\langle \gamma^2 \rangle^{1/2} \sim \sqrt{\frac{R}{D}} \left\langle \left(\frac{\delta \rho}{\rho} \right)^2 \right\rangle^{1/2}. \quad (33.52)$$

This would suggest that, for super-cluster scale structures with $R \sim 10\text{Mpc}$ and $\delta\rho/\rho \sim 1$, the root mean square shear would be on the order of 5%. More careful estimates — putting in factors like 4π and geometric factors properly — gives a prediction for shear of order 1% on scales of one degree.

This is a very small effect. A 1% shear means that a circular object will appear elliptical, with major axis about 1% larger than the minor axis. However, galaxies are already elliptical, with root mean square ellipticity of about 30%, so the effect cannot be detected for a single object. What makes the effect measureable is that the shear is *spatially coherent*. The effect from super-cluster and larger scale structure will be coherent over large angular scales. Now the number of background galaxies — the ‘cosmic wallpaper’ — becomes very large; one can readily detect $\gtrsim 10^5$ galaxies per square degree. By looking for a statistical tendency for the galaxy position angles to be anisotropically distributed one can measure shears on the order of $\sim 0.3/\sqrt{N_{\text{gal}}}$, which is comfortably smaller than the prediction on degree scales. This is just the statistical error; to keep the systematic errors below this requires very careful analysis of the images. Current measurements find consistent results on scales of $\lesssim 10'$, and here is little data on larger scales as yet, but several large-scale surveys are being carried out, so the outlook is promising.

Chapter 34

Non-Linear Cosmological Structure

In chapter 31 we explored the evolution of small amplitude perturbations of otherwise homogeneous cosmological models. This provides an accurate description of the evolution of structure from very early times. On sufficiently large scales, the structure is still in the linear regime today, but small scale structures have reached the point where $\delta\rho/\rho \gtrsim 1$ and have gone *non-linear*.

When dealing with the development of non-linear structure we can usually neglect radiation pressure and assume that the structures are much smaller than the horizon scale, so a Newtonian treatment is valid. However, the equations of motion are still relatively complicated and it is hard to find exact solutions except in highly idealized models such as spherical or planar 1-dimensional collapse. One approach to non-linear structure growth is to attempt to evolve the initial conditions forward from the linear regime numerically using either *N-body simulations*, to evolve the collisionless Boltzmann equation, or *hydro-dynamical simulations* to evolve the Euler, energy and continuity equations. The former is adequate to describe the evolution of collisionless dark matter matter, but the latter is required if one also wants to treat the baryonic matter. Another possibility is to extend perturbation theory beyond linear order. This is an area where there has been much activity by theorists in recent years. These calculations typically assume a Gaussian initial density field, and then compute the emergence of non-Gaussianity, e.g. the skewness, or the kurtosis of the density distribution. Such results are limited to the ‘quasi-linear’ regime; i.e. density contrasts $\delta \lesssim 1$. This is a rather limited range of validity. Also, since most interest is in theories with ‘hierarchical’ initial fluctuation spectra, when one scale is just going non-linear, there are smaller scale structures which will be highly non-linear. Usually such calculations deal with this by assuming some smoothing of the initial δ -field, but the validity of this is questionable.

Here I shall describe a number of approximate methods and models that directly address the ‘quite-strongly non-linear’ regime. These models are typically quite idealized, but they are still useful as they provide insight into the way structure has evolved, and is evolving today.

34.1 Spherical Collapse Model

A simple model for the non-linear evolution of an initially positive density fluctuation is the ‘top-hat model’ in which there is an initial spherical over-density, with constant $\delta = \delta\rho/\rho$. We have already analyzed this in the linear regime. The non-linear evolution of such a perturbation is illustrated in figure 34.1.

The inner radius obeys

$$\dot{R}^2 = 2GM/R - 2E_0 \quad (34.1)$$

with ‘cycloidal’ solutions

$$R = \frac{GM}{2E_0}(1 - \cos(\eta)) \quad (34.2)$$

$$t = \frac{GM}{(2E_0)^{3/2}}(\eta - \sin(\eta)) \quad (34.3)$$

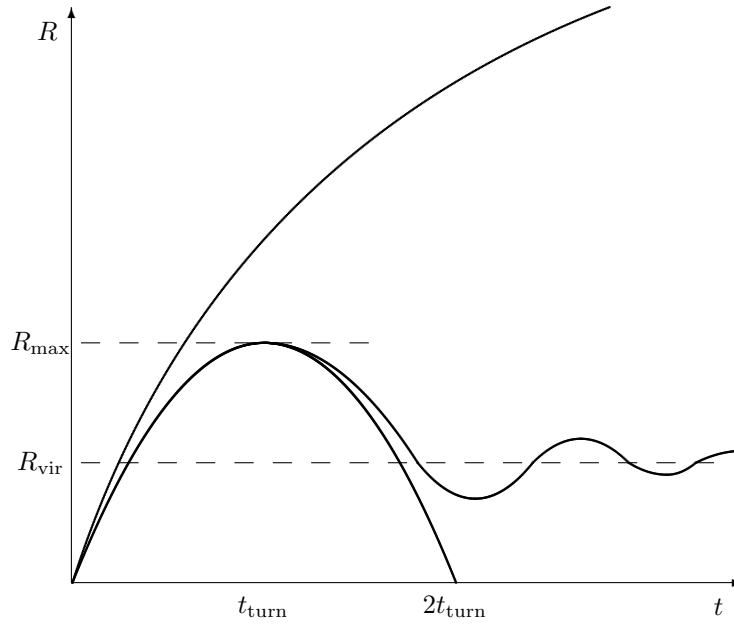


Figure 34.1: In the spherical ‘top-hat’ perturbation we excise some sphere of matter and replace it with a smaller concentric uniform density sphere. The upper line depicts the expansion of the exterior mass shell. The interior, assumed here to be gravitationally bound, behaves like part of a closed FRW model (lower curve). If the interior is precisely uniform it will collapse to a final singularity and form a black hole. However, for perturbations much smaller than the horizon scale, the specific gravitational binding energy at the point of maximum expansion is small, and a huge collapse factor is required to form a black-hole. More realistically, any initial irregularity or angular momentum will cause the collapse to ‘bounce’, and we expect the system to then settle down to an equilibrated, or virialized, state with final radius roughly half that of the sphere at the point of maximum expansion.

The inner radius peaks at the ‘turn-around time’ $t_{\text{turn}} = \pi GM(2E_0)^{-3/2}$ when the density is

$$\rho = \frac{3M}{4\pi R_{\text{max}}^3} = \frac{3\pi}{32Gt^2}. \quad (34.4)$$

Compare this with the density of an Einstein - de Sitter ($k = 0$, $\Omega = 1$) background

$$\bar{\rho} = \frac{1}{6\pi t^2}. \quad (34.5)$$

In such a background, the over-dense perturbation will turn around with density contrast

$$\left(\frac{\rho}{\bar{\rho}}\right)_{\text{turn}} = \frac{9\pi^2}{16} \simeq 5.55. \quad (34.6)$$

A perfectly spherical over-density would collapse to infinite density at $2t_{\text{turn}}$ and would form a black hole. In the more realistic case we would expect a collapsing ‘blob’ to become increasingly aspherical and *virialize* after contracting by about a factor 2 in radius. What we mean by virialization is reaching a state where the radius is no longer contracting as in the collapse phase. While only a rough guide to the real situation, this suggests that the system will virialize with a density ~ 8 times larger than at turnaround. Now the background cosmology is expanding with scale-factor $a \propto t^{2/3}$, so in the interval $t_{\text{turn}} < t < 2t_{\text{turn}}$ the background will have expanded by a factor $2^{2/3}$ and will therefore have decreased in density by a factor 4, giving a density contrast at virialization of

$$\left(\frac{\rho}{\bar{\rho}}\right)_{\text{virial}} = 18\pi^2 \simeq 180. \quad (34.7)$$

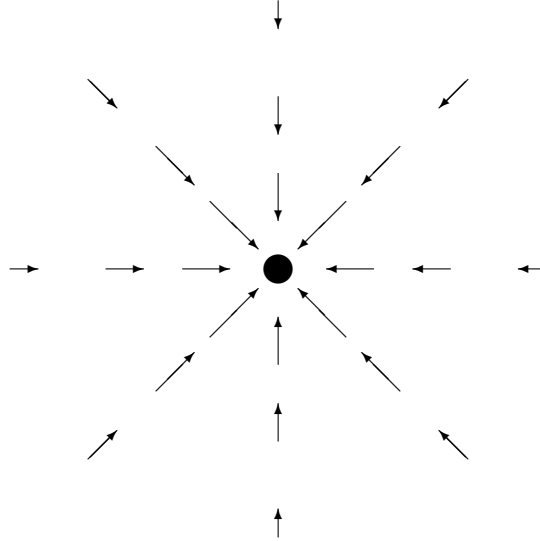


Figure 34.2: Illustration of the kind of divergence free flow pattern induced by a point mass. The peculiar velocity falls off as $\delta v \propto 1/R^2$, with the consequence that the flux of matter through any a shell of radius R is independent of R . For $R > 0$ the density remains unperturbed, but there is a net accumulation of mass at the origin.

This is for a perturbation of an Einstein - de Sitter background, which may not be realistic at very late times, but the analysis is readily generalized to open models, since the only thing the interior ‘knows’ about the background cosmology is the time since the big-bang.

The perturbation conserves energy, so the final virial velocity σ_v is related to the initial perturbation to the Newtonian potential perturbation $\delta\phi$ by

$$\frac{\sigma_v^2}{c^2} \sim \delta\phi. \quad (34.8)$$

Perhaps more interestingly, we can estimate the velocity dispersion for a recently virialized object of a given size, or *vice versa*. For an object with an over-density of 180, the circular velocity is

$$v_{\text{circ}}^2 = \frac{GM}{R} \simeq 180 \times \frac{4}{3}\pi G\bar{\rho}R^2 \simeq 90H^2R^2. \quad (34.9)$$

The line-of-sight velocity dispersion is $\sigma_v^2 \simeq v_{\text{circ}}^2/2$, so the velocity dispersion and radius (at density contrast 180) is

$$R_{180} \sim \frac{1}{7} \frac{\sigma_v}{H}. \quad (34.10)$$

For a rich cluster of galaxies like the Coma cluster, which has $\sigma_v \simeq 1000\text{km/s}$, this gives $R_{180} \simeq 1.5h^{-1}\text{Mpc}$. While clusters do not have sharp edges — there being matter in-falling at greater distances and denser material in the center which collapsed in the past — it is gratifying that this estimate of the size of a cluster agrees very nicely with the size that George Abell assigned to such objects.

34.2 Gunn-Gott Spherical Accretion Model

Another very illuminating model is that of Gunn and Gott (19??) who considered what happens if one introduces a point-like ‘seed’ of mass M_0 into an otherwise uniform Einstein -de Sitter universe.

First consider the linear theory. At large radii, the mass will induce a peculiar acceleration at physical distance R of

$$g = \frac{GM_0}{R^2}. \quad (34.11)$$

Acting over a Hubble time $t \sim 1/H$ this will generate a peculiar infall velocity

$$\delta v \sim gt \sim \frac{GM_0}{HR^2}. \quad (34.12)$$

This kind of $\delta v \propto 1/R^2$ flow (see figure 34.2) is ‘divergence free’, so there is no change in the density at large distances (think of two concentric comoving shells; the flux of matter across a surface is the velocity times the area and is independent of radius, so there is no build up except at the center). The amount of mass convected across a shell in one Hubble time is $\delta M \sim \bar{\rho} R^2 \delta v t$ which, with (34.12) and $H^2 = 8\pi G\bar{\rho}/3$, gives $\delta M \sim M_0$. Thus the seed induces, after one expansion time, a growing mode density perturbation $(\delta\rho/\rho)_i \sim M_0/M$, and this subsequently grows with time as $\delta\rho/\rho = \delta M(t)/M \propto a(t)$. This is assuming, for simplicity, an Einstein - de Sitter background.

The amount of mass accumulated in the center is therefore

$$\delta M(t) \sim M_0 \frac{a}{a_i} \propto a(t). \quad (34.13)$$

This mass represents a density contrast of order unity at a physical radius R such that $\delta M \sim \bar{\rho} R^3$, and slightly inside will lie the turnaround radius

$$R_{\text{turn}} \sim (\delta M/\bar{\rho})^{1/3} \propto a^{4/3}. \quad (34.14)$$

As time goes on, progressively larger shells will turn around, collapse and virialize in some complicated way with shell crossing etc. However, we may reasonably expect that the final specific binding energy of a shell of a certain mass will be equal, modulo some factor of order unity, to its initial specific binding energy $\delta\phi$. Now the initial binding energy is a power law in radius:

$$\delta\phi \sim GM_0/R \propto 1/R \propto M^{-1/3}, \quad (34.15)$$

whereas the final binding energy as a function of the final radius R_f is

$$\delta\phi \sim GM(R_f)/R_f \quad (34.16)$$

where $M(R_f)$ is the mass within radius R_f . Equating these gives the scaling law

$$M(R_f) \propto R_f^{3/4}. \quad (34.17)$$

If the mass within radius R_f is a power law in R_f then so also is the density:

$$\rho(R_f) \sim M(R_f)/R_f^3 \propto R_f^{-9/4}. \quad (34.18)$$

This analysis then tells us that the virialized system should have a power law density profile. What is interesting about this result is that it is very close to the $\rho(R) \sim R^{-2}$ density run for a flat rotation curve halo, and also similar to the profile of clusters of galaxies, which are also often modeled as ‘isothermal spheres’.

34.3 The Zel’dovich Approximation

In linear theory, and for growing perturbations in an Einstein - de Sitter model, particles move with peculiar velocity

$$\mathbf{v}(\mathbf{r}, t) = (t/t_0)^{1/3} \mathbf{v}_0(\mathbf{r}) \quad (34.19)$$

where now \mathbf{r} is a comoving spatial coordinate and \mathbf{v}_0 is the peculiar velocity at some initial time. This says that the peculiar velocity field just grows with time at the same rate $\mathbf{v} \propto t^{1/3}$ at all points in space.

The physical displacement of a particle in time dt is $d\mathbf{x} = \mathbf{v}dt$, so the comoving displacement is

$$d\mathbf{r} = \frac{d\mathbf{x}}{a} = \frac{\mathbf{v}}{a}dt \equiv \mathbf{u}dt \quad (34.20)$$

where the comoving peculiar velocity is $\mathbf{u} = \mathbf{v}/a$. The rate of change of comoving position with scale factor a is then

$$\frac{d\mathbf{r}}{da} = \frac{\mathbf{v}}{a} \frac{dt}{da}, \quad (34.21)$$

but with $\mathbf{v} \propto t^{1/3}$ and $a \propto t^{2/3}$, so $da/dt \propto t^{-1/3}$, this says that

$$\frac{d\mathbf{r}}{da} \propto t^0. \quad (34.22)$$

Therefore, if we define a new ‘time’ $\tau \propto a$, then the particles move ballistically in comoving coordinate space: $d\mathbf{r}/d\tau = \text{constant}$. Zel’dovich’s approximation is to assume that this ballistic motion continues into the non-linear regime.

The result is a *Lagrangian mapping* resulting in formation of *caustics*, or surfaces of infinite density. This is very analogous to the formation of caustics on the swimming pool floor, which we explored in our study of geometric optics in chapter 8. There the horizontal deflection of the rays — a 2-dimensional vector displacement — increases linearly with distance from the surface, and here the 3-dimensional comoving displacement increases linearly with ‘time’ τ . We can write the actual comoving position or *Eulerian coordinate* \mathbf{x} as a function of the initial or *Lagrangian coordinate* \mathbf{r} as

$$\mathbf{x}(\mathbf{r}) = \mathbf{r} + \tau \mathbf{U}(\mathbf{r}) \quad (34.23)$$

where \mathbf{U} is a suitably scaled version of \mathbf{u} .

Until caustics form, the density is

$$\rho \propto \left| \frac{\partial x_i}{\partial r_j} \right|^{-1} \quad (34.24)$$

where $|\partial \mathbf{x}/\partial \mathbf{r}|$ is the Jacobian of the transformation from Lagrangian to Eulerian coordinate. This is just conservation of mass: $dM = \rho_L d^3r = \rho_E d^3x$, with ρ_E and ρ_L the densities in Eulerian and Lagrangian space respectively. Now from (34.23), $\partial x_i/\partial r_j = \delta_{ij} + \tau \partial U_i/\partial r_j$, so we can also write the density as

$$\rho \propto \frac{1}{(1 + \tau \lambda_1)(1 + \tau \lambda_2)(1 + \tau \lambda_3)} \quad (34.25)$$

where the λ_i are the eigenvalues of the *deformation tensor* $\Phi_{ij} = \partial U_i/\partial r_j$.

If there is a negative eigenvalue $\lambda_1 < \lambda_2, \lambda_3$, then the density of a small comoving volume of matter will become infinite with collapse along the appropriate *principle axis* when $\tau = -1/\lambda_1$. *Pancakes* — or perhaps we should call them *blinis* — form with a multi-stream region sandwiched between the caustic surfaces. These pancakes grow rapidly and intersect to form a cellular network of walls or pancakes intersecting in lines with the matter in these lines draining into the nodes where all three of the eigenvalues of $\partial U_i/\partial r_j$ go negative.

The Zel’dovich approximation seems to give a good picture of formation of structure in the HDM model, but continued unaccelerated motion of particles after shell crossing is clearly unrealistic. A useful modification of Zel’dovich’s approximation is to assume that particles move ballistically until shell-crossing, at which point they stick together. This is described by *Burger’s equation*, and gives infinitesimally thin walls. This is also obviously unrealistic, but actually receives some justification if we think of the Universal expansion adiabatically stretching a self-gravitating sheet. If the thickness of the sheet is T and the surface density Σ , then the acceleration of a particle at the surface is $\ddot{r} \sim G\Sigma$, and the frequency of oscillation of particles through the sheet is

$$\omega \sim \sqrt{\frac{\ddot{r}}{T}} \sim \sqrt{\frac{G\Sigma}{T}}. \quad (34.26)$$

Now the surface density decreases as $\Sigma \propto 1/a^2$ for a sheet expanding in the transverse direction at the Hubble rate, so applying the law of adiabatic invariance $r = A \cos(\omega t)$ with amplitude $A \propto 1/\sqrt{\omega}$ and requiring $A \simeq T$ we find that the (physical) thickness must evolve as

$$T \propto a^{2/3}. \quad (34.27)$$

This increases with time, but not as fast as the scale factor $a(t)$, so in comoving coordinates the sheet should indeed become thin.

Another nice feature of the Zel'dovich approximation is that one can compute the non-linear power spectrum analytically in terms of the initial power spectrum, as described in detail in chapter 8.

34.4 Press-Schechter Mass Function

The *Press-Schechter approximation* is designed for hierarchical type initial fluctuation fields. It provides one with a useful approximation for the *differential mass function* $n(M) = dN(> M)/dM$, where the *cumulative mass function* $N(> M)$ is the comoving number density of bound structures with mass $> M$.

The idea is that one identify two quantities: The first is the fraction of space where the initial density contrast field $\delta(\mathbf{r})$, when filtered with a kernel of mass M , lies above the threshold δ_{crit} for formation of non-linear condensations.

$$f(\delta > \delta_{\text{crit}}; M) = \int_{\delta_{\text{crit}}/\sigma(M)}^{\infty} \frac{d\nu}{\sqrt{2\pi}} \exp(-\nu^2/2). \quad (34.28)$$

The second is the fraction of mass in objects more massive than M

$$f(> M) = \int_M^{\infty} dM M n(M). \quad (34.29)$$

Differentiating (34.28) and (34.29) with respect to M and equating we get

$$n(M) = \frac{\delta_{\text{crit}} d\sigma(M)/dM}{\sqrt{2\pi} M \sigma^2(M)} \exp(-\delta_{\text{crit}}^2/2\sigma^2(M)) \quad (34.30)$$

While hard to justify rigorously, the idea obviously contains an element of truth, and moreover seems to give predictions which agree with the results of N-body experiments.

If one assumes a power-law spectrum $P(k) \propto k^n$ then the variance as a function of smoothing mass M is also a power law, $\sigma^2(M) \propto M^{-(n+3)/3}$. In ‘hierarchical models’ (those with $n > -3$) the mass variance increases with decreasing mass. At sufficiently low masses we must have $\sigma \gg \delta_{\text{crit}}$ and the exponential factor becomes close to unity and the theory predicts a power-law differential mass function. For $n = -2$, for instance, which is the slope of the CDM spectrum around the mass scale of galaxies the theory predicts $n(M) \propto M^{-5/6}$. At high masses, when $\delta_{\text{crit}}/\sigma(M)$ starts to exceed unity the exponential factor becomes very small. The general prediction is for a power-law mass function which becomes exponentially cut off above some characteristic mass scale; the mass M_* where $\sigma(M_*) \simeq 1$. This is just the kind of behavior seen in the *galaxy luminosity function* and also in the *cluster mass function*.

34.5 Biased Clustering

In the Press-Schechter theory, collapsed objects are associated with regions where the initial over-density, smoothed on an appropriate mass-scale, is sufficiently large. A consequence of this is that objects on the high end of the mass function — those with $M \gtrsim M_*$ that is — will tend to have amplified large large-scale clustering properties. Their clustering is said to be positively biased. The effect is illustrated in figure 34.3 which shows how the density of over-dense regions is modulated by

long wavelength modes of the density field. This effect is fairly obvious, but what is less obvious is how the strength of the modulation increases as one raises the threshold. This is consequence of the peculiar property of a Gaussian distribution. The Gaussian distribution is $P(\nu) \propto \exp(-\nu^2/2)$. In the vicinity of some value $\nu = \nu_0$ the distribution for $\Delta\nu = \nu - \nu_0$ is $P(\Delta\nu) \propto \exp(-(\nu_0 + \Delta\nu)^2/2)$. Expanding the quadratic factor in the exponential, and assuming $\nu_0 \gg \Delta\nu$ gives

$$P(\Delta\nu) \sim \exp(-\nu_0 \Delta\nu). \quad (34.31)$$

Thus a Gaussian looks locally exponential: $P(\Delta\nu) \sim \exp(-\Delta\nu/\sigma_{\Delta\nu})$ with exponential scale length $\sigma_{\Delta\nu} = 1/\nu_0$ which decreases with increasing ν_0 . Thus the further out we go on the tail of a Gaussian the steeper the distribution becomes.

If we add a positive background field δ_b , the fractional change in the probability to exceed the threshold is then $\Delta P/P \simeq \nu_0 \delta_b / \sigma$. The fluctuation in the number density of upward excursions is then

$$1 + \frac{\delta n}{n} = 1 + b \delta_b \quad (34.32)$$

where the *bias factor* is

$$b = \frac{\nu_0}{\sigma} = \frac{\delta_{\text{crit}}}{\sigma^2}. \quad (34.33)$$

Since δ_{rmcrit} is constant here, the bias factor rapidly increases with the mass of the objects (because $\sigma^2(M)$ decreases with increasing mass). This is the linearized bias; valid for very small δ_b , such that $b\delta_b \ll 1$. It is not difficult to show that for $\delta_b \lesssim 1$, the density of upward fluctuations is proportional to $\exp(b\delta_b)$. Thus the density of objects is the exponential of the background field.

One solid application of this theory is to clusters of galaxies; these are the most massive gravitationally collapsed objects, and so are naturally identified with particularly high peaks. For a long time, the very strong clustering of such objects was a puzzle; they have a correlation length of about 20Mpc as compared to about 5Mpc for galaxies. Now we understand that this is just about what one would expect given Gaussian initial density fluctuations. It is tempting to apply this theory also to galaxies, but there the connection between theory and observation is more tenuous. However, at high redshift one would expect the rare, most massive galaxies to be the analog of very massive clusters today, and this theory then provides a natural explanation for the rather strong clustering of ‘Lyman-break’ galaxies at $z \sim 3$.

34.6 Self-Similar Clustering

Another useful approximation for ‘hierarchical’ initial conditions is the *self-similar evolution model*. The idea is that if the initial spectrum of fluctuations approximates a power law over some range of wave-number then the structure should evolve in such a way that the non-linear density field at one time is a scaled replica of the field at another time, where the scaling factor in mass say is given by $M_*(t)$; the nominal mass going non-linear.

For a power law spectrum $P(k) \propto k^n$, the mass variance scales as $\sigma^2 \sim (k^3 P(k))_{k=1/r} \propto r^{-(n+3)}$. The root mean squared mass fluctuations grow with time in proportion to the scale factor $a(t)$, and therefore σ scales with $M \propto r^3$ and a as

$$\sigma(M, t) \propto a M^{-(n+3)/6}. \quad (34.34)$$

The mass-scale of non-linearity ($\sigma \sim 1$) therefore scales as

$$M_* \propto a^{6/(n+3)} \quad (34.35)$$

and the scaling law is that the fraction of mass per log interval of mass is a function only of $M/M_*(t)$:

$$M^2 n(M, t) = F(M/M_*(t)) \quad (34.36)$$

(see figure (34.4)). This relation does not specify anything about the universal function $F(y)$, but it does allow one to predict the mass function at redshift $z > 0$ given the form at $z = 0$.

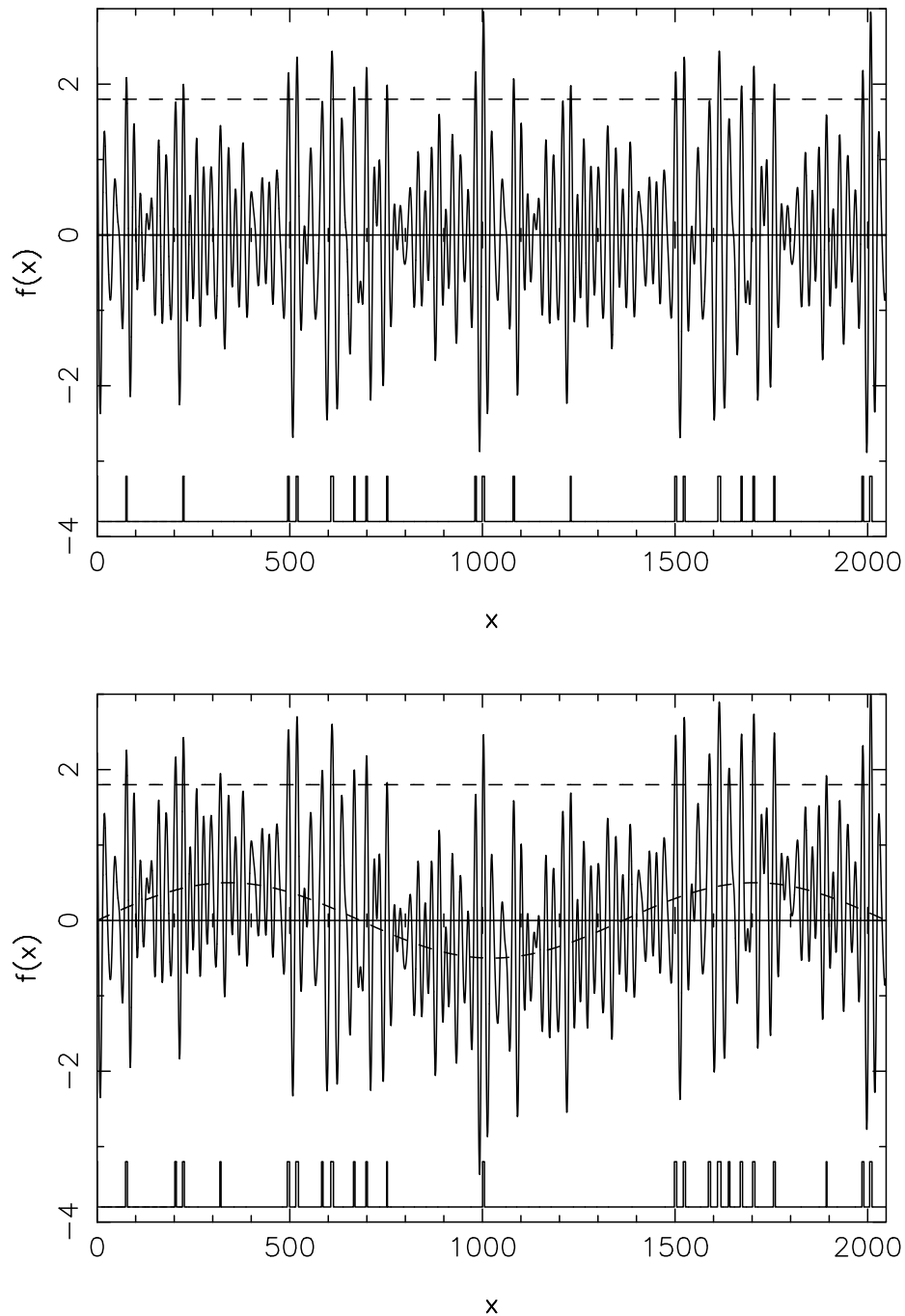


Figure 34.3: The upper panel shows a realization of a Gaussian random noise field. This is supposed to represent the initial Gaussian density perturbation field $\delta(\mathbf{r})$. The horizontal dashed line is supposed to represent the threshold density required in order for a region to have collapsed. The lower trace shows the ‘excursion set’ for this threshold (here taken to be 1.8 times the root mean squared fluctuation). This function is one or zero depending on whether $f(x)$ exceeds the threshold. The positive parts of the excursion set are randomly distributed with position. The lower panel shows the same thing, but where we have added a long-wavelength sinusoidal ‘background’ field. Clearly, and not surprisingly, the background field has modulated the density of the regions exceeding the critical threshold.

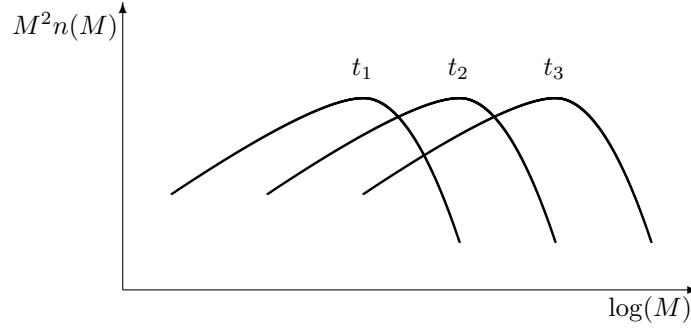


Figure 34.4: If the initial spectrum of density fluctuations has a power-law form, the differential mass function $n(M)$ evolves such that $M^2 n(M, t)$ is a universal function $F(M/M_*(t))$. Here $M_*(t)$ is the mass going non-linear at time t . Self-similarity does not tell one the form of the universal function $F(y)$ — sketched here as a Schechteresque function — but it does enable one to retrodict the mass function at earlier times from observations of the current mass function. Since the evolution of the characteristic mass-scale $M_*(t)$ depends on the initial spectral index, this provides a useful test of cosmological theory.

This model provides a simple, but apparently quite accurate, model for the evolution of structures in CDM-like models where the slope of the spectrum varies quite gradually with mass. The models are parameterized by n , the effective slope of the power spectrum on mass scales of interest. For CDM this varies from $n \simeq -1$ on the scale of clusters to $n \simeq -2$ on the scale of galaxies.

The scaling makes no assumption about the statistical nature of the density field and is applicable to e.g. the cosmic string model for structure formation.

34.7 Davis and Peebles Scaling Solution

The goes beyond the self-similar scaling and attempts to determine the slope of the two-point correlation function in the non-linear regime from the slope n of the initial power spectrum, assumed to be power-law like with $P(k) \propto k^n$.

The original discussion was couched in terms of the BBGKY hierarchy, but the essential result can be easily obtained from conservation of energy considerations, much as we did for the accretion onto a point mass.

With the initial spectrum for the density fluctuations δ and with $\nabla^2 \phi = 4\pi G \rho \delta$, so $\delta \phi_k = 4\pi G \rho \delta_k / k^2$ the root mean square potential fluctuations on scale r are

$$\langle \delta \phi^2 \rangle_r^{1/2} \sim \left[\int d^3 k k^{-4} k^n \tilde{W}_r(k) \right]^{1/2} \quad (34.37)$$

where $\tilde{W}_r(k)$ is the transform of the smoothing kernel, which falls rapidly for $k \gg 1/r$. This gives

$$\langle \delta \phi^2 \rangle_r^{1/2} \propto r^{(1-n)/2}. \quad (34.38)$$

In terms of mass scale, $M \propto r^{1/3}$ this is

$$\langle \delta \phi^2 \rangle_r^{1/2} \propto M^{(1-n)/6}. \quad (34.39)$$

One the other hand, in the non-linear regime, we have a power-law mass auto-correlation function $\xi(r) \propto r^{-\gamma}$. Now imagine the mass distribution to be a set of randomly distributed clumps of size r and over-density $\delta_* \gg 1$. The fraction of space occupied by the clumps is $f \sim 1/\delta_*$, so the density fluctuation variance is $\xi(r) \simeq \langle \delta^2 \rangle \sim f \delta_*^2 \sim \delta_*$. The mass of a lump is $M \sim \bar{\rho} \delta_* r^3$, so the

characteristic mass of clumps of size r is $M \propto r^{3-\gamma}$. The binding energy of clumps then scales with their radius and mass as

$$\delta\phi \sim M/r \propto r^{2-\gamma} \propto M^{(2-\gamma)/(3-\gamma)}. \quad (34.40)$$

Equating (34.39) and (34.40), we obtain the relation

$$\gamma = \frac{9+3n}{5+n} \quad (34.41)$$

which would fit with the empirically observed slope $\gamma \simeq 1.8$ for a white noise spectrum $n = 0$.

While the derivation here is similar to that for spherical accretion, the result is much less robust. While it makes perfect sense to say that the binding energy of structures when they first form is given, within a geometrical factor of order unity, by the initial binding energy, the calculation here assumes that even when much larger mass objects have collapsed, the small clumps still preserve the binding energy with which they are born. This is not likely to be the case, as there will be transfer of energy between the different scales of the hierarchy. As we have argued above, entropy considerations suggest that such interactions will tend to erase sub-structure. Numerical simulations do not provide much support for this theory.

34.8 Cosmic Virial Theorem

The *cosmic virial theorem* (Davis and Peebles again) attempts to relate the low order correlation functions for galaxies to the relative motions of galaxies and thereby obtain an estimate of the mass-to-light ratio of mass clustered along with galaxies.

In essence, their argument is as follows: Assume that galaxies cluster like the mass — this means that the excess mass within distance r of a galaxy grows like $M \propto \int_0^r d^3r \xi(r) \propto r^{3-\gamma}$. The potential well depth is then $\delta\phi \sim GM/r \propto r^{2-\gamma}$. One would expect the relative velocity of galaxies at separation r to scale as

$$\sigma^2(r) \propto r^{2-\gamma} \simeq r^{0.2} \quad (34.42)$$

This prediction seems to be remarkably well obeyed on scales from a few tens of kpc out to about 1 Mpc (and one would not expect the result to hold at larger separations where things have yet to stabilize anyway).

From the size of the peculiar motions, one infers that the mass-to-light ratio of material clustered around galaxies on scale ~ 1 Mpc or less is $M/L \simeq 300h$ in solar units. If representative of the universal value, this would imply $\Omega \simeq 0.2$. This is similar to the mass-to-light ratio from virial analysis of individual clusters of galaxies, and provides strong supporting evidence for copious amounts of dark matter. It also supports the hypothesis that the galaxies cluster like the mass, and therefore that the universal density parameter is $\Omega \simeq 0.2$ rather than the aesthetically pleasing $\Omega = 1$.

Part VII

Appendices

Appendix A

Vector Calculus

A.1 Vectors

The *prototype vector* is the distance between two points in space

$$\mathbf{d} = (d_x, d_y, d_z) = d_i. \quad (\text{A.1})$$

A three component entity v_i is a vector if it transforms under rotations in the same manner as d_i .

A.2 Vector Products

The *scalar product* or *dot product* of two vectors is

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^3 a_i b_i = a_i b_i \quad (\text{A.2})$$

The *cross product* or *outer product* is

$$\mathbf{c} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix} = \begin{bmatrix} a_y b_z - a_z b_x \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{bmatrix}. \quad (\text{A.3})$$

A.3 Div, Grad and Curl

The *gradient operator* is

$$\nabla = (\partial_x, \partial_y, \partial_z) = \partial_i \quad (\text{A.4})$$

with $\partial_i \equiv \partial/\partial x_i$.

The gradient of a scalar field $f(\mathbf{r})$ is a vector

$$\nabla f = \partial_i f. \quad (\text{A.5})$$

The *divergence* of a vector field $\mathbf{v}(\mathbf{r})$ is a scalar

$$\nabla \cdot \mathbf{v} = \left(\frac{\partial v_x}{\partial x}, \frac{\partial v_y}{\partial y}, \frac{\partial v_z}{\partial z} \right) = \frac{\partial v_i}{\partial x_i}. \quad (\text{A.6})$$

The *Laplacian operator* is $\nabla^2 \equiv \nabla \cdot \nabla$ and yields a scalar when applied to a scalar and a vector when applied to a vector.

The Laplacian of a spherically symmetric function $f(r)$ is

$$\nabla^2 f = \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{df}{dr} \right). \quad (\text{A.7})$$

The *curl* of a vector field $\mathbf{v}(\mathbf{r})$ is a vector

$$\nabla \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial_x & \partial_y & \partial_z \\ v_x & v_y & v_z \end{vmatrix} = \begin{bmatrix} \partial_y v_z - \partial_z v_y \\ \partial_z v_x - \partial_x v_z \\ \partial_x v_y - \partial_y v_x \end{bmatrix}. \quad (\text{A.8})$$

The *curl of the curl* of a vector field is

$$\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v} \quad (\text{A.9})$$

The *divergence of the curl* of a vector field vanishes

$$\nabla \cdot (\nabla \times \mathbf{v}) = 0 \quad (\text{A.10})$$

The *divergence of the cross-product* of two vector fields is

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B}) \quad (\text{A.11})$$

Another identity which is useful in fluid dynamics is

$$(\mathbf{u} \cdot \nabla) \mathbf{u} = \frac{1}{2} \nabla u^2 - \mathbf{u} \times (\nabla \times \mathbf{u}). \quad (\text{A.12})$$

The above vector identities are all readily verified by writing out the cross-products using the determinant form for the cross product.

A.4 The Divergence Theorem

The *divergence theorem* says that the volume integral of the divergence of a vector field $\mathbf{v}(\mathbf{r})$ is equal to the integral over the surface of the volume of the normal component of \mathbf{v} :

$$\int_V d^3r \nabla \cdot \mathbf{v} = \int_S d\mathbf{A} \cdot \mathbf{v} \quad (\text{A.13})$$

and can be proven by integrating by parts.

A.5 Stokes' Theorem

The integral of the normal component of the curl of a vector field \mathbf{v} over a surface is equal to the loop integral of the tangential component of the field around the perimeter:

$$\int d\mathbf{A} \cdot \nabla \times \mathbf{v} = \oint d\mathbf{l} \cdot \mathbf{v}. \quad (\text{A.14})$$

A.6 Problems

A.6.1 Vector Calculus Identities

The cross product of two vectors \mathbf{a} , \mathbf{b} can be expressed as the determinant of the matrix

$$\mathbf{c} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix} = \begin{bmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{bmatrix} \quad (\text{A.15})$$

and the same is true if the components of the vector \mathbf{a} are differential operators: $\mathbf{a} = \nabla = (\partial_x, \partial_y, \partial_z)$.

Use this result to show that:

1. The *curl of the curl* of a vector field is

$$\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v} \quad (\text{A.16})$$

2. The *divergence of the curl* of a vector field vanishes

$$\nabla \cdot (\nabla \times \mathbf{v}) = 0 \quad (\text{A.17})$$

3. The *divergence of the cross-product* of two vector fields is

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B}) \quad (\text{A.18})$$

4. and finally that

$$(\mathbf{u} \cdot \nabla) \mathbf{u} = \frac{1}{2} \nabla u^2 + \mathbf{u} \times (\nabla \times \mathbf{u}). \quad (\text{A.19})$$

Appendix B

Fourier Transforms

B.1 Discrete Fourier Transform

Consider a set of N uniformly spaced samples of a 1-dimensional scalar function F_X , $X = 0, N - 1$. Define the *discrete Fourier transform* as

$$\tilde{F}_K \equiv \sum_{X=0}^{N-1} F_X e^{i2\pi KX/N} \quad (\text{B.1})$$

Now consider the function

$$G_X \equiv \sum_{K=0}^{N-1} F_K e^{-i2\pi KX/N} \quad (\text{B.2})$$

substituting for F_K from (B.1) this becomes

$$G_X = \sum_{X'=0}^{N-1} F_{X'} \sum_{K=0}^{N-1} e^{i2\pi K(X'-X)/N} \quad (\text{B.3})$$

The second sum here is a simple geometric series, with value

$$\sum_{K=0}^{N-1} e^{i2\pi K(X'-X)/N} = \frac{1 - e^{i2\pi(X'-X)}}{1 - e^{i2\pi(X'-X)/N}} \quad (\text{B.4})$$

the numerator here is zero for all $X' - X$ (since $X' - X$ is an integer) whereas the denominator is finite unless $X' - X$ is a multiple of N . For X, X' in the range $0, N - 1$ then the sum vanishes unless $X' - X = 0$, in which case it has value N , so we have

$$\sum_{K=0}^{N-1} e^{i2\pi KX/N} = N\delta_X \quad (\text{B.5})$$

where δ_X is the *discrete delta-function* which is unity or zero if X is zero or non-zero respectively. It follows from (B.3) that $G_X = NF_X$, from which we obtain the *inverse discrete Fourier transform*

$$F_X = \frac{1}{N} \sum_{K=0}^{N-1} \tilde{F}_K e^{-i2\pi KX/N} \quad (\text{B.6})$$

Equations (B.1), (B.6) allow one to reversibly transform from real-space to Fourier-space representation and *vice versa*.

B.2 Continuous Fourier Transform

We can translate the foregoing results for discrete transforms by converting sums to integrals. Let the real-space domain have length L so the real-space coordinate is $x = X\Delta x$, where the ‘pixel’ size is $\Delta x = L/N$, and define $f(x) = F_{X=x/\Delta x}$. Similarly define the continuous wave number $k = K\Delta k$ with $\Delta k = 2\pi/L$ and set $\tilde{f}(k) = \Delta x \tilde{F}_{K=k/\Delta k}$ to obtain from (B.1)

$$\tilde{f}(k) = \sum_{X=0}^{N-1} \Delta x f(x) e^{ikx} \rightarrow \int_0^L dx f(x) e^{ikx} \quad (\text{B.7})$$

and the discrete inverse transform becomes

$$f(x) = \frac{1}{N\Delta k\Delta x} \sum_{K=0}^{N-1} \Delta k \tilde{f}(k) e^{-ikx} \rightarrow \int \frac{dk}{2\pi} \tilde{f}(k) e^{-ikx} \quad (\text{B.8})$$

it is also of interest to convert the expression for the discrete δ -function to a continuous integral:

$$\sum_{K=0}^{N-1} e^{iKX/N} \rightarrow \frac{N\Delta k}{2\pi} \delta_X = \frac{\delta_X}{\Delta x} \quad (\text{B.9})$$

The final expression here has value $1/\Delta x$ if we are in the zeroth pixel $X = 0$ or equivalently if $0 < x < \Delta x$ so in the continuum limit $\Delta x \rightarrow 0$ this is a representation of the delta function and we have

$$\int \frac{dk}{2\pi} e^{ikx} = \delta(x) \quad (\text{B.10})$$

where the ‘Dirac δ -function’ $\delta(x)$ has the property that for any function $h(x)$

$$\int dx' h(x') \delta(x - x') = h(x) \quad (\text{B.11})$$

Note that if a is held constant inside an integral

$$\int dy f(y) \delta(ay) = \int \frac{dz}{a} f(z/a) \delta(z) = f(0)/a = \int dy f(y) \delta(y)/a \quad (\text{B.12})$$

thus

$$\delta(ay) = \delta(y)/a \quad (\text{B.13})$$

which is a useful result.

The generalization to M-dimensional space is straightforward and we have

$$\begin{aligned} \tilde{f}(\vec{k}) &= \int d^M r f(\vec{x}) e^{i\vec{k} \cdot \vec{x}} \\ f(\vec{x}) &= \int \frac{d^M k}{(2\pi)^M} \tilde{f}(\vec{k}) e^{-i\vec{k} \cdot \vec{x}} \end{aligned} \quad (\text{B.14})$$

and

$$\int \frac{d^M k}{(2\pi)^M} e^{i\vec{k} \cdot \vec{x}} = \delta(\vec{x}) \quad (\text{B.15})$$

Note that if $f(\vec{x})$ is real then $\tilde{f}(-\vec{k}) = \tilde{f}(\vec{k})^*$.

B.3 Parseval's Theorem

We now derive *Parseval's theorem* for discrete functions and then for continuous functions. Consider the sum of the squared values of a discrete function F_X . Invoking (B.6) we can write this as

$$\begin{aligned}\sum F_X^2 &= \sum \frac{1}{N} \sum \tilde{F}_K e^{-i2\pi X K/N} \frac{1}{N} \sum \tilde{F}_{K'}^* e^{i2\pi X K'/N} \\ &= \frac{1}{N^2} \sum \sum \tilde{F}_K \tilde{F}_{K'}^* \sum e^{-i2\pi X (K-K')/N} \\ &= \frac{1}{N} \sum \sum \tilde{F}_K \tilde{F}_{K'}^* \delta_{K-K'} = \frac{1}{N} \sum |F_K|^2\end{aligned}\quad (\text{B.16})$$

and for continuous functions we have

$$\begin{aligned}\int dx f^2(x) &= \int dx \int \frac{dk}{2\pi} \tilde{f}(k) e^{-ikx} \int \frac{dk'}{2\pi} \tilde{f}(k')^* e^{ikx} \\ &= \int \frac{dk}{2\pi} \int \frac{dk'}{2\pi} \tilde{f}(k) \tilde{f}(k')^* \int dx e^{-i(k-k')x} \\ &= \int \frac{dk}{2\pi} \int \frac{dk'}{2\pi} \tilde{f}(k) \tilde{f}(k')^* 2\pi \delta(k-k') = \int \frac{dk}{2\pi} |\tilde{f}(k)|^2\end{aligned}\quad (\text{B.17})$$

B.4 Convolution Theorem

We define the convolution of two continuous functions $f(x)$, $g(x)$ to be

$$c(x) = \int dx' f(x') g(x-x') \quad (\text{B.18})$$

the transform of which is

$$\begin{aligned}\tilde{c}(k) &= \int dx c(x) e^{ikx} = \int dx \int dx' f(x') g(x-x') e^{ikx} \\ &= \int dx \int dx' \int \frac{dk'}{2\pi} \tilde{f}(k') \tilde{g}(k-k') e^{i(kx-k'x-(k-k')(x-x'))} \\ &= \int \frac{dk'}{2\pi} \int \frac{dk''}{2\pi} \tilde{f}(k') \tilde{g}(k'') \int dx e^{i(k-k'')x} \int dx' e^{i(k''-k')x'} \\ &= \int dk' \int dk'' \tilde{f}(k') \tilde{g}(k'') \delta(k-k'') \delta(k''-k') \\ &= \tilde{f}(k) \tilde{g}(k)\end{aligned}\quad (\text{B.19})$$

so the transform of a convolution of two functions is the product of the individual transforms. A direct corollary is that the transform of a product is the convolution of the individual transforms.

B.5 Wiener-Khinchin Theorem

The *auto-correlation* of a function f is

$$\begin{aligned}\xi(x) &= \frac{1}{L} \int dx' f(x') f(x'+x) \\ &= \frac{1}{L} \int dx' \int \frac{dk}{2\pi} \tilde{f}(k) e^{-ikx'} \int \frac{dk'}{2\pi} \tilde{f}(k') e^{-ik'(x+x')} \\ &= \frac{1}{L} \int \frac{dk}{2\pi} \int \frac{dk'}{2\pi} \tilde{f}(k) \tilde{f}(k') e^{-ik'x} \int dx' e^{-i(k+k')x'} \\ &= \frac{1}{L} \int \frac{dk}{2\pi} \int \frac{dk'}{2\pi} \tilde{f}(k) \tilde{f}(k') e^{-ik'x} 2\pi \delta(k+k') \\ &= \frac{1}{L} \int \frac{dk}{2\pi} \tilde{f}(-k) \tilde{f}(k) e^{-ikx} = \int \frac{dk}{2\pi} P(k) e^{-ikx}\end{aligned}\quad (\text{B.20})$$

where we have used $\tilde{f}(-k) = f^*(k)$ (since $f(x)$ is real) and where $P(k) \equiv |\tilde{f}(k)|^2/L$ is the *power spectrum*. Thus $\xi(x)$ is the FT of $P(k)$ and *vice versa*, which is the *Wiener-Khinchin theorem*. Note that for a statistically homogeneous random field $\xi(x)$ and $P(k)$ tend in an average sense to a limit which is independent of the sampling box size L .

B.6 Fourier Transforms of Derivatives and Integrals

If

$$F(x) \equiv \frac{df(x)}{dx} \quad (\text{B.21})$$

then by equation (B.8) we have

$$\begin{aligned} \int \frac{dk}{2\pi} \tilde{F}(k) e^{-ik \cdot x} &= \frac{d}{dx} \int \frac{dk}{2\pi} \tilde{f}(k) e^{-ik \cdot x} \\ &= \int \frac{dk}{2\pi} \tilde{f}(k) \frac{d e^{-ik \cdot x}}{dx} \\ &= \int \frac{dk}{2\pi} (-ik) \tilde{f}(k) e^{-ik \cdot x} \end{aligned} \quad (\text{B.22})$$

so the transform of the derivative of a function is $-ik$ times the transform of the function:

$$\tilde{F}(k) = -ik \tilde{f}(k) \quad (\text{B.23})$$

One can also obtain this result by integration by parts.

The transform of the integral of a function is similarly equal to the transform of that function divided by $-ik$. This is ill-defined for $k = 0$. This reflects the fact that an integral is determined only up to an additive constant, and the $k = 0$ term in a Fourier expansion is just this constant.

B.7 Fourier Shift Theorem

The Fourier transform of a shifted field $f'(x) = f(x + d)$ is given by

$$\tilde{f}'(k) = e^{ikd} \tilde{f}(k). \quad (\text{B.24})$$

B.8 Utility of Fourier Transforms

One use for Fourier transforms in astrophysics is to convert differential equations to algebraic equations; the ‘grad’ operator ∇ , becoming multiplication by $-i\mathbf{k}$ in the Fourier domain.

Fourier transforms are a great help in describing random processes — and especially statistically homogeneous random processes.

Fourier transforms are computationally very useful for convolving or de-convolving data. Say you want to convolve a $N \times N$ image with some extended smoothing kernel. At face value this would appear to take $\mathcal{O}N^4$ operations (for each of the N^2 ‘destination image’ pixels you need to sum over N^2 ‘source image’ pixels). The operation count is smaller for a smaller kernel, but still extremely expensive for large images. The great advantage of performing such smoothing in Fourier transform space is that the *fast Fourier transform* algorithm requires only $\mathcal{O}N^2 \log(N^2)$ operations. Thus, the preferred method to convolve an image with a kernel is to transform both with the FFT, multiply the transforms and then inverse transform the result.

B.9 Commonly Occurring Transforms

Here we describe some useful common transforms (see figure B.1). We have taken some liberties with normalization.

- The FT of a δ -function at the origin $f(\mathbf{r}) = \delta(\mathbf{r})$ is a constant. If the δ -function is shifted, so $f(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_0)$, the constant is multiplied by $\exp(i\mathbf{k} \cdot \mathbf{r})$.
- A 1-dimensional δ -function is the derivative of a *step function*. Hence the transform of a step function is proportional to $1/k$.
- The transform of a Gaussian $f(r) = \exp(-r^2/2\sigma^2)$ is another Gaussian $\tilde{f}(k) = \exp(-k^2\sigma^2/2)$. This generalized the result for a δ -function — which can be thought of as a Gaussian with $\sigma \rightarrow 0$. Fine-scale structures in real space correspond to extended features in transform space.
- The transform of a ‘box-car’ or ‘top-hat’ function $f(\mathbf{r}) = 1$ for $|r| < L/2$ is the ‘sinc’ function: $\tilde{f}(k) = \sin(kL/2)/(kL/2)$.
- The transform of a ‘comb’ function $c_{\Delta r}(r) = \sum_n \delta(r - n\Delta r)$ is another comb function: $\tilde{c}(k) = c_{2\pi/\Delta r}(k) = \sum_n \delta(k - 2\pi n/\Delta r)$. Comb functions are often referred to as ‘Shah’ functions.

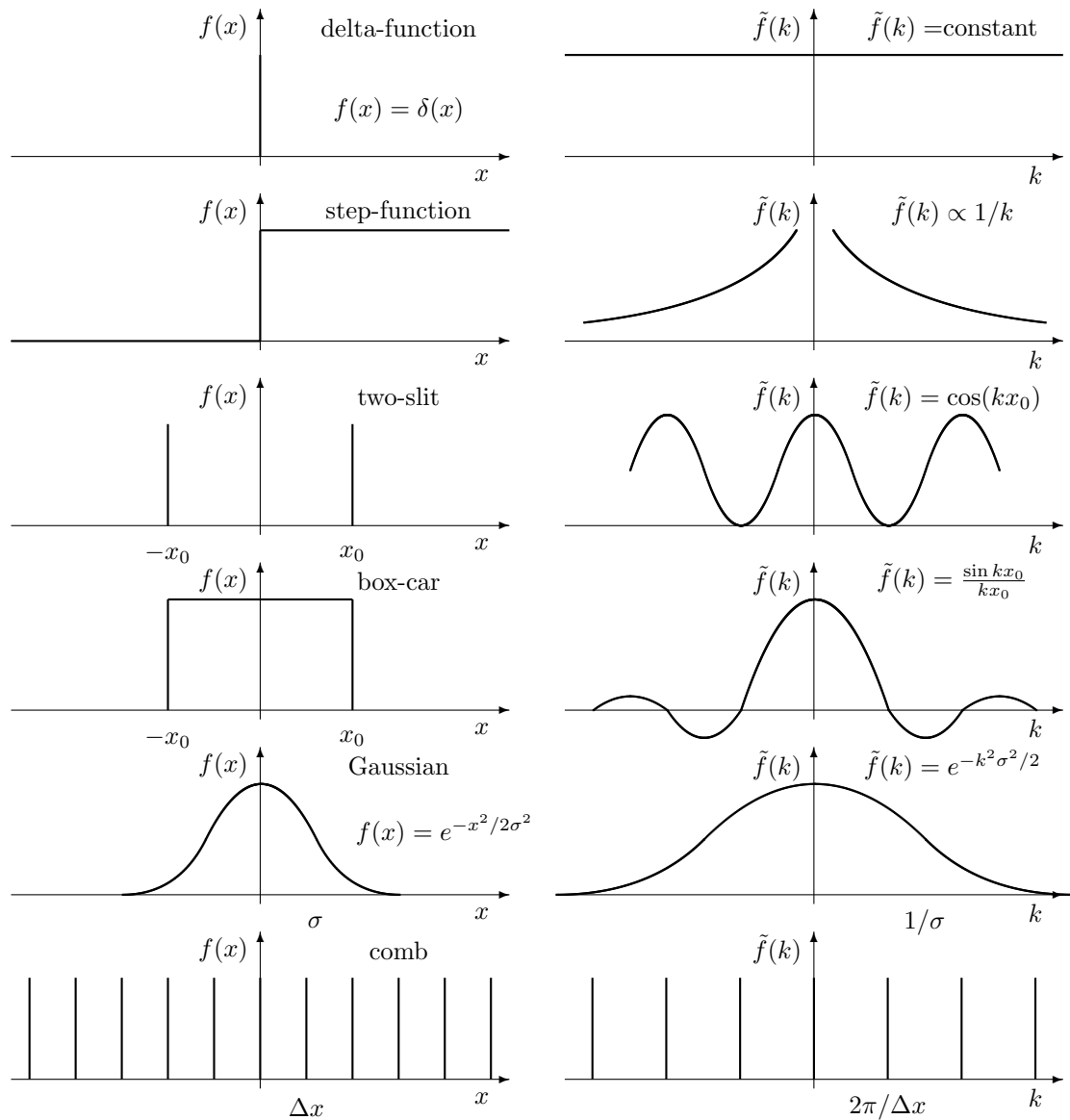


Figure B.1: A variety of Fourier transform pairs.

B.10 The Sampling Theorem

As a useful and interesting application of some of the foregoing results consider the transform of a ‘pixellated’ image. Say we have some source image $f(r)$ which we measure at a set of locations separated by Δr (see figure B.2). The result is a set of δ -function samples $f_{\text{pix}}(r) = c_{\Delta r} \times f$. The transform of these pixel values is, by the convolution theorem, $\tilde{f}_{\text{pix}}(k) = c_{2\pi/\Delta r} \otimes \tilde{f}$. This is the sum of a set of replicas of the transform of the source image placed at locations $k = 2\pi n/\Delta r$.

Now suppose that the original signal is ‘band-limited’, so $\tilde{f}(k) \neq 0$ only for $|k| \leq k_{\text{max}}$. This is not at all an unreasonable assumption; any astronomical telescope produces images which are strictly band-limited. If the cut-off frequency is less than one half of the spacing between the of the comb $c_{2\pi/\Delta r}$ then the replicas do not overlap. This means that the transform of the source image can be recovered exactly from the transform of the pixels, simply by multiplying by a box-car function of width $2k_{\text{max}}$:

$$\tilde{f}(k) = W(k) \times \tilde{f}_{\text{pix}}(k) \quad (\text{B.25})$$

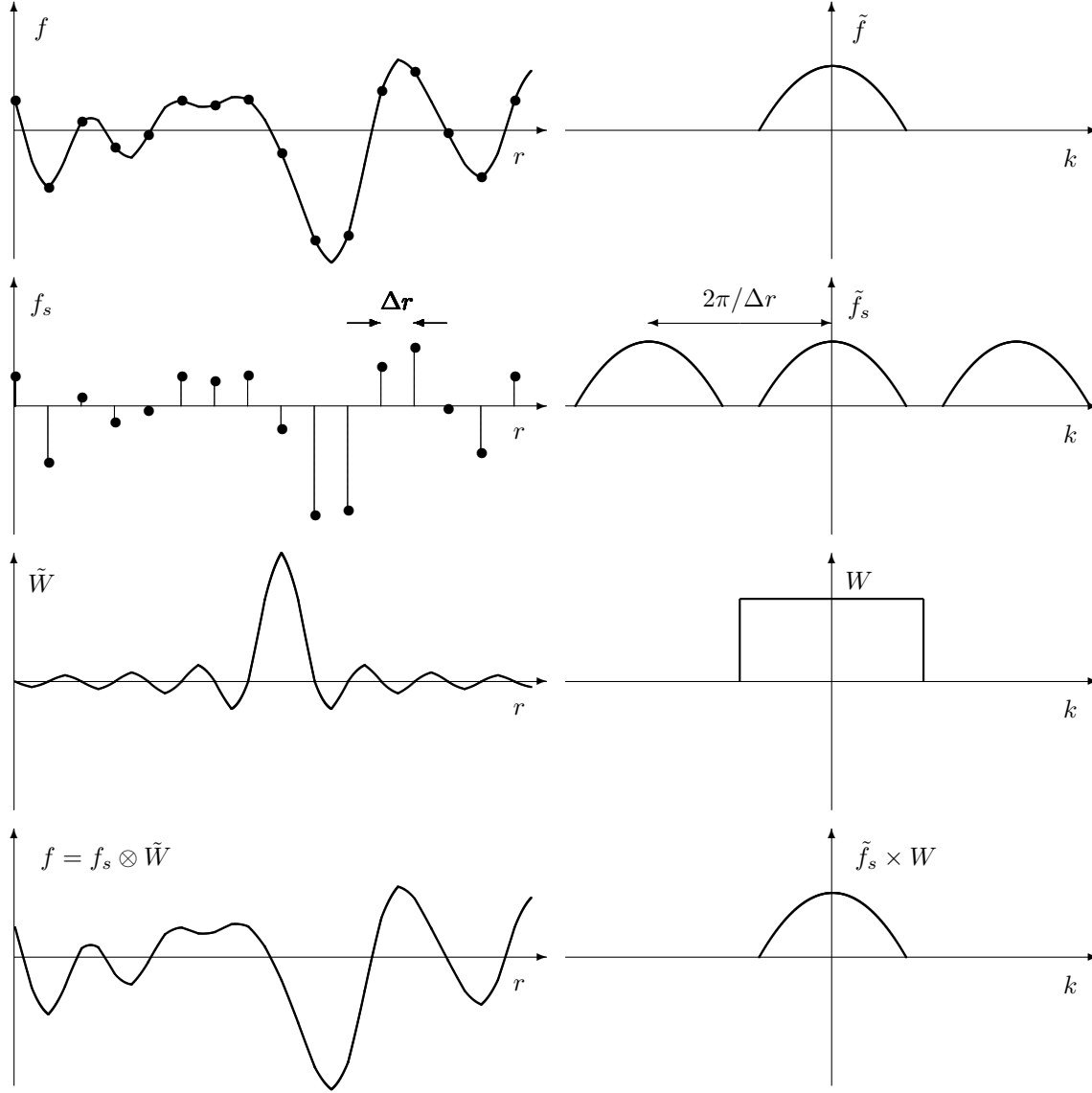


Figure B.2: Illustration of the sampling theorem. The upper left plot shows a random function $f(r)$ and its transform $\tilde{f}(k)$ is shown schematically in the upper right plot. Sampling $f(r)$ at a set of uniformly spaced points yields the function $f_s(r)$ (a set of δ -functions) whose transform $\tilde{f}_s(k)$ is the superposition of a set of replicas of the transform of the original field on a grid of spacing $\Delta k = 2\pi/\Delta r$. If the sampling rate is sufficiently high then these replicas do not overlap. It is then possible to recover the transform of the original field simply by windowing \tilde{f}_s with a box-car $W(k)$. Inverse transforming recovers the original field $f(r)$. Equivalently, in real-space, one can recover the original field by convolving the sampled field with a ‘sinc’ function which is the transform of the box car.

where $W(k) = 1$ if $|k| \leq k_{\max}$ and zero otherwise.

The value of the source image at position r_0 is

$$f(r_0) = \int \frac{dk}{2\pi} \tilde{f}(k) e^{-ikr_0} = \int \frac{dk}{2\pi} W(k) \tilde{f}_{\text{pix}}(k) e^{-ikr_0} \quad (\text{B.26})$$

but $\tilde{f}_{\text{pix}}(k) = \sum_p f_p e^{ikr_p}$ where f_p is the value of the pixel at position r_p . Hence the source image

value can be recovered exactly from the finite number of samples f_p as

$$f(r_0) = \sum_p f_p \int \frac{dk}{2\pi} W(k) e^{ik(r_p - r_0)} = \sum_p f_p \tilde{W}(r_p - r_0). \quad (\text{B.27})$$

The transform of the box-car is $\tilde{W}(r) = \text{sinc}(k_{\text{max}}r)$, so this says we can recover the source image value at any position — not necessarily at a measured point — by combining the measured values with weights equal to $\text{sinc}(k_{\text{max}}r)$. This is commonly referred to as *sinc interpolation*.

The pixel spacing required in order to be able to sinc-interpolate is $\Delta r \leq \pi/k_{\text{max}}$. Equivalently, with given pixel spacing Δr , the source image should not contain any signal at wavelengths less than $\lambda_{\text{min}} = 2\Delta r$ — i.e. the shortest wavelength allowed just fits within two pixels. If this condition is satisfied, the image is said to be *critically sampled* and can be shifted and re-sampled without any loss of information.

One can show that for a telescope with primary diameter D , observing at wavelength λ , critical sampling requires that the angular pixel size be less than $\Delta\theta = \lambda/2D$. For HST ($D = 2.4\text{m}$) at visible wavelength of say $0.5\mu\text{m}$ this requires a pixel size of about $0''.02$. For the wide field imager WFPC, the size of the wide-field pixels are $0''.1$, so the instrument is not critically sampled. However, with multiple exposures, it is possible to reconstruct a critically sampled image.

A remarkable feature of sinc interpolation is that if one uses it to interpolate from one image grid onto another which is just shifted with respect to the first, then if there is noise in the source image which is uncorrelated from pixel to pixel, then the noise in the interpolated image is also uncorrelated.

B.11 Problems

B.11.1 Fourier Transforms

Compute the Fourier transforms of the following 1-dimensional functions. Also sketch both real-space and Fourier-space functions.

1. A delta-function: $f(x) = \delta(x - x_0)$.
2. A step function $f(x) = 1, 0$ for $x > 0, x < 0$ respectively.
3. A Gaussian: $f(x) = \exp(-x^2/2\sigma^2)$.
4. A ‘double-slit’: $f(x) = \delta(x - d) + \delta(x + d)$.
5. A ‘dipole’: $f(x) = \delta(x - d) - \delta(x + d)$.
6. A ‘box-car’: $f(x) = 1$ if $|x| < L/2$, $f(x) = 0$ otherwise.
7. A ‘comb-function’: $f(x) = \sum_n \delta(x - n\Delta x)$.

Appendix C

The Boltzmann Formula

Consider a gas of N distinguishable particles, each of which can be in one of a set of discrete energy states $E_i = i\Delta E$.

The number of different ways to assign these particles to a particular configuration, ie a certain set of *occupation numbers* $\{n_i\}$ for the different energy levels is

$$W = \frac{N!}{n_1!n_2!\dots n_{i-1}!n_i!n_{i+1}!\dots}. \quad (\text{C.1})$$

In the evolution of such a gas the occupation numbers will change as particles scatter off one another, but must obey the constraints

$$\sum_i n_i = N \quad \text{and} \quad \sum_i n_i E_i = E_{\text{total}}. \quad (\text{C.2})$$

If the system explores the available states in a random manner then the probability should be proportional to W . Now there is a certain set of occupation numbers \bar{n}_i for which W (and therefore also $\log W$) is maximized. We expect that the system will typically be found with occupation numbers very similar to \bar{n}_i . For this most probable configuration \bar{n} has a specific dependence on energy E_i which we will now deduce.

Consider a scattering event where two particles initially in level i end up in levels $i-1$, $i+1$, as illustrated in figure C.1. This clearly obeys the number and energy conservation. The new complexion is

$$W' = \frac{N!}{n_1!n_2!\dots (n_{i-1}+1)!(n_i-2)!(n_{i+1}+1)!\dots}. \quad (\text{C.3})$$

The ratio of complexions is

$$\frac{W'}{W} = \frac{n_i(n_i-1)}{(n_{i-1}+1)(n_{i+1}+1)} \simeq \frac{n_i^2}{n_{i-1}n_{i+1}} \quad (\text{C.4})$$

where we have assumed that the occupation numbers are large.

For the most probable configuration the complexion should be *stationary* with respect to small changes (such as the single scattering event described above), so $W'/W = 1$ or

$$\frac{n_{i-1}}{n_i} = \frac{n_i}{n_{i+1}}. \quad (\text{C.5})$$

This is true regardless of the choice of i . Thus, in the most probable configuration, the ratio of the occupation of one level to that of the next higher level is a constant

$$n(E + \Delta E)/n(E) = \text{constant} \quad (\text{C.6})$$

and the functional form which satisfies this is

$$n(E) = \alpha \exp(-\beta E) \quad (\text{C.7})$$

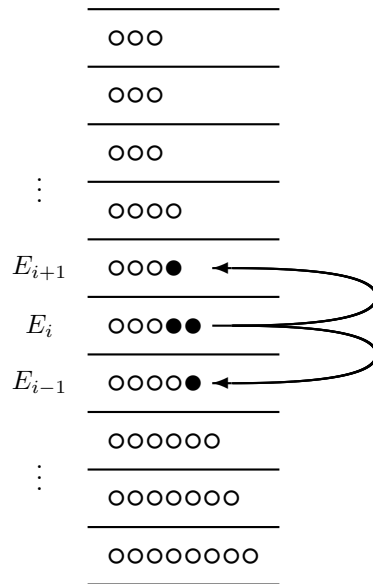


Figure C.1: The circles represent schematically the number of particles in each energy level. A possible ‘scattering’ event — i.e. one which respects conservation of both energy and particle number — is illustrated. Here two particles initially in energy level E_i end up in the adjacent energy levels. For large occupation numbers, and near equilibrium, the probability for the state should be stationary with respect to such exchanges. This requirement leads to the Boltzmann formula.

where α and β are constants, whose values are fixed once we specify the total number of particles and the total energy.

The occupation numbers $n(E)$ are proportional to the probability to find some arbitrary particle with energy E . Thus we obtain the *Boltzmann law*

$$p(E) \propto \exp(-\beta E). \quad (\text{C.8})$$

Appendix D

Dispersive Waves

Many physical systems have equations of motion that admit planar waves

$$\phi(\mathbf{x}, t) = \phi_{\mathbf{k}} e^{i(\omega_{\mathbf{k}} t - \mathbf{k} \cdot \mathbf{x})} \quad (\text{D.1})$$

as either exact or approximate solutions. In this, and in what follows, the actual wave is the real part of this expression. Examples covered here include sound waves, electro-magnetic waves, scalar fields, deep ocean waves, waves in a plasma etc.

A pure plane wave propagates with velocity

$$\mathbf{c} = \omega_{\mathbf{k}} \mathbf{k} / k^2 \quad (\text{D.2})$$

since

$$\phi(\mathbf{x}, t + \tau) = \phi_{\mathbf{k}} e^{i(\omega_{\mathbf{k}}(t+\tau) - \mathbf{k} \cdot \mathbf{x})} = \phi_{\mathbf{k}} e^{i(\omega_{\mathbf{k}} t - \mathbf{k} \cdot (\mathbf{x} - \mathbf{c}\tau))} = \phi(\mathbf{x} - \mathbf{c}\tau, t) \quad (\text{D.3})$$

so the wave at time $t + \tau$ is identical to the wave at time t but shifted by a distance $\Delta \mathbf{x} = \mathbf{c}\tau$. Equation (D.2) is the velocity at which surfaces of constant *phase* $\psi(\mathbf{x}, t) = \omega_{\mathbf{k}} t - \mathbf{k} \cdot \mathbf{x}$ march across space, and the speed of this motion $|\mathbf{c}| = \omega_{\mathbf{k}}/k$ is called the *phase velocity* (though it should really be called the phase-speed).

For *non-dispersive* systems, the phase speed is independent of the wave-number. These systems admit, for instance, 1-dimensional solutions which propagate preserving their wave-form, and also the response to a δ -function source is an outgoing spherical wave with a sharp δ -function pulse profile. For many of the systems mentioned above, however, the phase speed depends on the wavelength. This has the consequence that *wave-packets* travel at a different velocity from the wave-crests — the *group velocity* — and that the response to a δ -function source becomes a ‘chirp’.

D.1 The Group Velocity

The *dispersion relation* is the relation between the spatial and temporal frequencies for a wave, usually as an explicit equation for the frequency ω in terms of the wave-number \mathbf{k} :

$$\omega = \omega(\mathbf{k}). \quad (\text{D.4})$$

We will often use the notation $\omega_{\mathbf{k}} \equiv \omega(\mathbf{k})$. Also, for many systems, the frequency is independent of the direction of the wave, and then $\omega = \omega(k)$. Non-dispersive systems have the ‘trivial’ linear relation $\omega(k) = ck$.

Consider a disturbance which is the sum of two plane waves with wave-vectors \mathbf{k}_1 and \mathbf{k}_2 :

$$\phi(x, t) = e^{i(\omega_{\mathbf{k}_1} t - \mathbf{k}_1 \cdot \mathbf{x})} + e^{i(\omega_{\mathbf{k}_2} t - \mathbf{k}_2 \cdot \mathbf{x})}. \quad (\text{D.5})$$

Now $e^{ia} + e^{ib} = e^{i(a+b)/2} (e^{i(a-b)/2} + e^{-i(a-b)/2}) = 2e^{i(a+b)/2} \cos((a-b)/2)$, so we can write this as

$$\phi(x, t) = 2e^{i(\bar{\omega} t - \bar{\mathbf{k}} \cdot \mathbf{x})} \times \cos\left(\frac{\omega_{\mathbf{k}_1} - \omega_{\mathbf{k}_2}}{2} t - \frac{\mathbf{k}_1 - \mathbf{k}_2}{2} \cdot \mathbf{x}\right) \quad (\text{D.6})$$

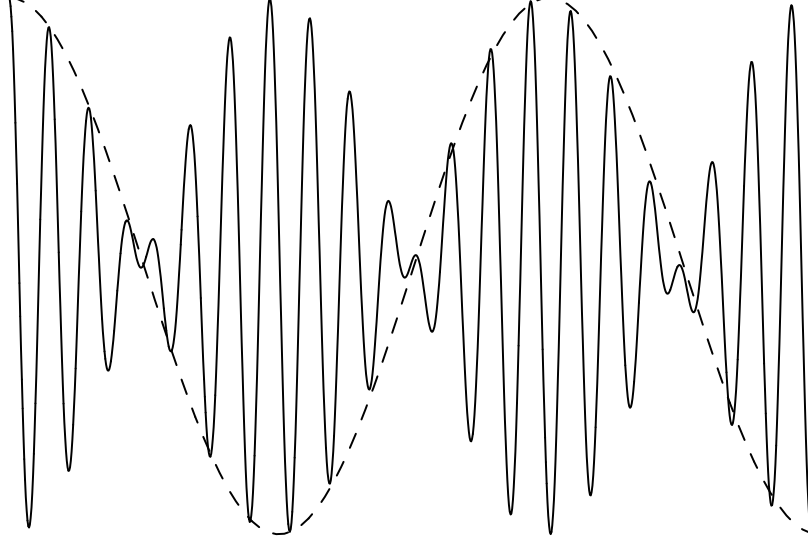


Figure D.1: Wave which is the sum of two neighboring frequencies $\cos(k_1x) + \cos(k_2x)$. This is the product of two sinusoids, one the ‘carrier’ at the mean frequency $\bar{k} = (k_1 + k_2)/2$ and the other the envelope at the modulating frequency $(k_1 - k_2)/2$. This phenomenon is known as *beating*. The wave consists of packets or ‘sets’. The number of waves per set is $\sim \bar{k}/\Delta k$. If the system is dispersive, the modulation pattern travels at a speed $v_{\text{group}} = \Delta\omega/\Delta k$ which differs from the phase velocity.

where $\bar{\omega}$ and $\bar{\mathbf{k}}$ are the average temporal and spatial frequencies. If the two frequencies are similar, so $\mathbf{k}_1 - \mathbf{k}_2 \ll \mathbf{k}_1, \mathbf{k}_2$, then this is a rapidly oscillating plane wave with $e^{i(\bar{\omega}t - \bar{\mathbf{k}} \cdot \mathbf{x})}$ being modulated by a relatively slowly varying amplitude $2\cos((\Delta\omega t - \Delta\mathbf{k} \cdot \mathbf{x})/2)$ as illustrated in figure D.1.

Evidently, this modulation pattern travels at velocity

$$\mathbf{v}_g = \frac{\Delta\omega\Delta\mathbf{k}}{(\Delta k)^2} \quad (\text{D.7})$$

For small Δk this tends to a well-defined limit, for which these sets, or groups, of waves travel at the *group velocity*

$$\mathbf{v}_g = \frac{d\omega(\mathbf{k})}{d\mathbf{k}}. \quad (\text{D.8})$$

For a non-dispersive system, the group- and phase-velocities are the same, but in general they differ. For example:

- For *deep ocean gravity waves*, for instance, the dispersion relation is $\omega_k = \sqrt{gk}$ so $v_g = c/2$. For these waves, the wave-crests march through the sets, appearing at the tail and vanishing at the nose.
- For free *de Broglie waves*, with dispersion relation $\omega(\mathbf{k}) = \hbar k^2/2m$, the phase-velocity is $c = \hbar k/2m$ group velocity is $\mathbf{v}_g = \hbar \mathbf{k}/m = 2c$.
- For electro-magnetic waves in a cold plasma, or for Klein-Gordon waves, with $\omega(\mathbf{k})^2 = c^2 k^2 + m^2 c^4/\hbar^2$, the phase velocity is $v_p = \omega/k = c\sqrt{1 + m^2 c^2/\hbar^2 k^2}$ which diverges as $k \rightarrow 0$, while the group velocity is $\mathbf{v}_g = d\omega/d\mathbf{k} = c^2 \mathbf{k}/\omega$.

D.2 Wave Packets

This argument can be generalized to describe the propagation of waveforms with more complicated profiles. Let us define the real function

$$W(\mathbf{x}) = \sum_{\mathbf{q}} \tilde{W}(\mathbf{q}) e^{i\mathbf{q} \cdot \mathbf{x}}, \quad (\text{D.9})$$

where reality of $W(\mathbf{x})$ requires $\tilde{W}(-\mathbf{q}) = \tilde{W}^*(\mathbf{q})$. Now construct a traveling wave $\phi(\mathbf{x}, t)$ as the sum of modes like (D.1) and with $\phi_{\mathbf{k}} = \tilde{W}(\mathbf{k} - \bar{\mathbf{k}})$:

$$\phi(\mathbf{x}, t) = \sum_{\mathbf{k}} W(\mathbf{k} - \bar{\mathbf{k}}) e^{i(\omega_{\mathbf{k}} t - \mathbf{k} \cdot \mathbf{x})} = e^{i(\bar{\omega} t - \bar{\mathbf{k}} \cdot \mathbf{x})} \sum_{\mathbf{q}} \tilde{W}(\mathbf{q}) e^{i(\Omega_{\mathbf{q}} t - \mathbf{q} \cdot \mathbf{x})} \quad (\text{D.10})$$

where $\mathbf{q} = \mathbf{k} - \bar{\mathbf{k}}$ and

$$\Omega_{\mathbf{q}} = \omega(\bar{\mathbf{k}} + \mathbf{q}) - \omega(\bar{\mathbf{k}}) = q_i \frac{d\omega}{dk_i} + \frac{1}{2} q_i q_j \frac{d^2\omega}{dk_i dk_j} + \dots \quad (\text{D.11})$$

with partial derivatives evaluated at $\mathbf{k} = \bar{\mathbf{k}}$. If we keep only the term of first order in \mathbf{q} here, the wave is

$$\phi(\mathbf{x}, t) = W(\mathbf{x} - \mathbf{v}_g t) e^{i(\bar{\omega} t - \bar{\mathbf{k}} \cdot \mathbf{x})}. \quad (\text{D.12})$$

where $\mathbf{v}_g = d\omega/d\mathbf{k}$ is the group velocity. This is again a rapid oscillation at the carrier frequency $\bar{\mathbf{k}}$ modulated by the moving envelope function $W(\mathbf{x} - \mathbf{v}_g t)$.

This therefore describes a *wave packet* with profile $W(\mathbf{x})$ which propagates along at the group velocity, preserving its shape. Now small \mathbf{q} means a large packet, since the values of \mathbf{q} for which $\tilde{W}(\mathbf{q})$ are appreciable are on the order of $q \sim 1/L$, with L the extent of the packet. Thus we expect this non-evolving packet approximation to become better the larger the packet. For small packets, or for long times, the packet will evolve. We can estimate the evolution time-scale for the wave-packet if we keep the next order term in the Taylor expansion (D.11). We then have

$$\phi(\mathbf{x}, t) = W(\mathbf{x} - \mathbf{v}_g t, t) e^{i(\bar{\omega} t - \bar{\mathbf{k}} \cdot \mathbf{x})}. \quad (\text{D.13})$$

with

$$W(\mathbf{x}, t) = \sum_{\mathbf{q}} \tilde{W}(\mathbf{q}) e^{iq_i q_j (d^2\omega/dk_i dk_j) t/2} e^{i\mathbf{q} \cdot \mathbf{x}}. \quad (\text{D.14})$$

The transform of the wave-envelope in this approximation then becomes

$$\tilde{W}(\mathbf{q}, t) = \tilde{W}(\mathbf{q}) e^{iq_i q_j (d^2\omega/dk_i dk_j) t/2}, \quad (\text{D.15})$$

so the evolution of the packet profile is small for times $t \ll (q^2 d^2\omega/dk^2)^{-1}$. The evolution of a small wave packet is shown in figure D.2. We can identify three time-scales here; τ_{phase} , τ_{cross} and τ_{evol} , being the inverse of the carrier frequency, the time for the packet to pass through its own length, and the time for the wave-packet to evolve. These are

$$\tau_{\text{phase}} \sim \frac{1}{\omega} \quad \tau_{\text{cross}} \sim \frac{L}{v_{\text{group}}} \sim \frac{1}{q d\omega/dk} \quad \tau_{\text{evol}} \sim \frac{1}{q^2 d^2\omega/dk^2} \quad (\text{D.16})$$

We have used here $L \sim 1/q$. For the deep ocean waves, for example, where $\omega(k) = \sqrt{gk}$ these time-scales are in the ratio $1 : \bar{k}/q : (\bar{k}/q)^2$. But $\bar{k}/q \sim N$, the number of waves in the packet, so a packet containing say 100 waves will take ~ 100 times the wave period to pass through its length, but will propagate ~ 100 times its length before substantially changing its shape. For a great many naturally occurring dispersion relations $\tau_{\text{evol}}/\tau_{\text{cross}} \sim \bar{k}/q$, so large wave packets ($L \gg 1/\bar{k}$) persist for a long time.

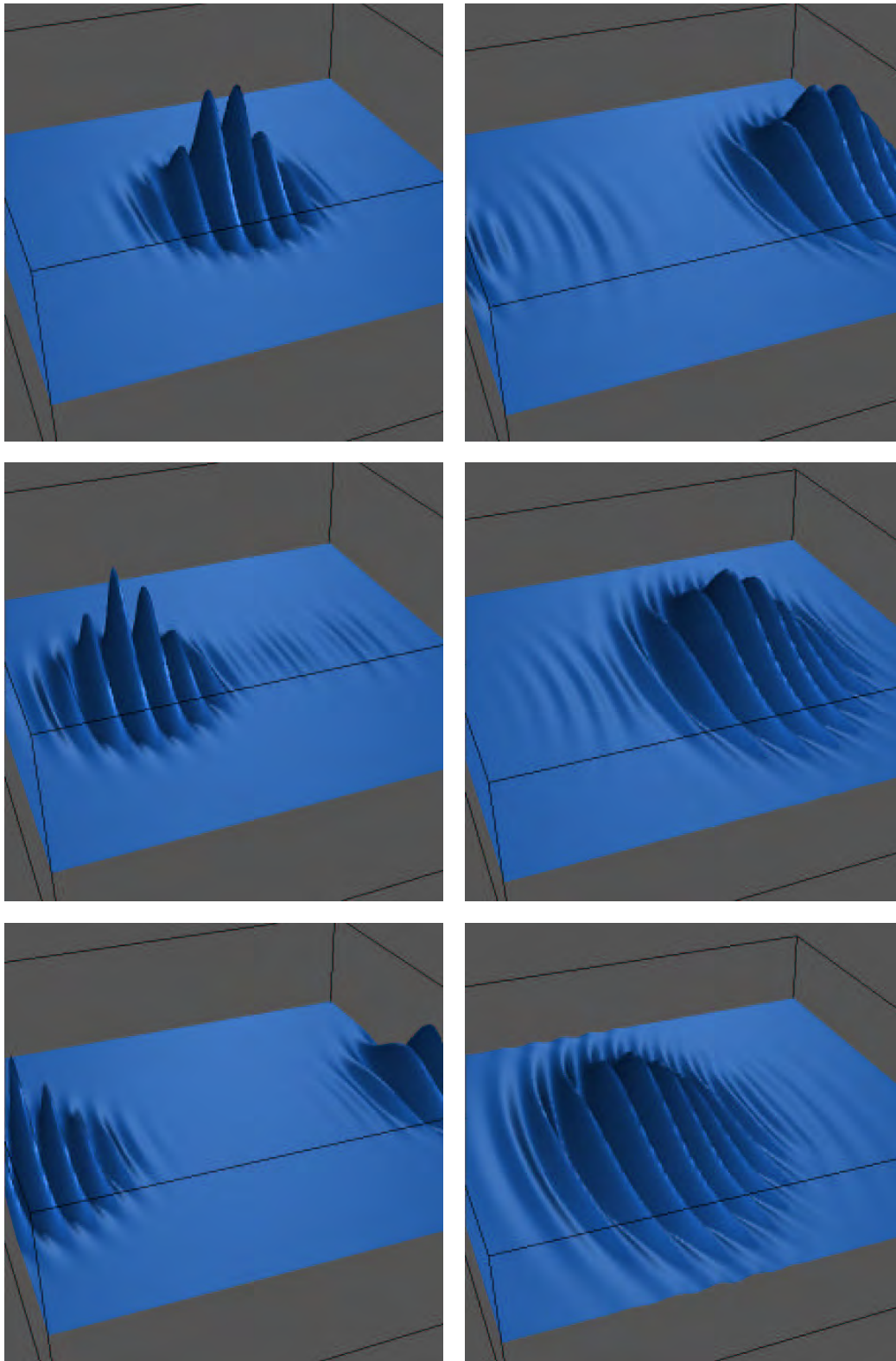


Figure D.2: Evolution of a wave packet. Initial packet is at top left and is moving to the left. Final packet, after propagating once around the (periodic) box is at lower right. The spreading of the packet is evident. The dispersion relation is that of deep ocean waves.

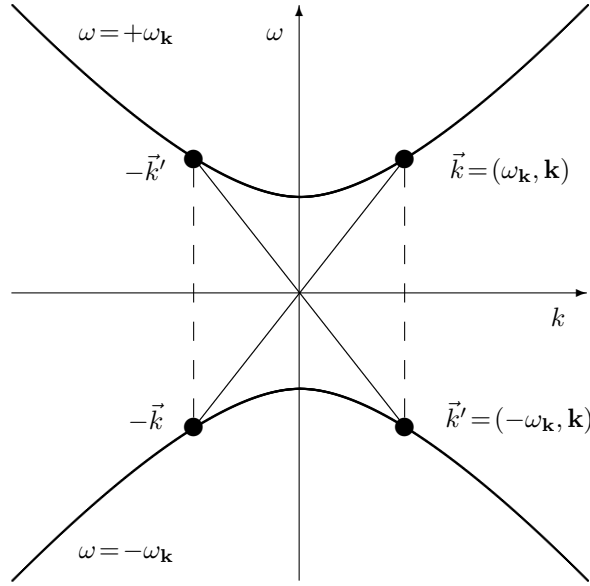


Figure D.3: Solutions of free-field wave equations have a 4-dimensional Fourier transform which is confined to a pair of 3-dimensional hyper-surfaces $\omega = \pm\omega_{\mathbf{k}}$. For a non-dispersive wave these are a pair of cones which meet at the origin. Here we show the form of the hyper-surface for dispersion relation like that of a massive field: $\omega_{\mathbf{k}} = \sqrt{c^2 k^2 + \mu^2}$.

D.3 Evolution of Dispersive Waves

A common problem is to evolve some system of wave equations forward in time given some initial field configuration. In general, one needs to cast the equations into some discretized form and then step these equations forward. However, for non-interacting wave equations, life is much simpler since the general field is a sum of traveling wave solutions like (D.1). Thus, to evolve from some initial field configuration at $t = 0$ to some later time we need only determine the coefficients $\phi_{\mathbf{k}}$, and then multiply these by the complex phase factor $e^{i\omega_{\mathbf{k}}t_f - \mathbf{k} \cdot \mathbf{x}}$ to synthesize the field $\phi(\mathbf{x}, t_f)$. One might guess that the coefficients $\phi_{\mathbf{k}}$ are just the Fourier transform of the initial field $\phi(\mathbf{x}, t = 0)$. However, this is not quite correct; consider the waves $\phi_1 = \cos(\omega t - \mathbf{k} \cdot \mathbf{x})$ and $\phi_2 = \cos(-\omega t - \mathbf{k} \cdot \mathbf{x})$. These waves travel in directions \mathbf{k} and $-\mathbf{k}$ respectively, but have identical transform at $t = 0$. However, these fields have different time derivative at $t = 0$, so if we supply $\phi(\mathbf{x}, t = 0)$ and $\dot{\phi}(\mathbf{x}, t = 0)$ then this should suffice to determine both the future evolution. This is not too surprising; these wave-like solutions often arise as the solution of second order differential equations for which, to fully specify the initial state, we need to specify both the initial displacement $\phi(\mathbf{x}, 0)$ and the initial velocity $\dot{\phi}(\mathbf{x}, 0)$.

A general field in space time $\phi(\vec{x}) = \phi(\mathbf{x}, t)$ has a 4-dimensional Fourier transform

$$\tilde{\phi}(\vec{k}) = \int d^4x \phi(\vec{x}) e^{i\vec{k} \cdot \vec{x}} \quad (\text{D.17})$$

with inverse transform

$$\phi(\vec{x}) = \int \frac{d^4k}{(2\pi)^4} \tilde{\phi}(\vec{k}) e^{-i\vec{k} \cdot \vec{x}}. \quad (\text{D.18})$$

For a free-field, however, the transform vanishes except on the hyper-surfaces $\omega = \pm\omega_{\mathbf{k}}$ (see figure D.3). The modes $\tilde{\phi}(\vec{k})$ and $\tilde{\phi}(-\vec{k})$ together describe a disturbance propagating in the direction \mathbf{k} , while $\tilde{\phi}(\vec{k}')$ and $\tilde{\phi}(-\vec{k}')$ propagate in the direction $-\mathbf{k}$. For a real field $\phi(\vec{x})$, $\tilde{\phi}(-\vec{k}) = \tilde{\phi}^*(\vec{k})$, so the negative energy modes are fully determined once we specify the positive energy mode amplitudes. To specify the field then we need to provide the real and imaginary components of the complex field

$\tilde{\phi}(\omega_{\mathbf{k}}, \mathbf{k})$ at each point in \mathbf{k} -space. This corresponds to the need to provide both $\phi(\mathbf{x}, t = 0)$ and $\dot{\phi}(\mathbf{x}, t = 0)$ at each point in real space.

Let us write the field as a sum over positive and negative frequency components:

$$\phi(\mathbf{x}, t) = \int \frac{d^3k}{(2\pi)^3} \phi^+(\mathbf{k}) e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} + \int \frac{d^3k}{(2\pi)^3} \phi^-(\mathbf{k}) e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} \quad (\text{D.19})$$

Taking the spatial Fourier transforms of ϕ and $\dot{\phi}$ at $t = 0$, we find

$$\tilde{\phi}(\mathbf{k}) = \int d^3x \phi(\mathbf{x}, 0) e^{i\mathbf{k} \cdot \mathbf{x}} = \phi^+(\mathbf{k}) + \phi^-(-\mathbf{k}) \quad (\text{D.20})$$

and

$$\tilde{\dot{\phi}}(\mathbf{k}) = \int d^3x \dot{\phi}(\mathbf{x}, 0) e^{i\mathbf{k} \cdot \mathbf{x}} = i\omega_{\mathbf{k}}\phi^+(\mathbf{k}) - i\omega_{\mathbf{k}}\phi^-(-\mathbf{k}). \quad (\text{D.21})$$

Solving for $\phi^+(\mathbf{k})$ and $\phi^-(\mathbf{k})$ gives

$$\phi^+(\mathbf{k}) = \frac{1}{2} \left[\tilde{\phi}(\mathbf{k}) + \frac{\tilde{\dot{\phi}}(\mathbf{k})}{i\omega_{\mathbf{k}}} \right] \quad (\text{D.22})$$

and

$$\phi^-(\mathbf{k}) = \frac{1}{2} \left[\tilde{\phi}(-\mathbf{k}) + \frac{\tilde{\dot{\phi}}(-\mathbf{k})}{i\omega_{\mathbf{k}}} \right]. \quad (\text{D.23})$$

For example, consider a 1-dimensional, initially static, disturbance

$$\begin{aligned} \phi(x, y, z, t = 0) &= f(x) \\ \dot{\phi}(x, y, z, t = 0) &= 0. \end{aligned} \quad (\text{D.24})$$

Evaluating the transforms and using the above pair of equations gives

$$\phi_+(\mathbf{k}) = \phi_-(-\mathbf{k}) = \delta(k_y)\delta(k_z)\tilde{f}(k_x)/2. \quad (\text{D.25})$$

The future development of the field is

$$\phi(\mathbf{x}, t) = \frac{1}{2} \int \frac{dk_x}{2\pi} \tilde{f}(k_x) e^{i(\omega_{\mathbf{k}}t - k_x x)} + \text{c.c} \quad (\text{D.26})$$

For a non-dispersive system (i.e. one for which $\omega_{\mathbf{k}} = ck$) this is

$$\phi(\mathbf{x}, t) = \frac{1}{2} \int \frac{dk_x}{2\pi} \tilde{f}(k_x) e^{ik_x(x-ct)} + \text{c.c} = \frac{f(x+ct)}{2} + \frac{f(x-ct)}{2}. \quad (\text{D.27})$$

As another example, consider a localized impulse which we model as $\phi(\mathbf{x}, 0) = \delta(\mathbf{x})$, $\dot{\phi}(\mathbf{x}, 0) = 0$. Now we have $\phi^+(\mathbf{k}) = \phi^-(\mathbf{k}) = \text{constant}$ and the future development of the wave is

$$\phi(\mathbf{x}, t) \propto \int d^3k e^{i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} + \int d^3k e^{-i(\omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x})} \quad (\text{D.28})$$

For given \mathbf{x} , t the major contribution to say the first integral here comes from the region of \mathbf{k} -space where the phase $\psi(\mathbf{k}; \mathbf{x}, t) = \omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x}$ is stationary, i.e. where $\partial\psi/\partial\mathbf{k} = 0$. Performing the derivative, we see that the phase is stationary for wave number $\mathbf{k} = \mathbf{k}_0(\mathbf{x}, t)$ such that $\mathbf{v}_g(\mathbf{k}_0) = \mathbf{x}/t$. In the vicinity of this stationary point, we can expand the phase as

$$\psi(\mathbf{k}; \mathbf{x}, t) \simeq \omega_{\mathbf{k}_0}t - \mathbf{k}_0 \cdot \mathbf{x} + \frac{1}{2} q_i q_j \frac{d^2\omega_{\mathbf{k}}}{dk_i dk_j} t \quad (\text{D.29})$$

where $\mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0$ and the partial derivatives are to be evaluated at $\mathbf{k} = \mathbf{k}_0(\mathbf{x}, t)$. In this approximation the wave solution is

$$\phi(\mathbf{x}, t) \sim e^{i(\omega_{\mathbf{k}_0} t - \mathbf{k}_0 \cdot \mathbf{x})} \int d^3 q \exp \left(\frac{i}{2} q_i q_j \frac{d^2 \omega_{\mathbf{k}}}{dk_i dk_j} t \right) \quad (\text{D.30})$$

with a similar contribution from the second integral. The integral here a slowly varying envelope which modulates the relatively rapidly varying complex exponential factor. At fixed position \mathbf{x} , the wave has a sinusoidal variation with time varying frequency $\omega(\mathbf{k}_0(\mathbf{x}, t))$ and with time varying amplitude. This is called a ‘chirp’.

Specializing to the case of an isotropic dispersion relation $\omega(\mathbf{k}) = \omega(k)$, the quadratic form appearing here is

$$q_i q_j \frac{d^2 \omega_{\mathbf{k}}}{dk_i dk_j} = \frac{1}{k} \frac{d\omega}{dk} q_{\perp}^2 + \frac{d^2 \omega}{dk^2} q_{\parallel}^2 \quad (\text{D.31})$$

where $\mathbf{q}_{\parallel} \equiv (\hat{\mathbf{k}} \cdot \mathbf{q}) \hat{\mathbf{k}}$ and $\mathbf{q}_{\perp} \equiv \mathbf{q} - \mathbf{q}_{\parallel}$. The value of the integral is then $\sim k(d\omega/dk)^{-1}(d^2\omega/dk^2)^{-1/2}$, and similarly in two dimensions it is $\sim \sqrt{k/(d\omega/dk)(d^2\omega/dk^2)}$. As an application, consider deep ocean waves for which $\omega = \sqrt{gk}$. The group velocity is $d\omega/dk = \frac{1}{2}\sqrt{g/k}$ and the stationary phase condition $\mathbf{v}_g(\mathbf{k}_0) = \mathbf{x}/t$ gives $\mathbf{k}_0(\mathbf{x}, t) = gt^2\mathbf{x}/4x^3$ and the frequency is $\omega_0(\mathbf{x}, t) = \omega(\mathbf{k}_0) = gt/2x$. The determinant is $|d^2\omega/dk_i dk_j| \sim g/k_0^3$. The envelope is

$$\int d^2 q \exp \left(\frac{i}{2} q_i q_j \frac{d^2 \omega_{\mathbf{k}}}{dk_i dk_j} t \right) \propto k_0(x, t)^{3/2} t^{-1} \quad (\text{D.32})$$

so the wave develops as

$$\phi(x, t) \sim k_0^{3/2} t^{-1} e^{i\omega_0 t - \mathbf{k}_0 \cdot \mathbf{x}} \sim \frac{t^2}{x^3} \cos(gt^2/4x). \quad (\text{D.33})$$

Had we instead applied an initial impulsive velocity $\dot{\phi}(\mathbf{x}, t=0) = \delta(\mathbf{x})$ with zero displacement $\phi(\mathbf{x}, t=0) = 0$ then this generates a spectrum $|\phi^+(k)|^2 \propto 1/\omega^2 \propto 1/k$ rather than a flat spectrum. Similarly, if the initial impulse is not point-like but is extended over some region of size l then this will give a cut-off in the power spectrum at $k \gtrsim 1/l$. For example, for a Gaussian velocity impulse with profile $\dot{\phi}(\mathbf{x}, 0) \propto \exp(-x^2/2\sigma^2)$ the wave is

$$\phi(x, t) \sim \frac{t^1}{x^2} \exp(-g^2 t^4 \sigma^2 / 8x^4) \cos(gt^2/4x). \quad (\text{D.34})$$

This is plotted in figure D.4. The 2-dimensional disturbance generated from a localised random disturbance is shown in figure D.5

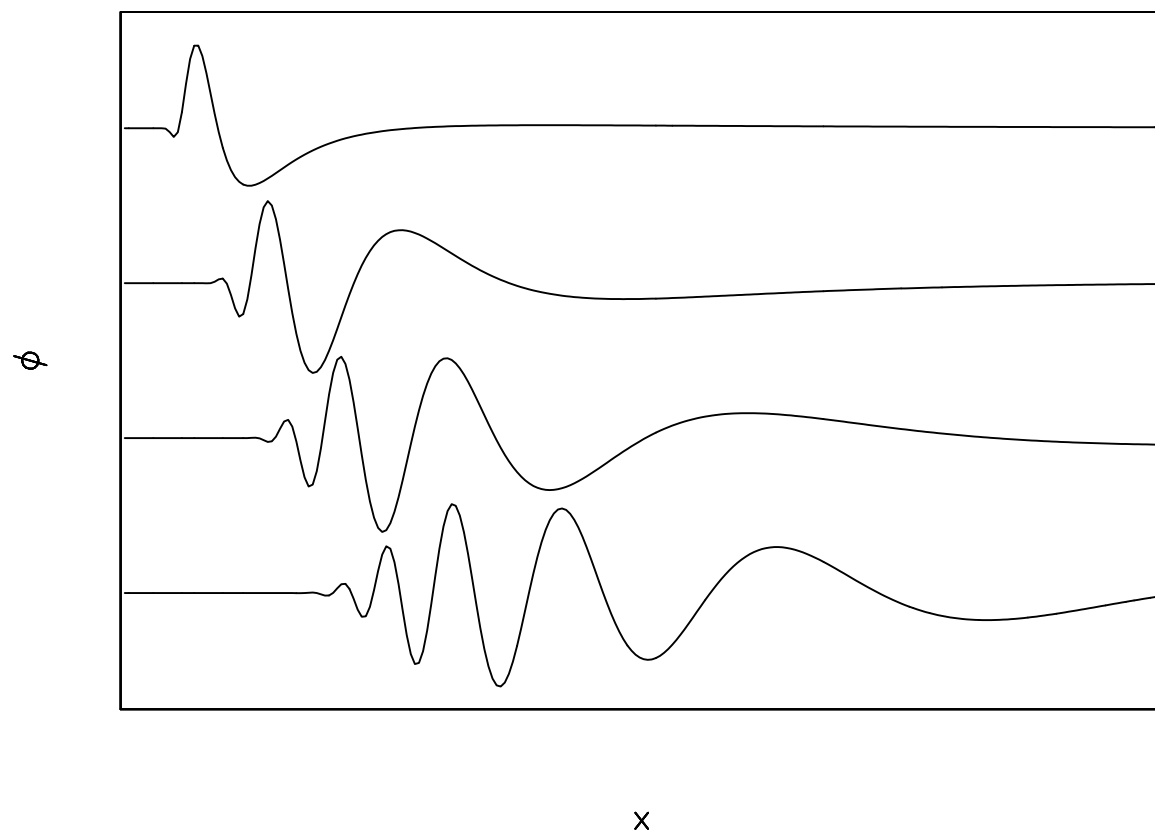


Figure D.4: Evolution of a spherical deep ocean water wave. The quantity plotted is $\phi = (t^2/x^2) \cos(t^2/4x) \exp(-t^4/8x^4)$, for $t = 15, 30, 45, 60$. This is the ‘chirp’ generated by an impulse with a Gaussian profile with scale $\sigma = 1$.

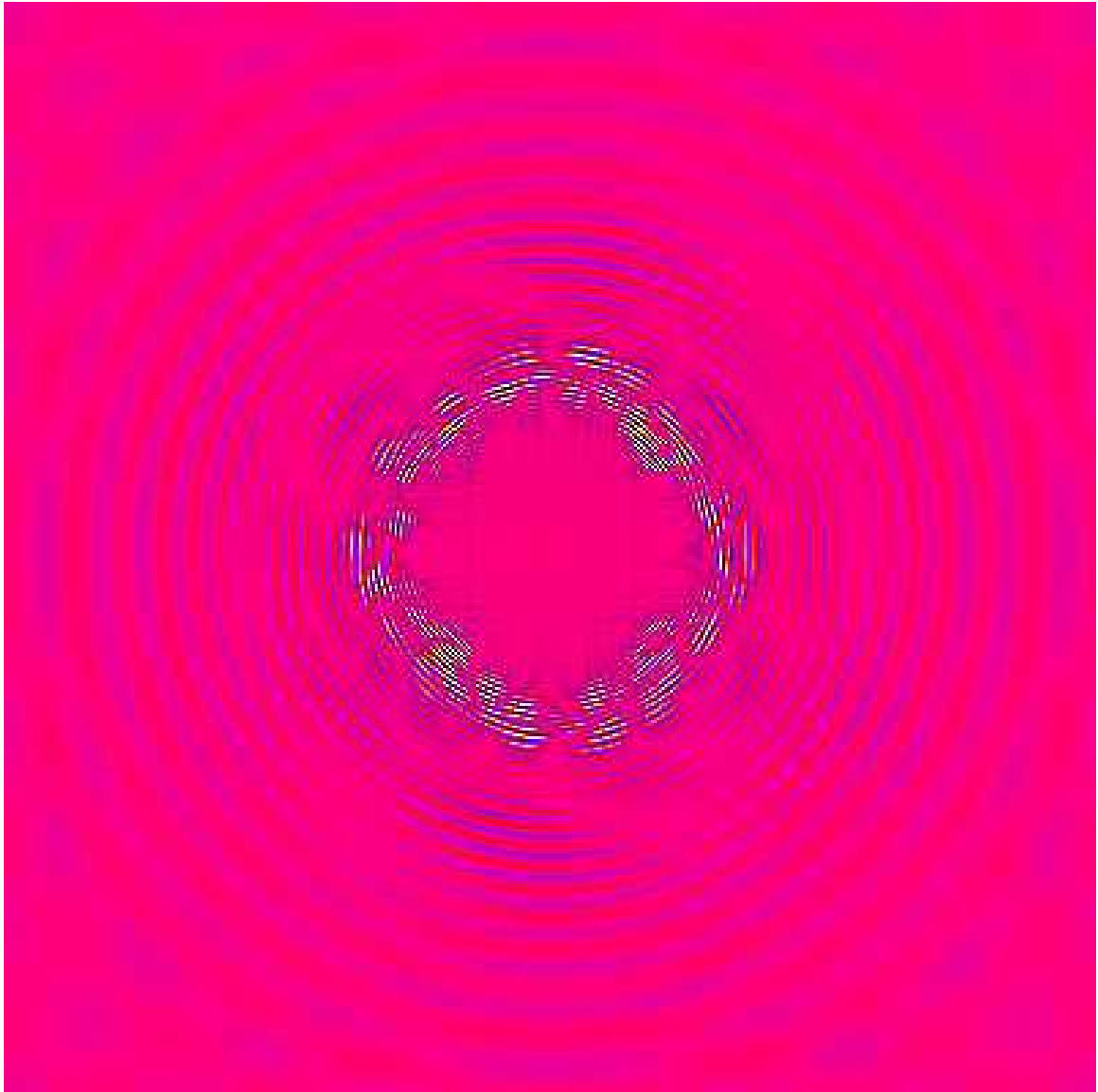


Figure D.5: Evolution of deep ocean water waves from a localised ‘storm’. The initial conditions were uniform aside from a small region near the origin in which the field was given a random ‘white noise’ displacement. This was then evolved as described in the text.

Appendix E

Relativistic Covariance of Electromagnetism

Here we review the relativistic formulation of the laws of electromagnetism.

The electronic charge e is a constant and is the same in all frames of reference. This means that the charge density $\rho = en_e$ transforms so that ρ/E is conserved, ie ρ transforms like the time-component of a four-vector. Similarly, the current density \mathbf{j} transforms like the spatial components of a 4-vector so we can combine ρ and \mathbf{j} to form the *four-current*

$$\vec{j} = j^\mu = (c\rho, \mathbf{j}). \quad (\text{E.1})$$

Some features of the transformation of the four-current are quite familiar. For example, it is clear that a static charged rod, if put into motion becomes a current. Less familiar perhaps is the fact that an electrically neutral rod, but containing a current in the form of electrons moving against a static background of ions will, in a boosted frame, have a non-zero charge density. This has to do with the fact that the density of a set of particles depends on the frame of reference.

The equation of charge conservation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0 \quad \Longleftrightarrow \quad j^\mu{}_{,\mu} = 0 \quad (\text{E.2})$$

which is covariant (it says that the scalar quantity $j^\mu{}_{,\mu}$, the four-divergence of the four-current, vanishes).

Maxwell's equations in terms of the potential can also be expressed in a covariant manner:

$$\left. \begin{aligned} \square \mathbf{A} &= -4\pi \mathbf{j}/c \\ \square \phi &= -4\pi \rho \end{aligned} \right\} \quad \Longleftrightarrow \quad A^{\mu,\nu}{}_{,\nu} = -\frac{4\pi}{c} j^\mu \quad (\text{E.3})$$

with

$$\vec{A} = A^\mu = (\phi, \mathbf{A}) \quad (\text{E.4})$$

another four-vector.

The Lorentz gauge condition can also be expressed in a covariant manner:

$$\nabla \cdot \mathbf{A} + \dot{\phi}/c = 0 \quad \Longleftrightarrow \quad A^\mu{}_{,\mu} = 0. \quad (\text{E.5})$$

It turns out that the fields \mathbf{E} and \mathbf{B} are the non-vanishing parts of the *Faraday tensor* $F_{\mu\nu}$ which is the anti-symmetrized derivative of the four-potential: $F_{\mu\nu} = A_{[\mu,\nu]} = A_{\mu,\nu} - A_{\nu,\mu}$ with components

$$F_{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{bmatrix}. \quad (\text{E.6})$$

The transformation law for the \mathbf{E} , \mathbf{B} is

$$\begin{aligned} \mathbf{E}'_{\parallel} &= \mathbf{E}_{\parallel} & \mathbf{B}'_{\parallel} &= \mathbf{B}_{\parallel} \\ \mathbf{E}'_{\perp} &= \gamma(\mathbf{E}_{\perp} + \boldsymbol{\beta} \times \mathbf{B}) & \mathbf{B}'_{\perp} &= \gamma(\mathbf{B}_{\perp} - \boldsymbol{\beta} \times \mathbf{E}). \end{aligned} \quad (\text{E.7})$$

see R+L for simple physical examples illustrating these transformation laws.

The relativistically correct expression of the Lorentz force law is

$$\frac{dU^{\mu}}{d\tau} = \frac{q}{mc} F^{\mu}{}_{\nu} U^{\nu} \quad (\text{E.8})$$

whose time and space components can be written as

$$\begin{aligned} \frac{dP^0}{dt} &= \frac{1}{\gamma} \frac{dP^0}{d\tau} = q\mathbf{E} \cdot \mathbf{v} \\ \frac{d\mathbf{P}}{dt} &= \frac{1}{\gamma} \frac{d\mathbf{P}}{d\tau} = q[\mathbf{E} + (\mathbf{v} \times \mathbf{B})/c] \end{aligned} \quad (\text{E.9})$$

which is identical to the non-relativistic form, except that the momentum here is the relativistic momentum $\mathbf{P} = \gamma m \mathbf{v}$.

E.1 EM Field of a Rapidly Moving Charge

In the rest frame of a charge (primed frame) the electric potential is $\phi' = q/r'$ and the magnetic potential \mathbf{A}' vanishes, so applying a boost along the x -axis gives the potential in the lab-frame (unprimed frame)

$$A^{\mu} = \Lambda^{\mu}{}_{\nu} A'^{\nu} = \begin{bmatrix} \gamma q/r' \\ -\gamma \beta q/r' \\ 0 \\ 0 \end{bmatrix} \quad (\text{E.10})$$

so

$$\vec{A} = \begin{bmatrix} \phi \\ \mathbf{A} \end{bmatrix} = \begin{bmatrix} \gamma q/r' \\ -\beta \gamma \hat{\mathbf{x}} q/r' \end{bmatrix} \quad (\text{E.11})$$

with

$$r' = \sqrt{x'^2 + y'^2 + z'^2} = \sqrt{\gamma^2(x - vt)^2 + y^2 + z^2} \quad (\text{E.12})$$

Computing the electric field $\mathbf{E} = -\dot{\mathbf{A}}/c - \nabla\phi$ yields

$$\mathbf{E} = \frac{\gamma q}{r'^3} \begin{bmatrix} x - vt \\ y \\ z \end{bmatrix}. \quad (\text{E.13})$$

For a field point at $x = z = 0$, $y = b$, we have $E_z = 0$ and

$$E_y(t) = \frac{\gamma q b}{(\gamma^2 v^2 t^2 + b^2)^{3/2}} \quad \text{and} \quad E_x(t) = \frac{\gamma q v t}{(\gamma^2 v^2 t^2 + b^2)^{3/2}}. \quad (\text{E.14})$$

If we define the dimensionless time $y = \gamma v t/b$ then we can write

$$\frac{b^2 E_y(t)}{\gamma q} = \frac{1}{(1 + y^2)^{3/2}} \quad \text{and} \quad \frac{b^2 E_x(t)}{\gamma q} = \frac{y}{\gamma (1 + y^2)^{3/2}} \quad (\text{E.15})$$

These functions are shown in figure E.1. Both of these are impulses of extent $\Delta y \sim 1 \rightarrow \Delta t \sim b/\gamma v$. For highly relativistic motion this means that the electric field configuration is not spherical (in which case the period of the impulse would be $\Delta t \sim b/v$) but is compressed along the direction of motion by a factor γ . Also, for highly relativistic motion $E_x \ll E_y$. The y -component has maximum value $E_{\max} = \gamma q/b^2$.

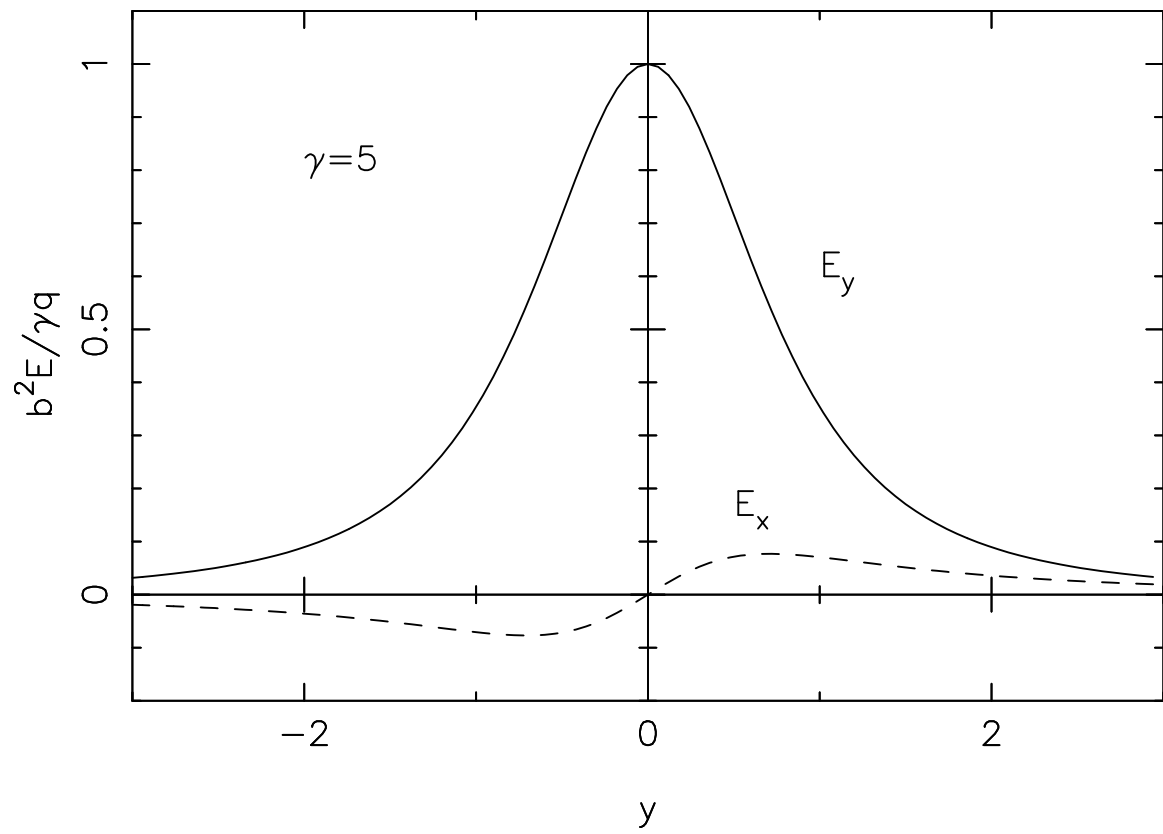


Figure E.1: Electric field of a rapidly moving charge plotted versus dimensionless time coordinate $y = \gamma vt/b$ for Lorentz factor $\gamma = 5$.

Appendix F

Complex Analysis

F.1 Complex Numbers and Functions

- An example of a *complex number* is $z = x + iy$, where x, y are real and $i = \sqrt{-1}$.
- A complex number corresponds to a point (x, y) on the *Argand Diagram*.
- The complex number z can also be represented in ‘polar coordinates’ as $z = ae^{i\varphi} = a(\cos(\varphi) + i\sin(\varphi))$, with $a = \sqrt{x^2 + y^2}$ and $\tan(\varphi) = y/x$.
- Complex numbers can be added like 2-vectors, but unlike ordinary 2-vectors can also be directly multiplied, $zz' = aa'e^{i(\varphi+\var')}$, and divided $z/z' = (a/a')e^{i(\varphi-\var')}$.
- Multiplying a complex number z' by $z = ae^{i\varphi}$ increases its length by a factor a and rotates it in the Argand plane by an angle φ .
- A *complex function* $f(z)$ associates with each point z a complex number $f(z)$.

F.2 Analytic Functions

A complex function $f(z)$ can be written as

$$f(z) = u(x, y) + iv(x, y) \quad (\text{F.1})$$

where $u(x, y)$ and $v(x, y)$ are real.

A very special, but nonetheless extremely important, class of complex functions are *analytic complex functions* defined as those for which

$$\begin{aligned} \partial u / \partial x &= \partial v / \partial y \\ \partial u / \partial y &= -\partial v / \partial x. \end{aligned} \quad (\text{F.2})$$

These are called the *Cauchy-Riemann conditions*.

If we use the notation

$$f = u + iv = \begin{bmatrix} u \\ v \end{bmatrix} \quad (\text{F.3})$$

the differential of an arbitrary complex function $f(z) = u + iv$ is

$$\Delta f = \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} \partial u / \partial x & \partial u / \partial y \\ \partial v / \partial x & \partial v / \partial y \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}. \quad (\text{F.4})$$

Now if $f(z)$ is analytic we can use the Cauchy-Riemann conditions to eliminate $\partial v / \partial x$, $\partial v / \partial y$ to give

$$\Delta f = \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} \partial u / \partial x & \partial u / \partial y \\ -\partial u / \partial y & \partial u / \partial x \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}. \quad (\text{F.5})$$

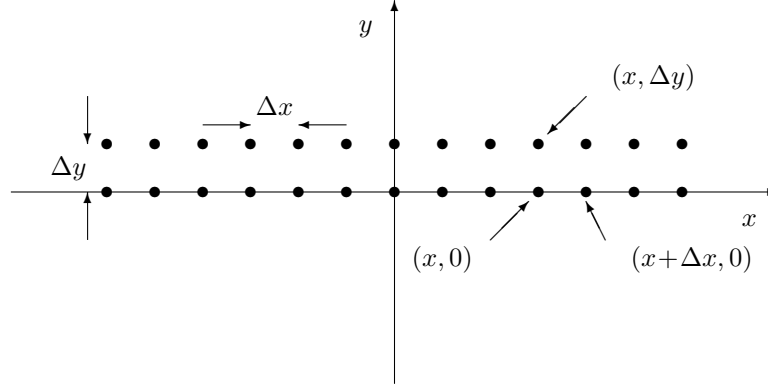


Figure F.1: If an analytic function $f(z) = u(x, y) + iv(x, y)$ is known at all points along the x -axis then the value at some point $(x, \Delta y)$ is given in terms of the values of u and v at the neighboring points on the x -axis.

This is rather interesting; the matrix appearing here looks a lot like a rotation matrix: $R = \{\{\cos \theta, -\sin \theta\}, \{\sin \theta, \cos \theta\}\}$. Indeed, if we define $a' = \sqrt{(\partial u / \partial x)^2 + (\partial u / \partial y)^2}$ and $\tan(\varphi') = (\partial u / \partial y) / (\partial u / \partial x)$ the differential of f becomes

$$\Delta f = \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = a' \begin{bmatrix} \cos(\varphi') & -\sin(\varphi') \\ \sin(\varphi') & \cos(\varphi') \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (\text{F.6})$$

But this is just a multiplication of $\Delta z = \Delta x + i\Delta y$ by a factor a' and a rotation through an angle φ' , i.e. it is a multiplication of Δz by another complex number $f' = a'e^{i\varphi'}$:

$$\Delta f = f' \Delta z \quad (\text{F.7})$$

Thus an analytic function $f(z)$ has derivative

$$f'(z) = \lim_{\Delta z \rightarrow 0} \frac{\Delta f}{\Delta z} \quad (\text{F.8})$$

which is also a complex function.

F.3 Analytic Continuation

A general complex function $f(z) = u(x, y) + iv(x, y)$ is clearly a two dimensional complex function; we need to specify the values of the real and imaginary parts at all points on the plane. Analytic functions have the magical property that they are essentially one dimensional, in the sense that if the values of u, v are known along some line in the Argand diagram then they are known everywhere.

To see how this comes about, let's discretize the Argand plane with a grid spacing $\Delta x = \Delta y$, as illustrated in figure F.1. We will assume that the grid spacing is very small compared to the scale of any variations in u or v . Let's also assume that u and v are known at all points along the x -axis. The values of u and v at some point $z = (x, \Delta y)$ (i.e a distance Δy above the point $z = (x, 0)$) are

$$\begin{aligned} u(x, \Delta y) &= u(x, 0) + \partial u / \partial y \Delta y = u(x, 0) - \partial v / \partial x \Delta y \\ v(x, \Delta y) &= v(x, 0) + \partial v / \partial y \Delta y = v(x, 0) + \partial u / \partial x \Delta y \end{aligned} \quad (\text{F.9})$$

where we have used the Cauchy-Riemann conditions (F.2). But with $\partial u / \partial x = (u(x + \Delta x, 0) - u(x, 0)) / \Delta x$ etc. and $\Delta x = \Delta y$ these are

$$\begin{aligned} u(x, \Delta y) &= u(x, 0) - v(x + \Delta x, 0) + v(x, 0) \\ v(x, \Delta y) &= v(x, 0) + u(x + \Delta x, 0) - u(x, 0) \end{aligned} \quad (\text{F.10})$$

Since all of the values on the right hand side are known for all x , this determines the values of $f(z)$ for all x along the line $y = \Delta y$. These in turn determine $f(z)$ along the line $y = 2\Delta y$ and so on. If we let Δx , Δy tend to zero, this becomes exact, and we can thereby *analytically continue* the function $f(z)$ off the real axis, or indeed off any line across the Argand plane.

The foregoing is valid provided $f(z)$ is analytic, but will cease to apply in regions where $f(z)$ does not obey the Cauchy-Riemann conditions (F.2). As we shall see, there are complex functions which are analytic except at one or more points in the Argand plane where the function $f(z)$ becomes infinite. These points are known as the *poles* of the function $f(z)$.

F.4 Contour Integration

The great utility of complex analysis for astrophysicists is that it allows us to evaluate, or provide analytic formulae for, integrals occurring in physical problems. The essential result used here is that the value of a *contour integral*, $I = \oint dz f(z)$ taken around some closed loop in the Argand plane is invariant under deformation of the integration path (provided the path is not forced to cross over a pole).

We can see how this applies in the foregoing example. Let $I(y_0)$ denote the integral of f along the line $y = y_0$ in the Argand plane, so for this path $dz = dx$. In the discretized complex plane, for $y_0 = 0$ this is

$$I(0) = \int dz f(z) = \sum_x \Delta x (u(x, 0) + iv(x, 0)) \quad (\text{F.11})$$

but the integral $I(\Delta y)$ is identical since, with $u(x, \Delta y)$, $v(x, \Delta y)$ given by (F.10) the terms involving v in the u integration cancel and *vice versa*. The value of $I(y_0)$ is therefore independent of which contour we use.

More generally, in the vicinity of the origin we have

$$f(z) \simeq f(0) + \frac{\partial u}{\partial x}x + \frac{\partial u}{\partial y}y + i\frac{\partial v}{\partial x}x + i\frac{\partial v}{\partial y}y, \quad (\text{F.12})$$

so with $dz = dx + idy$ we have

$$f(z)dz = \left(\frac{\partial u}{\partial x}x + \frac{\partial u}{\partial y}y + i\frac{\partial v}{\partial x}x + i\frac{\partial v}{\partial y}y \right) (dx + idy). \quad (\text{F.13})$$

The partial derivatives are to be considered constant here. Now when we perform the integration $\oint dz f(z)$ around a closed loop, terms like $(\partial u/\partial x) \oint dx x$ vanish, since $x dx = dx^2/2$ is a total derivative, so $\oint dx^2 = 0$. The only terms which can contribute are terms like $(\partial u/\partial y) \oint dx y$, which, in general, would give a contribution proportional to the area within the contour. However, for an analytic function the Cauchy-Riemann conditions (F.2) allow us to replace $\partial v/\partial y$ with $\partial u/\partial x$ and $\partial v/\partial x$ with $-\partial u/\partial y$, and we have

$$\oint dz f(z) = \left[\frac{\partial u}{\partial y} - \frac{\partial u}{\partial x} \right] \left(\oint dx y + \oint dy x \right) = \left[\frac{\partial u}{\partial y} - \frac{\partial u}{\partial x} \right] \oint d(xy) = 0. \quad (\text{F.14})$$

However, the choice of origin was arbitrary, so we find that, quite generally, $\oint dz f(z)$ vanishes for any small loop (provided $f(z)$ is analytic on and within the loop).

If we have some large loop then we can slightly deform it to obtain the integral around a new large loop plus the integral around a small closed loop, this being the difference between the two large loops, and which vanishes. Repeating this process allows us to continuously deform a contour integral without changing its value. If a function is analytic everywhere within some loop, for example, then the loop can be shrunk to zero and it follows that the integral around the original loop vanishes.

However, what if the loop contains a pole? For example, consider $\oint dz f$ with $f(z) = f'(z)/(z - z_0)$ where $f'(z)$ is analytic and where the path encircles the point z_0 as illustrated in figure F.2. The

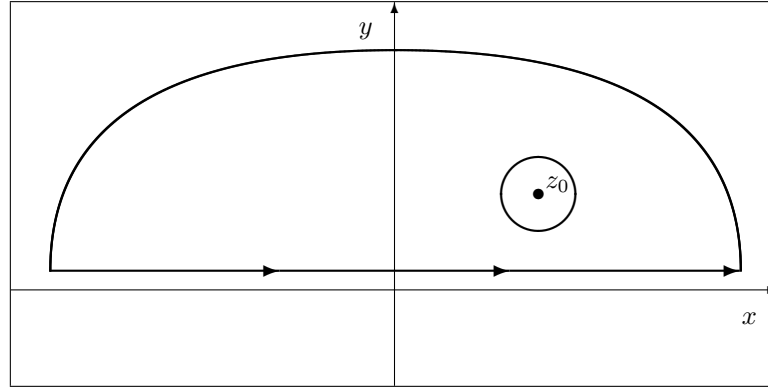


Figure F.2: Example of a contour integral.

function $f(z)$ has a singularity at $z = z_0$, but can be shown to be analytic elsewhere. This means that we can shrink the integral down to a small circular loop of radius ϵ enclosing the pole. With $z - z_0 = \epsilon e^{i\varphi}$, so $dz = \epsilon e^{i\varphi} d\varphi$, the integral then becomes

$$\oint dz f = \oint \frac{dz f'(z)}{z - z_0} \simeq f'(z_0) \oint \frac{dz}{\epsilon e^{i\varphi}} = f'(z_0) \int d\varphi = 2\pi f'(z_0) \quad (\text{F.15})$$

where the second and third loop integrals are taken around the small circle. This approximation becomes exact in the limit $\epsilon \rightarrow 0$. We say that the integral becomes the *residue* of the integrand at the pole.

Index

- aberration, 121
- absolute magnitude, 334
- acceleration equation, 303
- action, 37, 186
- adiabatic baryonic, 349
- adiabatic compression, 259
- adiabatic equation of state, 73
- adiabatic expansion law, 73
- adiabatic flows, 251
- adiabatic index, 73
- adiabatic invariance, 42, 234
- adiabatic perturbation, 347
- aether drift, 19, 210
- Ampere's law, 93
- analytic complex functions, 425
- analytic continuation, 426
- analytic functions, 425
- angular diameter distance, 331
- angular surveys, 369
- anti-commutation relation, 231
- apodizing, 113
- apparent magnitude, 334
- Argand Diagram, 425
- astroid caustic, 103
- atmospheric seeing, 270
- atmospheric turbulence, 51, 113
- aureole, 114
- auto-correlation, 403
- auto-correlation function, 46
- axion, 233
- Babinet's principle, 111
- baryon asymmetry, 319
- Bayes' theorem, 57
- beating, 412
- Bernoulli effect, 253
- biased clustering, 390
- big crunch, 306
- Biot-Savart law, 132
- Birchoff's theorem, 304
- black body radiation, 69, 71
 - entropy of, 73
- Boltzmann equation, 241
- Boltzmann formula, 409
- Boltzmann transport equation, 242
- Boltzmann's constant, 63
- Bose condensation, 78
- Bose-Einstein distribution, 78, 233, 244
- bottom up scenario, 351
- box-car function, 404
- bremsstrahlung, 149
 - relativistic, 153
- brightness, 64
- brightness temperature, 77
- Brownian noise, 51
- bulk-flows, 369, 378
- Burger's equation, 389
- canonical momentum, 40
- canonical quantization procedure, 211
- carburetor, 253
- Cauchy-Riemann conditions, 425
- caustics, 102
 - random, 104
- central limit theorem, 55
 - multi-variate, 56
- centroid of the PSF, 118
- Cerenkov radiation, 143
- chaotic inflation, 322
- Chapman-Enskog procedure, 261
- charge conservation
 - covariant form, 421
- chion propagator, 225
- circular polarization, 97
- circulation, 254
- classical wave-particle correspondence principle, 209
- cluster mass function, 390
- clusters of galaxies, 154
- CMB dipole anisotropy, 378
- coarse-grained phase-space density, 299
- coefficient of shear viscosity, 261
- coherence length, 46, 48
- cold dark matter model, 351
- collision cross-section, 241
- collision term, 242
- collision time, 248
- collisional shock, 274
- collisionless Boltzmann equation, 241
- color temperature, 77

- comb function, 404
- commutation relation, 211
- comoving coordinates, 233, 308
- comoving peculiar velocity, 340
- comoving sound horizon, 343
- comoving spatial coordinate, 340
- complex
 - analysis, 425
 - function, 425
 - number, 425
- complex scalar field, 196
- Compton y -parameter, 172
- Compton cooling, 174
- Compton drag, 174
- Compton effect
 - inverse, 169
- Compton scattering, 167
- Compton wavelength, 168, 201
- conditional probability, 57
- conduction heat flux, 247
- conformal time, 309, 314, 330
- conservation laws, 246
- conservation of energy, 40
- conservation of momentum, 40
- conservation of particle number, 199
- conserved quantities, 246
- continuity equation, 246, 251
 - 6-dimensional, 241
- continuum limit, 185
- contour integral, 427
- contraction, 26
- contravariant vector, 25
- convective derivative, 242
- convective stability, 252
- convolution theorem, 403
- cooling
 - Compton, 174
 - synchrotron, 158
- cooling flows, 154
- cooling time, 154
- correlation function, 46
 - 2-point, 46
 - 3-point, 46
- correspondence principle, 201, 235
- cosmic drag, 342
- cosmic microwave background anisotropies, 380
- cosmic photosphere, 380
- cosmic ray, 143
- cosmic strings, 361, 365
- cosmic variance, 373
- cosmic virial theorem, 394
- cosmological density parameter, 306
- cosmological models, 303
- cosmological parameters, 305
- cosmology, 303
- covariance matrix, 56
- covariant vector, 25
- creation and destruction operators, 211
- critical density, 305
- critical frequency, 159
- critical sampling, 407
- cross product, 397
- cumulative mass function, 390
- curl, 397
- curvature eigenvalue, 313
- curvature of space-time, 313
- curvature tensor, 287
- d'Alembertian, 127
- damping term, 234
- dark halos, 293
- dark matter, 294
- Davis and Peebles scaling solution, 393
- de Broglie waves, 412
- de Laval nozzle, 273
- Debye length, 278, 279
- deep ocean gravity waves, 412
- deformation rate tensor, 261
- deformation tensor, 389
- density of states, 74
- detailed balance, 89
- differential mass function, 390
- differential momentum flux, 65
- diffraction
 - Fraunhofer, 108
 - Fresnel, 108
- diffraction theory, 107
- diffusion scale, 270
- dipole radiation, 132
- Dirac δ -function, 402
- discrete delta-function, 401
- discrete Fourier transform, 401
- dispersion measure, 280
- dispersion relation, 96, 411
 - gravity waves, 256
- displacement current, 93
- distance modulus, 334
- distinguishable particles, 409
- distortion tensor, 104
- div, 397
- divergence, 397
- divergence theorem, 398
- domain walls, 230, 361
- Doppler effect, 29
- Doppler formula, 30
- dot product, 397
- drunkards walk, 51, 84
- dynamical friction, 299

- dynamical time, 289
- eddies, 269
- Eddington approximation, 87
- effective temperature, 78
- eikonal equation, 101
- Einstein - de Sitter model, 305
- Einstein coefficients, 88
- Einstein relations, 88
- electromagnetic potentials, 127
- electromagnetic radiation, 63
- electromagnetic spectrum, 63
- electromagnetic waves, 95
- electron
 - classical radius of, 137
- emission measure, 154
- emissivity, 81
- energy conservation, 247
- energy eigenstate, 212
- energy equation, 303
- energy flux, 64
- enthalpy, 252
- entropy
 - statistical mechanical, 245
- equation of state, 208
- equivalence principle, 304
- escape speed, 288
- Euler equation, 248
- Euler's equation, 251
- Euler-Lagrange equation, 39, 187
- Eulerian coordinate, 102, 389
- Faber-Jackson relation, 378
- Faraday rotation, 280
- Faraday tensor, 422
- Faraday's law, 93
- fast Fourier transform, 404
- Fermat's principle, 102
- Fermi-Dirac distribution, 233, 244
- Feynman diagram, 220
- field theory, 179
- field-momentum, 190
- figure error, 121
- figure errors, 113
- fine structure constant, 150
- finger of god, 375
- first law of thermodynamics, 248
- flat rotation curve, 289
- flatness problem, 319
- flicker noise, 115, 271
- flicker-noise, 51
- fluid equations, 248
- Fock-Planck equation, 173
- fold catastrophe, 102
- force equation, 247
- four acceleration, 27
 - proper, 27
- four current, 421
- four momentum, 27
- four vector
 - scalar product, 25
- four velocity, 26
- four-vector, 24
 - norm of, 25
- Fourier transform, 401
- Fraunhofer diffraction, 108, 110
- free fields, 215
- free-free emission, 149
- free-free emissivity, 152
- Fresnel diffraction, 108
- Fresnel integrals, 109
- Fresnel length, 108
- Fresnel zone, 123
- Fried length, 115, 120, 270
- fundamental plane, 379
- galaxy clustering, 369
- galaxy luminosity function, 390
- Galilean relativity, 19
- gauge transformation, 127
- Gaunt factor, 151
- Gauss' law, 93, 288
- Gaussian fields, 57
- Gaussian random fields, 55
- general theory of relativity, 287
- generalized coordinate, 37
- generalized k-correction, 335
- generating function, 55
- geometric optics, 64, 101, 107
- geometrodynamics, 287
- global gauge transformation, 197
- grad, 397
- gradient operator, 25, 397
- grand unification, 321
- grand unified theories, 319
- gravitational binding energy, 288
- gravitational instability, 267
- gravitational potential, 287
- gravity waves, 256
- gravo-thermal catastrophe, 298
- group velocity, 96, 188, 280, 411, 412
- Gunn-Gott spherical accretion model, 387
- gyro frequency, 157
- gyro-frequency, 280
- H-theorem
 - Boltzmann's, 245
- Hamilton's equations, 41, 242

- Hamiltonian, 41
- Hamiltonian density, 186
- Harrison-Zel'dovich spectrum, 355
- Hawking radiation, 360
- heat flux, 261
- Heaviside radiation, 143
- Heisenberg picture, 213
- helicity, 97
- horizon problem, 319
- hot dark matter model, 351
- hot-dark-matter, 348
- Hubble parameter, 305
- Huygens' wavelets, 107
- hydro-dynamical simulations, 385
- hydrostatic equilibrium, 252
- ideal fluids, 251
- image wander, 118
- inertial range, 269
- inflation, 235, 319
- inflationary scenario, 319
- inflaton field, 321
- inner scale, 269
- integral constraint, 372
- interaction Hamiltonian, 192
- interaction picture, 213
- invariant volumes, 31
- inverse Compton effect, 169
- inverse discrete Fourier transform, 401
- inverse square law, 65, 287
- isentropic perturbation, 348
- isocurvature perturbation, 348
- isothermal perturbation, 347
- Jeans wavelength, 343
- Jeans' equations, 292
- jump conditions, 275
- k-correction, 335
- Kelvin's circulation theorem, 253
- Kelvin-Helmholtz instability, 267
- Keplerian profile, 289
- Kibble mechanism, 366
- kinematic constraints, 227
- kinematic Sunyaev Zel'dovich effect, 378
- kinematic SZ effect, 174
- kinematic viscosity, 263, 270
- kinetic theory, 232, 241
- Kirchoff's law, 83, 89
- Klein-Gordon equation, 230
 - in expanding coordinates, 234
- Klein-Nishina formula, 167
- Kolmogorov spectrum, 269
- Kolmogorov turbulence, 115
- Kompaneets equation, 173
- ladder operators, 212
- Lagrangian coordinate, 102, 389
- Lagrangian density, 186
- Lagrangian derivative, 246
- Lagrangian dynamics, 37
- Lagrangian mapping, 102, 389
- Laplacian operator, 397
- Larmor's Formula, 133
- law of inertia, 39
- Lenz's law, 93
- Lienard-Wiechart potentials, 129
- Limber's equation, 54, 376
- line element, 313
- Liouville operator, 242
- Liouville's equation, 242
- local supercluster, 379
- Lorentz force law, 93, 422
- Lorentz gamma-factor, 20
- Lorentz gauge condition, 128
 - covariant form, 421
- Lorentz transformation, 23
- Lorentz-Fitzgerald length contraction, 21
- Lorentzian profile, 138
- luminosity density, 307
- luminosity distance, 331
- luminosity function, 370
- Mach number, 269, 273
- magnetic potential, 127
- magnitude scale, 334
- mass conservation, 246
- mass shell, 28
- mass to light ratio, 293
- mass-shell condition, 168
- Maxwell's equation, 93
- Maxwell's equations, 127
 - covariant form, 421
- Maxwellian distribution, 151
- mean free path, 83, 84
 - effective, 85
- mean harmonic radius, 293
- mean occupation number, 75
- median
 - variance of, 59
- metric tensor, 287
- Michelson and Morley experiment, 19
- micro-lensing, 104
- microwave background radiation, 174, 312
- Minkowski metric, 25
- mirror roughness, 113
- missing matter, 295
- modulation transfer function, 113
- momentum conservation, 247
- momentum distribution function, 32

- momentum flux, 65
- momentum space density, 32
- monochromatic specific intensity, 64
- monopole problem, 319
- multi-pole expansion, 133
- N-body simulations, 385
- N-point correlation functions, 46
- N-point distribution functions, 45
- natural units, 230
- Navier-Stokes equations, 248
- negative specific heats, 297
- neutrino Jeans length, 348
- Newtonian gravity, 287
- Noether's theorem, 197
- non-thermal radiation, 157
- normal modes of oscillation, 179
- number operator, 212
- occultation experiments, 119
- occupation number representation, 179
- occupation numbers, 179, 409
- Olber's paradox, 69
- Oort limit, 295
- operators
 - creation and destruction, 211
- optical depth, 82
 - effective, 85
- optical transfer function, 112
- orbital period, 289
- outer product, 397
- outer scale, 269
- parsec, 334
- Parseval's theorem, 47, 403
- particle number, 200
- passive additives, 270
- past light cone, 130
- peculiar gravity, 341
- peculiar velocity, 340, 341
- phase screen, 108
- phase sheet, 298
- phase space, 241
- phase space density, 232
- phase structure function, 115
- phase transition, 362
- phase velocity, 96, 188, 411
- phase-space density, 33, 241
- phonons
 - scattering, 217
- photo-electric effect, 63
- photon discreteness, 150
- physical coordinates, 308
- pitch angle, 158
- Planck energy, 313
- Planck mass, 313
- Planck spectrum, 71, 73
- Planck time, 313
- Planck's constant, 63
- plasma frequency, 277
- point spread function, 112
- Poisson sampling model, 370
- Poisson's equation, 252, 288
- polarization, 96
- polarization of the CMB, 381
- potential flows, 255
- potentials
 - Lienard-Wiechart, 129
 - retarded, 128
- power law density profile, 289
- power spectrum, 47, 403
- Poynting vector, 95
- Press-Schechter approximation, 390
- Press-Schechter mass function, 390
- pressure tensor, 292
- principle of adiabatic invariance, 344
- principle of least action, 37
- principle of special relativity, 26
- probability density functional, 45
- proper distance, 21
- proper time, 20
- prototype 4-vector, 24
- pupil plane, 112
- quadrupole radiation, 134
- quantum electrodynamics, 231
- quantum fields, 211
- quantum mechanical statistical-factor, 233
- radiation constants, 77
- radiation damping, 138
- radiation force, 83
- radiation power spectrum, 96
- radiation pressure, 66
- radiation reaction, 137
- radiative flux, 64
- radiative transfer, 81
 - equation of, 82
- Raleigh-Jeans law, 76
- Raleigh-Jeans region, 75
- random field
 - projection of, 53
- random fields, 45
- random process, 45
- random walks, 84
- Rayleigh-Taylor instability, 267
- redshift, 329
- redshift of equality, 312
- redshift space distortion, 369

- redshift surveys, 369
- relativistic beaming, 29
- relativistic decays, 30
- relativistic field theory, 229
- relaxation time, 278, 291
- residue, 428
- resonance conditions, 194
- retarded potentials, 128
- retarded time, 130
- Reynold's number, 264, 269
- Richardson's criterion for stability, 271
- Rosseland approximation, 85
- rotation measure, 282

- S-matrix expansion, 214
- Sachs-Wolfe effect, 380
- sampling theorem, 405
- sampling theorem, 113
- sampling variance, 373
- scalar field, 229
 - Lagrangian density, 229
 - non-relativistic, 235
- scalar fields, 233
- scalar modes, 341
- scalar potential, 127
- scalar product, 397
- scaling solution, 363
- scattering rate, 220
- Schroedinger equation, 213, 235
- Schroedinger picture, 213
- scintillation, 119
- screening length, 279
- second-quantization, 227
- self interaction, 192
- self interactions, 219
- self-ordering fields, 361
- self-similar clustering, 391
- self-similar evolution model, 391
- Shah function, 404
- shock adiabat, 275
- shock waves, 273
- Silk damping, 348
- simple harmonic oscillator, 211
- sinc function, 404
- sinc interpolation, 407
- single scattering albedo, 85
- singular isothermal sphere, 289
- slow-roll condition, 324
- smoothing kernel, 51
- sonic boom, 143
- sound horizon, 343
- sound speed, 258
- sound waves, 258
- source function, 82

- space-time 4-volume, 21
- special relativity, 19
- specific energy density, 65
- specific heat, 297
- specific intensity, 64
 - conservation of, 33
 - constancy of, 67
- specific internal energy, 247
- specific volume, 275
- speckle interferometry, 118
- speckle noise, 117
- spectral energy distribution, 335
- spectral index, 51
- speed of light, 63
- spherical collapse model, 293, 385
- spinor, 231
- spontaneous symmetry breaking, 230
- Stefan-Boltzmann law, 72
- step function, 404
- stimulated emission, 88, 219
- stochastic process, 45
- Stokes' parameters, 97
- Stokes' theorem, 398
- stress energy tensor, 206
- stress-energy tensor, 287
- structure function, 51
- Sunyaev-Zel'dovich effect, 173
- super-luminal motion, 143
- supersonic flows, 273
- surface brightness, 64, 293
- surface brightness fluctuation, 379
- synchrotron
 - radiation pulse, 162
- synchrotron radiation, 157
 - spectrum of, 158

- Taylor-Sedov shock solution, 276
- tensor, 25
 - addition, 26
 - contraction of, 26
 - contravariant, 25
 - covariant, 25
 - mixed rank, 25
- texture, 367
- thermal bremsstrahlung, 151
- thermal bremsstrahlung absorption, 152
- thermal conductivity, 86
- thermal instability, 268
- thermal radiation, 63, 71
- thermalization length, 85
- thermodynamic entropy, 71
- Thomson cross-section, 137
- Thomson scattering, 135
- three-point correlation function, 46

- tidal field tensor, 304
- tidal radius, 300
- time dependent Schroedinger equation, 202
- time dilation, 19
- time dilation factor, 20
- time dilation formula, 20
- tip-tilt correction, 118
- top down scenario, 351
- top-hat function, 404
- topological defects, 230, 361
- topological invariant, 365
- total energy equation, 247
- transfer function, 349
- transition amplitude, 219, 243
- transition probability, 219
- transition rate, 219
- translational invariance, 46
- transmission function, 334
- transport coefficients, 261
- Tully-Fisher relation, 378
- turbulence, 269
- turbulent cascade, 269
- two point correlation function, 46
- two-point correlation function, 46

- vector calculus, 397
- vector modes, 341
- Vega magnitude scale, 334
- velocity dispersion tensor, 292
- velocity potential, 255
- venturi, 253, 273
- virial theorem, 293
- virtual particle, 226
- viscosity, 261
- viscous dissipation, 248
- viscous fluids, 261
- viscous shear tensor, 261
- viscous stress-tensor, 247
- vorticity, 253

- wave momentum, 220
- wave operator, 127
- wave packet, 107, 413
- wave uncertainty principle, 107
- wave-momentum, 190
- wave-optics, 122
- wave-packet, 188
- weak gravitational lensing, 382
- weak lensing, 369
- weak-field limit, 287
- white-noise, 48
- Wien displacement law, 77
- Wien law, 77
- Wiener-Khinchin theorem, 47, 403

- winding number, 365
- world line, 23

- Zel'dovich approximation, 388